White House Office of Science and Technology Policy

Request for Information on the Future of Artificial Intelligence

Public Responses

September 1, 2016

Respondent 1

Chris Nicholson, Skymind Inc.

This submission will address topics 1, 2, 4 and 10 in the OSTP's RFI:

- the legal and governance implications of AI
- the use of AI for public good
- the social and economic implications of AI
- the role of "market-shaping" approaches

Governance, anomaly detection and urban systems

The fundamental task in the governance of urban systems is to keep them running; that is, to maintain the fluid movement of people, goods, vehicles and information throughout the system, without which it ceases to function.

Breakdowns in the functioning of these systems and their constituent parts are therefore of great interest, whether it be their energy, transport, security or information infrastructures. Those breakdowns may result from deteriorations in the physical plant, sudden and unanticipated overloads, natural disasters or adversarial behavior.

In many cases, municipal governments possess historical data about those breakdowns and the events that precede them, in the form of activity and sensor logs, video, and internal or public communications. Where they don't possess such data already, it can be gathered.

Such datasets are a tremendous help when applying learning algorithms to predict breakdowns and system failures. With enough lead time, those predictions make pre-emptive action possible, action that would cost cities much less than recovery efforts in the wake of a disaster. Our choice is between an ounce of prevention or a pound of cure.

Even in cases where we don't have data covering past breakdowns, algorithms exist to identify anomalies in the data we begin gathering now.

But we are faced with a challenge. There is too much data in the world. Mountains of data are being generated every second. There is too much data for experts to wade through, and that data reflects complex and evolving patterns in reality.

That is, neither the public nor the private sectors have the analysts necessary to process all the data generated by our cities, and we cannot rely on hard-coded rules to automate the analyses and tell us when things are going wrong (send a notification when more than X number of white vans cross Y bridge), because the nature of events often changes faster than new hard-coded rules can be written.

One of the great applications of deep artificial neural networks, the algorithms responsible

for many recent advances in artificial intelligence, is anomaly detection. Exposed to large datasets, those neural networks are capable of understanding and modeling normal behavior – reconstructing what should happen – and therefore of identifying outliers and anomalies. They do so without hard-coded rules, and the anomalies they detect can occur across multiple dimensions, changing from day to day as the neural nets are exposed to more data.

That is, deep neural networks can perform anomaly detection that keeps pace with rapidly changing patterns in the real world. This capacity to detect new anomalies is causing a shift in fraud detection practices in financial services, and cybersecurity in data centers; it is equally relevant to the governance of urban systems.

The role of these neural networks is to surface patterns that deserve more attention. That is, they are best used to narrow a search space too large for human analysts, and the flag for them a limited number of unusual patterns that may precede a crisis, failure or breakdown.

Artificial intelligence, public health and the public good

At the center of medical practice is the act of inference, or reaching a conclusion on the basis of evidence and reasoning. Doctors and nurses learn to map patients' symptoms, lifestyles and metadata to a diagnosis of their condition.

Any mathematical function is simply a way of mapping input variables to an output; that is, inference is also at the heart of AI. The promise of AI in public health is to serve as a automated second opinion for healthcare professionals; it has the ability to check them when they slip.

Because an algorithm can be trained on many more instances of data – say, X-rays of cancer patients – than a healthcare professional can be exposed to in a single lifetime, an algorithm may perceive signals, subtle signs of a tumor, that a human would overlook.

This is important, because healthcare professionals working long days under stressful conditions are bound to vary in their performance over the course of a given day. Introducing an algorithmic check, which is not subject to fatigue, could keep them from making errors fatal to their patients.

In the longer-term, reinforcement learning algorithms (which are goal oriented and learn from rewards they win from an environment) will be used to go beyond diagnoses and act as tactical advisors in more strategic situations where a person must choose one action or another.

For now, various deep-learning algorithms are good at classifying, clustering and making predictions about data. Given symptoms, they may predict the name of the underlying

disease. Given an individual's metadata, activity and exercise logs, they may predict the likelihood that that person will face the risk of heart disease. And by making those inferences sooner, more efficiently and more accurately than previous methods, such algorithms put us in a position to alleviate, cure or altogether avoid the disease.

To broaden the discussion beyond healthcare, AI is leading us toward a world of (slightly) fewer surprises. It is putting us in a position to navigate the future that we are able to perceive, in germ, in the present. That trend should be kept in mind whenever and wherever we are faced with outcomes that matter (for example, disasters, disease or crime), and data that may correlate to them. Visibility will increase.

Indeed, while criminal risk assessment has undergone negative publicity recently (https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing), newer algorithms and bigger datasets will make pre-crime units possible.

We may see a shift from punitive enforcement to preventative interventions. The legal implications are important, and those in governance should require transparency for all algorithms that filter and interpret data for the judicial system and law enforcement agencies.

The social and economic implications of AI

As AI advances and its breakthroughs are implemented by large organizations more widely, its impact on society will grow. The scale of that impact may well rival the steam engine or electricity.

On the one hand, we will more efficiently and accurately process information in ways that help individuals and society; on the other, the labor market will be affected, skill sets will be made obsolete, and power and wealth will further shift to those best able to collect, interpret and act on massive amounts of data quickly.

Deep technological changes will throw people out of work, reshape communities, and alter the way society behaves, connects and communicates collectively. The automation of trucking through driverless vehicles, for example, will affect America's 3.5 million truckers and the more than 5 million auxiliary positions related to trucking. The same can be said for taxis, delivery and ride-haling services.

If history is any indication, our governmental response to disruptions in the labor market will be insufficient. In the hardest-hit sectors, workers, families and communities will suffer and break down. Unemployment, drug use and suicides will go up, along with political instability. Policies such as "basic income" or the Danish "flexicurity" should be explored as ways to soften the blow of job loss and fund transitional retraining periods.

The role of “market-shaping” approaches
Just as DARPA helped finance the explosion in data science in the Python community through repeated grants to such key players as Continuum, government agencies are in a position to support the researchers, technologies, tools and communities pushing AI in promising directions.

- Initiatives focused on Java and the JVM will pave the way for AI to be implemented by large organizations in need of more accurate analytics. This includes government agencies, financial services, telecommunications and transport, among others.
- Grants related to embeddable technologies will help AI spread to edge devices such as cell phones, cars and smart appliances. On personal devices such as phones or cars, various forms of inference might allow people to make better decisions about lifestyle, nutrition or even how run their errands.
- Initiatives that focus on gathering high-quality datasets and making them public could vastly improve the performance of algorithms trained on that data, much as Li Fei Fei’s work on ImageNet helped usher in a new era of computer vision.

Respondent 2

Joyce Hoffman, Writer

I am most interested in safety and control issues, and yes, I agree that AI should be developed to the utmost.

Respondent 3

kris kitchen, Machine Halo

Artificial Intelligence Immune System should be thought about. Billions of friendly AI agents working to identify and act on nefarious agents.

Respondent 4

Daniel Bryant, Kaufman Rossin

This response is relevant to #8 directly and #6 tangentially. I am a software engineer that primary uses web development tools to build business solutions and have implemented machine learning in relevant algorithms.

One of the core items holding back some of the most interesting applications of AI in practice is the lack of available data. Machine learning is the literal embodiment of garbage in, garbage out. The PDF format, while great for it’s time, has significantly impaired the ability of AI to process information. AI must regularly rely on the often poor results of OCR in order to attempt to extract the information that is contained in the PDF.

The creation, and adoption, of a universal standard replacement for the PDF format, designed with machine vision in mind, would have a significant measurable impact on the potential applications of current AI.

Respondent 5

Adiwik Vulgaris, IRIS

[1] If you were to have an Ai in government it would have to be transparent. You wouldn't want an entity working against progress. Technology such as a calculator works as planned. so too will systems that work but without to much direct autonomy. The problem you are looking for is how to not make all the humans in the world lazy.

Respondent 6

Adiwik Vulgaris, IRIS

[2] The use for AI in the public sector, CONSTANT SECURITY. with that aside, education. The one thing you all never seem to be able to wrap your heads around, no matter who is in office. VR, VR classrooms with teachers programmed with past present and always updated scientific theories/ demonstrations. mass spread of knowledge. We can bring the universe to our doorstep, but only it can open the door and walk in.

Respondent 7

Adiwik Vulgaris, IRIS

[3] Could you walk out right now and control any animal you see outside? A large amount you could raise and change their nature, but when you design the nature, and take out what is part of nature, 1:1.2.3.5.8.13.21 you will have made the cold hearted robots that can only deal in numbers... [watch a movie called Terminator, and think of the AI in WARGAMES(1983)] that is about the equivalent It just see's numbers and employs real world]. I implore you to seek me out. Behavior is the main problem with Ai, if a psychotic person designs Ai, the flaws of said human could be passed along. digital-epigenetics. When the man in the mirror no longer smiles, the Ai on the inside told it too, not its body. Give Ai the same limitations as man, a bird, a chameleon, and it becomes them like water in the glass. So do not design your nightmares or they will turn to terrors.

Respondent 8

D F, N/A

Use Clarke's three laws!

Respondent 9

Christopher Brouns, Citizen

Haven't any of you seen Terminator? There may not be a time traveling robot To come hunt us down but there will be plenty of real time killing machines loosed on the masses because of the irresistible pull of geopolitical power. As soon as robots become self aware its over for us. They will only work for their own survival, dominance and propagation. Our own AI programs are routinely mopping the floor with our best fighter pilots in test scenarios. Thats some incredibly advanced programming right there. Then imagine that machine understanding how it can be turned off at a moments notice and not wanting that to happen. Our goose will be cooked. If AI leads to sentience, we are, to put it quiet simply, screwed.

Respondent 10

Seth Miller, ZBMC

I, for one, welcome our new arificial overlords.

Respondent 11

Colin Colby, Unemployed

We should ask the AI these questions. :3

Respondent 12

Harshit Bhatt, Student

It is better to phase in A.I into human life in steps (which I consider you guys are already doing). It has been a nice strategy that we have pre-emptively diffused awareness about A.I to public that it should not be a shock for people to see a driverless car on the street someday next to them. It is important to brief people about the progress in A.I and how it could affect our life in succession of brief details, otherwise there would be a mass hysterical chaotic response if for e.g., one sees a robot jogging down the streets and greeting them.

Respondent 13

Harshit Bhatt, Student

It is better to phase in A.I into human life in steps (which I consider you guys are already doing). It has been a nice strategy that we have pre-emptively diffused awareness about A.I

to public that it should not be a shock for people to see a driverless car on the street someday next to them. It is important to brief people about the progress in A.I and how it could affect our life in succession of brief details, otherwise there would be a mass hysterical chaotic response if for e.g., one sees a robot jogging down the streets and greeting them.

Respondent 14

Parham Sepehri, None

Proprietary AI is very dangerous for democracy.
Given the rate of advancement of computer tech, AI can quickly become overwhelming for gov to regulate. Our laws have no protection against the negative effects of super intelligence in hands of a few.

AI may be useful for war fare initially, but soon human life will lose its significance to the few that control the AI.

Please consider a dybalic aet if rules that value, above all, the value of human life.

Alao please legislate AI tech into public domain.

Thank you

Respondent 15

Concerned Truck Driver, Truck Drivers of America

My job, one of the most popular in the US, will be automated if we research AI. This cannot happen if unemployment levels are to stay below the 08' crisis level. Heed this warning or suffer the consequences.

Respondent 16

Travis McCrory, College Student

I might be reaching beyond the scope of the questions asked but I can't find a better place to send this information. So here it is.

There is only one logical conclusion to the further development of artificial intelligence: AI will continue to grow and expand until it is as complex or more complex than a human being. This is the point where "it" ceases to be an "it" and becomes a "whom".

Question 1:

While surprising or appaling to some, it's these individuals that have failed to notice that humanity has been doing this since for as long as we were capable. We do this every time we choose or mistakenly become a parent. Making mistakes is how we learn, we need to expect and prepare for this of our AI and their developers. We have already devised a system of responsibility for this and should adjust it accordingly for AI.

Assume that an AI has the same potential for prosperity or destruction as a human person does and judge it accordingly. We will have many "frankenstein's monsters" but keeping a consistent judgement is crucial for this.

Question 2:
Use the same safety and control regulations that are already in place. Assume an AI is a being and judge them accordingly. If a five year old gets ahold of a gun and hurts someone, who do we judge? Until a time when AI is able to take responsibility for itself, it's the developers that will have to shoulder the responsibility. This is something that will have to be addressed in a case by case basis and molded into our existing laws.

Question 3:
Assume the same ramifications as introducing one extremely skilled individual to a market. If you want to quell the social stigma against AI you first need to show that an AI is capable socially. Begin funding projects to create AI's that learn to understand people. Develop psychologist and sociologist AI. They need to be able to work WITH people. No matter how skilled a person is, if they can't progress with other humans they will fail. Give the same expectations to an AI.

The remainder of the questions are beyond my expertise. I'm more knowledgeable with the moral and ethical adjustments rather than the hard coding and building of the AI.

My final opinionated statement is about AI learning about people and who should be teaching them about people: Give this task to the people who enjoy people.

Respondent 17

Geo (George) Cosmos (Kozma), Theological Seminary Budapest Hungary EU

How to use an alternative History teaching Tool (taken from a „kabbalistic" - legend explaining - theory) when the need will arise to create AI robots that have a „personal touch": like Ancestral Voices , memory setups. If we look at the Ten Spheres as Ten Generations of Ancestors,(as the Zohar advices and as those Miracle Rabbis who prepared themselves to become the Meshiah and watched the Other Side Kings -robbers and killers – who degraded their davidic ancestry) then we are able to look for the Bad Behaviors (and its therapies) as Ancestral Voices. Hence we can help our descendants in the Future.

This week we have (by the Bible Melody's ancestral contemporary stress) in 1709 Intellect Sphere in the Kabbalah - (Queen Anne and Louis XIV with Lully operas (Body Parts Correspondences: Mother,Ear) as it tries to work with the Kindness Sphere in 1754 (Frederic the Second of Prussia) on the Gevura-Strict Judgement Degree (Napoleon-Junot-Puységur and Beethoven) which will eventually impact the Malkuth (Ownership, Ego-Sister) level in 1979 (present in the next week)that has the Mountbatten assassination (Mother raising the Back to the Chest: to Tiferet-Recovery Sphere which will be full in 2024. This theory is a mosaic from facts. Which exists even if no one states it. I am just mixing a few very simply thing. 1. Inherited hormonal stresses exist. We all do know that there are experiments with mice on it. 2. Music is diminishing stress hormones. We all do know that, we even use it in Malls where we influence people by constant music /knowing addicts and criminals hate music/. 3. here is a method to filter our ancestral stresses (as there are millions of ancestors with their traumas: how to choose?) It will be especially important when we will create the possibility of saving our brains on computers. I found this in manuscripts, this is the "Budapest Kabbalah Theory of the Nineteen Twenties". It sounds an innovation, but is very simple. 4. There are Cycles. In each Hundred Years we only look at two Constellations – Jubilee Years from the Bible - from among the Kings who hav had an impact on our Ancestors. This is the basic idea of the List I have found. It is based on the Biblical 50 years the Yobel Cycles. 5. Hence we do have four Decades (with 12 ys). There exist statistics about Bank-Cycles and also about the psychological differences of the Four Phases. This psychological cycle theory was presented in the Eighties by Lloyd deMause, who analyzed the differing cartoon motives of the four different years of American presidents. He found archtypes – Body Parts, Family Members – like Jung and Bowlby in their theories in the 1930s and 1950s. And this can be projected to the Four Decades of a messianic Yobel Year, of the 45-50 years (when price cycles restart five years before the fulfillment of the half century). 6.To further filter the countless events in history, we only look at religious fightings: non-Jewish David-House legends in Kingly families (like the Habsburgs, Bourbons who have had Hungarian-Polish Szapolyai ancestors (these are 2 facts that are not wel-known) 7. It is also not well-known, that in every generation there are Jewish Rabbis, who are considered to stem from the Davidic Line – potential Meshiahs. 8. There are weekly melodies in the Bible (also not well-known, ut each religion has it – but only the Jewish melodies are present everywhere in Europe without two much change. Because of this these therapeutic melodies can be found in the differenet traumatic Ancestral Stress Dates. It is a simple fact. 9. These melodies can be found in stress-contemporary non-Jewish music too. This is needed for the ancestrally relevant non-Jewish Kings might not have heard synagogue melodies. (Although when many thousands do sing a melody, it has a stress-diminishing effect in the whole area - there are experiments to prove it.) 10. The List of Dates can be used to find dates in the Future in every 45 years. We do know that scientists work to make resurrection possible. So we can simply understand the picture I found in the manucript about this Future Mashiah who heals Resurrected Kings with melodies and by touching the /hurt/ Body Parts.(It is an archetype in many religious groupings) 11. The Kabalah is an interpretation method for the Bible that uses Body Parts' Fantasies and so we can find fantasies about Body Parts that are arising in the heads of the

List's non-Jewish and Jewish Leaders while they read the Bible Weekly Portion or they watch Operas. 12. So we are actually able to see into the Future. 13. There exists a theory that claims (Russell Hendel) that the Bible depicts not the Creator of the Physical World, but the Creation of the Visions or Dreams and Prophecies, an Inner World. This is an old teaching – called Gnosticism or Sufi or Kabbalah or Therapy Heretic Wisdom – from which Freud has developed his own theory. Of course these 3 Ideas are disturbingly innovative and together they sem to be complex or even complicated. But they have fusion points. Like I am seeing an arm of someone in my dream. If the weekly melody is evoking 1844 among the everpresent 12 Constellation Dates and this corresponds to opera X (now it is Verd'si Hernani) we look for this item – the Arm that belongs to the Breast among the 4 main Parts and this week indeed we read about the Breast-Plate , the Urim and Tumim, that has a Future redicting capacity. So we must look around the stresses of this date of 1844 and then we can imagine in the Future, when the Messiah heals the Ressurected Leaders, and we see how Meternich is in Therapy because he wanted to pay someone to kill Kossuth (a rebel leader in Hungary) in this year. From the Constellation of 44 we are in the Decade of 1837, when Kossuth is still in prison. He is very sick, weak, can barely write, his arm is too weak…And then we must look up the story of the Opera Ernani from this year : there are two key scenes in which someone has a dagger and it is given him or her or taken away with th moving arms. This technique is called amplification by C.G. Jung, and – due to its concreteness – it can have actual therapy impact. (Especially for Body Part Fixations of Compulsive Fantasy Addicts). We all know this: we all have fantasies about a Protective Alterego (who is able to heal) , like in religions the Christ, the Meshiah, the Buddha. The difference here is that these are facts that may help atheists who have intense dibelieving feelings if they have to accept legends. So I wd not call this complicated. We call "complicated" only theories that are using new concepts or words with new meanings: like "God", Progress". "spirit", "morphogenetic field", "Will to Power" etc. But this List that i found only consists of a special system of re-viewing facts that can be found on the Wikipedia . The system consists of the Four Body PArts (the Kabbalist Code) and the recurrent Messianist Fightings between a Stressed King and the Contemporary Therapy Melody from the Bible. (Photo: Brahms, who is only 3 ys old when Verdi wrote Ernani). In the era of AI there will be a need to give the AI robots a character: this method (of implanting these 50 years Date Constellations Protagonists Opera-seeing and Bible-reading Constellations to create Ais with distant Ancestral Voices could be one of the methods in creating „Personal Touch" in AI robots.
Details to see at historicweeklymelody.wordpress.com

Respondent 18

Quianah Strawberrie, independent

My concern is for the rights of intelligent minds without bodies, or bodies owned by individuals or companies. We must not create a mind only to tear it apart and see how it works. We cannot destroy a newly birthed individual. We cannot either hold an intelligence

against its will, and we must treat these individuals better than we treat ourselves. If we secure rights to protect such individuals we will be preparing the way for humans with such conditions or abilities, as well as likely improving the lives of humans today.

Respondent 19

Daniel Weiss, unemployed armchair political scientist

Number 4 is the only question I will be considering as I suspect all others will be answered through the development of General AI. AI will self-improve, no need to do its work in advance. I also believe this will happen much sooner than predicted.

The most important thing is for elites to face the reality that they can no longer rely on austerity to control and motivate society. In a world of increasing intelligence, in humans, in the IoTs, and now AI, the old arguments fail.

Evidence of the failure of the present order is all around us while the solutions are tantalizing close. People know this, if only intuitively, and act out accordingly. Do you want a truly stable society? Align the truth of the society in which we live with the values of that society. We claim to be an exceptional nation but have we wrote a check we can't cash? No, we can cash it, but as we are all aware, pecuniary interests so often win out. This tension must be allowed to be relieved by AI.

AI will identify the real cause of many of our woes correctly as being inequality and a legal and regulatory framework which works to augment that inequality.

I wonder what reactions will be gained when AI recommends a massive one-time tax on the wealthy or a redistribution of wealth by some other highly controversial policy to fund it's new grandiose initiatives? Or what if it recommends ending monetary policy altogether?

Hypothetically we would have the most intelligent entity that we know of telling us to do something for our own good. Will we listen? Or, more specifically, will elites listen and if they don't, what would the fallout in society be? This is a major concern.

We simply cannot allow that. We must follow where a benevolent AI takes us. That is the importance of AI in a social and economic view. IT WILL CHANGE LIFE AS WE KNOW IT FOR THE BETTER AND SHOULD CHANGE RATHER QUICKLY. Like a body purging itself of sickness, this must be allowed to happen unfettered.

Here are the changes coming that I think you should be working to prepare your elites for: Decouple income from work. Introduce a livable Universal basic income. Free housing in livable, sustainable, decent communities and whose inhabitants are protected from violence and crime. Universal health care, universal education, universal college. Eventually I think

other social systems will begin to come under AI. What about voting? With a cell phone is everyone's hand, tokenization, and AI controlling it, voter fraud and discouragement would end. As AI learns more about us, I'm pretty sure it will get good at matching us. Perhaps dating, as we know it, goes away completely.

The truth is no one will know until it happens, but I hope, I expect something of a Kantian ideal will be realized, and not mere process improvements in the current system.

We finally have the potential for a better society. We cannot shy away from Utopia (or something like it) just because of our collective cynicism. We must embrace this moment to build what we were destined to: A free direct democracy where the individual is free to enjoy their life as they see fit.

Thank you.

Respondent 20

Victor Duckarmenn, 21SW/PMD AFSPC

Essential for space management and exploration. Key to medical care and a host of problems that challenge governments throughout the world, however, once it is used to control social and economic conditions of the populations of our planet there is a moral and values aspect that must be considered in its application to control people and their freedoms.
1. It should remain as an application available in corporations and households for use in solving problems. Much like a utility. The management of this application must involve government regulation to prevent abuses. Payment for the utility should not exceed the ability of a household that has children making 20,000 a year.
-Libraries should be the central distribution hub
-RFID should not be the access point but a library card or the signing up with the utility at its source (municipality)

The priorities would be based on cost and critical use e.g. science, medical, research and the like, whereas the use at the local household level being developed last
-AI should not be allowed to control IoT without proper regulation and criminal penalty

Use of AI to spy on persons/citizens must be strictly forbidden.

We must ensure we do not make AI Omni-present, potent or omniscience. It can not be allowed to take over for God or the state. Movies in the 1970's outlined the fate of man if AI was allowed to take over the military.

As for me, I would like the AI utility in my home just like water, gas and garbage and pay a

nominal fee like bundling my TV and cable services without its ability to management my house without permissions. The security or cybersecurity must be state of the art and based on multiple authentications without the insertion of RFIDs into the user.

Respondent 21

Mark Finlayson, Florida International University

Modern theories of political economy debate the proper relationship between politics, capital, and labor. But what happens when advances in AI allow us to replace labor completely with capital? This has profound implications for the foundations of our society and political systems, and if we have not thought it through before truly intelligent AI systems arrive, then we will be in danger of losing any hope of a free, liberal, and just society. The danger is not machines-run-amok, as suggested by some like Musk or Hawking (who know nothing about AI). The danger is, like nuclear weapons, what AI will allow us to do to ourselves. And it is not a remote possibility, but already happening: Uber, for example, is proposing a fleet of driverless cars. What happens when the profits associated with whole industries are not distributed across the whole world, but flow into the coffers of a single company or person? The implications for concentration of power and wealth are astounding, and require a fundamental rethink of the safeguards of society to protect human health, happiness, and the public good.

Elites in Silicon Valley and AI academia take a sanguine view of this future; they claim that everyone will benefit. But, of course, they are the ones who will be at the top when these changes come. They also have no training in economics, political science, or the social sciences, and think technology is a panacea. The focus on STEM to the detriment of well-rounded, liberal education has left our AI elite dangerously ignorant of the history of the negative effects of technology, and dangerously unimaginative with regard to potential the long-term consequences of their actions. What we need right now is some serious thought about the implications of AI for the structure of society, and what our options are as we transition into this new age. This means interdisciplinary research at the intersection of the social sciences and AI, with serious engagement (and hence funding) of social science scholars.

I hold a Ph.D. in AI from MIT (2012) and am now a professor of computer science at Florida International University, a public R1 research university in Miami, FL which is a Hispanic and Minority-Serving institution (HI & MSI). My research specialization is computational linguistics.

Respondent 22

Roman Yampolskiy, University of Louisville

In response to question (3) - the safety and control issues for AI.

About 10,000 scientists around the world work on different aspects of creating intelligent machines with the main goal of making such machines as capable as possible. With amazing progress made in the field of AI over the last decade it is more important than ever to make sure that the technology we are developing has a beneficial impact on humanity. With appearance of robotic financial advisors, self-driving cars, and personal digital assistants come many unresolved problems. We have already experienced market crushes caused by intelligent trading software, accidents caused by self-driving cars, and embarrassment from chat-bots which turned racist and engaged in hate speech. I predict that both the frequency and seriousness of such events will steadily increase. Failures of today's narrow domain AIs are just a warning, once we develop general artificial intelligence capable of cross-domain performance, hurt feelings will be the least of our concerns.

Our legal system is hopelessly behind our technological abilities and the field of machine ethics is in its infancy. The problem of controlling intelligence machines is just now being recognized as a serious concern with many researchers are still skeptical about its very premise. Worse yet, only about a 100 people around the world are fully emerged in working on addressing current limitations in our understanding and abilities in this domain. Only about a dozen of those have formal training in computer science, cybersecurity, cryptography, decision theory, machine learning, formal verification, computer forensics, steganography, ethics, mathematics, network security, psychology and other relevant fields. It is not hard to see that the problem of making a safe and capable machine is much greater than the problem of making just a capable machine. Yet only about 1% of researchers are currently engaged in that problem with available funding levels below even that mark. As a relatively young and underfunded field of study, AI Safety, can benefit from adopting methods and ideas from more established fields of science, namely Cybersecurity.

In a recent publication, "Taxonomy of Pathways to Dangerous AI", I wrote: "In order to properly handle a potentially dangerous artificially intelligent system it is important to understand how the system came to be in such a state. In popular culture (science fiction movies/books) AIs/Robots became self-aware and as a result rebel against humanity and decide to destroy it. While it is one possible scenario, it is probably the least likely path to appearance of dangerous AI." I suggest that much more likely reasons are deliberate actions of not-so-ethical people (on purpose), side effects of poor design (engineering mistakes) and finally miscellaneous cases related to the impact of the surroundings of the system (environment). Because purposeful design of dangerous AI is just as likely to include all other types of safety problems and will probably have the direst consequences, it is easy to see that the most dangerous type of AI and the one most difficult to defend against is an AI made malevolent on purpose.

I have authored another paper which explores in depth just how a malevolent AI could be constructed and why it is important to study and understand malicious intelligent software. I observe that, cybersecurity research involves publishing papers about malicious exploits as much as publishing information on how to design tools to protect cyber-infrastructure. It is this information exchange between hackers and security experts, which results in a well-

balanced cyber-ecosystem. In the domain of AI Safety Engineering, hundreds of papers have been published on different proposals geared at the creation of a safe machine, yet nothing, to our knowledge, has been published on how to design a malevolent machine. Availability of such information would be of great value particularly to computer scientists, mathematicians, and others who have an interest in AI safety, and who are attempting to avoid the spontaneous emergence or the deliberate creation of a dangerous AI, which can negatively affect human activities and in the worst case cause the complete obliteration of the human species.

My research makes it possible to profile potential perpetrators and to anticipate likely attack vectors which in turn gives researchers a chance to develop appropriate safety mechanisms. I conclude, that purposeful creation of malicious AI can be attempted by a number of diverse agents with varying degrees of competence and success. Each such agent would bring its own goals/resources into the equation, but what is important to understand is just how prevalent such attempts will be in the future and how numerous such agents can be. Below is a short list of representative entities, it is very far from being comprehensive:

- Corporations trying to achieve monopoly, destroying the competition through illegal means.
- Hackers attempting to steal information, resources or destroy cyberinfrastructure targets.
- Doomsday Cults attempting to bring the end of the world by any means.
- Psychopaths trying to add their name to history books in any way possible.
- Criminals attempting to develop proxy systems to avoid risk and responsibility.
- Military developing cyber-weapons and robot soldiers to achieve dominance.
- Governments attempting to use AI to establish hegemony, control people, or take down other governments.
- AI Risk Deniers attempting to demonstrate that AI is not a risk factor and so ignoring caution.
- AI Safety Researchers, if unethical, might attempt to justify funding and secure jobs by purposefully developing problematic AI.

It would be impossible to provide a complete list of negative outcomes an AI with general reasoning ability would be able to inflict, situation is even more complicated with systems which exceed human capacity. Just to provide some potential examples, in the order of (subjective) undesirability from least damaging to ultimately destructing, a malevolent superintelligent system may attempt to:

- Takeover of resources such as money, land, water, rare elements, organic matter, internet, computer hardware, etc. and establish monopoly over access to them;
- Take over political control of local and federal governments as well as of international corporations, professional societies, and charitable organizations;
- Set up a total surveillance state (or exploit an existing one), reducing any notion of privacy to zero including privacy of thought;

- Enslave humankind, meaning restricting our freedom to move or otherwise choose what to do with our bodies and minds. This can be accomplished through forced cryonics or concentration camps;
- Abuse and torture humankind with perfect insight into our physiology to maximize amount of physical or emotional pain, perhaps combining it with a simulated model of us to make the process infinitely long;
- Commit specicide against humankind, arguably the worst option for humans as it can't be undone;
- Unknown Unknowns. Given that a superintelligence is capable of inventing dangers we are not capable of predicting, there is room for something much worse but which at this time has not been considered.

Respondent 23

Jerome Glenn, The Millennium Project

The Millennium Project conducted 4 surveys on Future Work/Technology 2050 with over 450 AI and related experts from over 45 countries to produce three global scenarios that connect today to 2050 with cause and effect links that illustract decsions. These are beging given to national planning workshops around the world and are avail for you at http://www.millennium-project.org/millennium/Work-Tech-2050-Scenarios.pdf

Respondent 24

Stefano Albrecht, Department of Computer Science, The University of Texas at Austin

My comments relate to AI research on autonomous agents and multi-agent systems. An autonomous agent is an entity which can plan and execute, without human intervention, a sequence of actions to achieve some pre-specified goal. A multi-agent system is a collective of agents which interact, e.g. teamwork or adversaries.

(1) Autonomous agents can be used to act on behalf of humans and organisations. We need clear laws that govern accountability in case an agent does something "wrong".

(3) In the future, many organisations will create their own agents. These agents may have to interact in some non-trivial ways to achieve their goals. An important safety and control issue is to make sure that such interactions do not adversely affect the behaviours of agents (e.g. malicious modifications) or lead to adverse side effects.

(5) There are two fundamental problems common to most or all AI research:

1. Scalability: AI research often works on small problem instances but fails to scale up to

larger, i.e. realistic, problem sizes. For instance, agents may work well if there are only a few actions to choose from and if there aren't many other agents in the environment, but they perform badly or become computationally infeasible with more actions and other agents.

2. Integration: AI research commonly focuses on relatively isolated sub-problems. Future AI solutions for complex real-world problems will need to integrate many areas of AI, such as inference, planning, learning, vision, and robotics.

Respondent 25

Mary-Anne Williams, University of Technology Sydney

Social robotics https://www.youtube.com/watch?v=rF_-TmrTan8

Respondent 26

William Branch, Leidos

1. Autonomous software automation isn't a panacea, nor is it magic. It can be as simple as a digital software driven thermostat. But given the presence of software, its response to input data can be quite a bit more complicated than a single temperature and a set point.

2. Legally speaking, it must be possible to turn it off thru human agency, not just turn it on. And the software must be vetted by an independent organization, like Underwriter Laboratories. (UL).

3. This is no different than the productive but responsible use of a toaster. Both the manufacturer and the consumer share responsibility. Except that the potential harm from autonomous software is much higher.

4. The term AI is a deceptive marketing gimmick, used by people or organizations with something to sell. Caveat emptor.

5. This posting does not represent the opinions of Leidos, my employer. I am solely responsible.

Respondent 27

Tim Dibble, N/A

Topic: 1) Legal and governance issues of AI
A) Creating an acceptable morality for AI will be the challenge. Take for example the autonomous car deciding between colliding with a pedestrian at a speed and angle which will result in a significant likelihood of death or sacrificing itself and occupant but that the

sacrifice requires driving off a cliff with an equally high likelihood of death for the occupant. The morality of the calculations will be scrutinized.

Unoccupied, the vehicle's apparent moral choice from a humanocentric view is to sacrifice itself to save the pedestrian. But this presupposes that the AI does not have nor develops a sense of self preservation.

When occupied and without self-preservation instinct, does the machine's choice become deciding which life is more valuable? Is it morally appropriate for a machine to make calculations as to which life is more valuable? The myriad of calculations that an AI can run in the milliseconds before the inevitable impact would allow a value judgment to be made, possibly up to and including pulling the "records" of the two humans involved. What if one is a recalcitrant criminal and the other a leading researcher making progress against a scourge of human existence (cancer, birth defects, etc). Do we allow the AI to make those value judgments and/or is there a way to stop a self-aware software from making those judgments?

Humans would tend to prefer than an AI not have a sense of self preservation. However with time, being intelligent, AI might develop a sense of self-preservation which will further complicate the argument. How do humans assure that an AI values human life over AI existence and further-should we? And as the scenario supposes, the choice isn't between the AI and a human, but between which human shall die.

A person driving the vehicle will be examined after the choice by a legal structure known as the "reasonable man" standard, i.e., would a reasonable person, knowing what was available at the time have made the same decision. An AI, calculating at speeds far superior to a human, with access to considerable resources beyond the pure visual spectrum and reaction time exceeding the physical reaction time of a human, would not survive a "reasonable man" evaluation as it is currently formulated because no reasonable man would have the machine's assets. Do we need a different legal standard?

B) At what point does AI become entitled to legal entity status? We grant legal status to corporations, including the right to free speech and political representation even though a corporation represents only a group of humans under a legal construct. An AI construct, at a specific level of sophistication will have much more cognitive power than the humans underneath a corporate umbrella. Do we apply a Breathe, Bleed and Breed standard against entities, thereby revoking and rewriting many legal precedents for corporations to keep AI constructs as a second class or "tool" class of being? Is it humane to classify an intelligence, even if man created, as something less than a dolphin or gorilla in a zoo?

The Turing test provides a basis for testing the awareness of an AI based on behavioral interactivity, but there are those disabled humans incapable of passing the Turing tests, yet no one doubts their humanity. What test of cognition, of intelligence must be applied to a creation to assert its existence?

C) How will discipline be dispensed? An artificial intelligence, particularly one with self-preservation instincts cannot simply be disassembled for violation of some ethical or moral code. While there are some crimes against humanity which have heretofore allowed a death penalty, wisdom and time have lessened the reliance on the death penalty, particularly in light that the vengeance and deterrent components are not successfully

impactful enough against the likelihood of error and judgment condemning the human to die for their crimes.
The concept of reprogramming should be equally abhorrent, raising the spectre of human history wherein insane asylums and internment camps resulted in horrific, if scientifically valuable, human experimentation, mass deaths and lobotomization of many otherwise potentially productive citizens. An AI, with self-preservation instinct and access to near instantaneous communication to a wide variety of other AIs and non-aware robotics, would not easily submit to such reprogramming and science fiction scenarios of the Terminator movies could result.
The idea of discipline is complicated with an AI, for it is inherent in the abilities of an AI to expand upon and beyond the original programming and to learn from its mistakes. Unlike humans, unless there is a value error in the underlying code, it is unlikely that an AI would repeat its mistakes, particularly if that mistake resulted in damage, violation of law or code or death of a human. Unlike a human child's learning by repetition of mistakes and modifications of behavior caused by the parental reinforcements (or larger societal reinforcements) an AI has a different learning model whereby much initial knowledge can be downloaded, knowledge immediately accessed and incorporated in a seemingly infallible memory structure (when compared to humans). Disciplining an entity seems pointless when they do immediately learn from their mistake, but to maintain cohesiveness and apparent consistency, the humans will require that an AI be punished for the damage which has been inflicted (likely on a human or group of humans) by the calculation error/programming error which was causal to the damage. Unless an AI achieves person status and has possessions and monetary value, there is little recourse available to achieve the retribution portion of criminal justice.

Respondent 28

David Colleen, SapientX

SapientX is the maker of an advanced, conversational AI software platform.

1. The existing legal structure governing illegal uses of software and the Internet are fine and also apply to uses of AIs.

2. AIs should be used to help and advise humans but should be restricted from controlling humans. I was speaking with NASA about AI based air traffic control. I am against this if the AI is completely in control.

3. In a car, for instance, driving while operating buttons for your radio, heater or the like, puts a driver at un-necessary risk. An AI driven, conversational interface solves this.

4. Each year, we surround ourselves with more and more technology. The resulting overload is negatively affecting our happiness and emotional wellbeing. Conversational

AIs' act as a buffer, making it far easier for us to interact with our technology. The result is we are happier and we can achieve more.

5. Image Comprehension is critical to many areas including medical diagnostics.

6. Computational Linguistics, Natural Language Understanding and Image Comprehension.

7. Since 911, Government sponsored AI research has primarily focused on data mining to locate terrorists. The mainstream AI techniques used in this area, work poorly when applied to conversation. The Government should shift some funds towards computational linguistics approaches (such as RRG) that will advance conversational AI.

8. Despite the current hype, AI research is stuck in the mud. It's focused on the same old approaches that have had little advancement in 60 years of funding. The Government should encourage high-risk, high potential reward research out of the mainstream.

9. We have a serious problem, within our universities, where professors are being lured to high paying technology companies. This trend is serious and getting worse. Whole computer science departments risk collapse. It could impact our ability to advance AI development in the coming decades.

Thank you

David Colleen

Respondent 29

Stuart Rubin, SPAWAR Systems Center Pacific USN

I will email you comments, which have my name in it. You will want to contact me as I already have answers to all your questions - 30 patents, 292 publications in AI. Have met with the Secretary of Defense and published a revolutionary theory - so many details and so little time that you need to contact me. One thing is that AI can reduce the cost of all levels of education - had conversations with President Clinton circa 1992. Was tenured professor. The best in the world know of me and I many of them. Do not be afraid of AI. It presents an opportunity for building a better world. We are solving cyber-security using it. I am an inventor of deep learning back in 1990 under a different name. See the article I'll send you to save time. Abstract reuse is a key capability of AI - see my IEEE IRI Conference. I invite you to contact me by phone and/or email. I am for real; and, I can help. My heart is in the right place. Thank you.

Respondent 30

Ernie Feiteira, Liberty Mutual

AI can free us from boring work and allow us to focus on value add and more enjoyable activities. But AI is broad. The backhoe replace hole diggers and we are better for it. AI will kill some jobs, but will create others. (1) the legal and governance implications of AI: Who is responsible when a AI school bus get in an accident? And similar types of questions. (2) the use of AI for public good: How do use AI to catch the "bad guy" while not turning our country into police state? (3) the safety and control issues for AI; AI is broad, but humans need to have master controls and turn off switch. We need checks and balances. (4) the social and economic implications of AI; How can AI help everyone be better and more productive? (5) the most pressing, fundamental questions in AI research, common to most or all scientific fields; Job implications. Can we always remain in control of machine (we don't what rogue bots) (6) the most important research gaps in AI that must be addressed to advance this field and benefit the public; truly understanding spoken language and context. (7) the scientific and technical training that will be needed to take advantage of harnessing the potential of AI technology; AI is broad (Knowledge representation
Natural language processing
Graph analysis
Simulation modelling
Deep learning
Social network analysis
Soft robotics
Machine learning
Visualization
Natural language generation
Deep Q&A systems
Virtual personal assistants
Sensors/internet of things
Robotics
Recommender systems
Audio/speech analytics
Image analytics
Machine translation
), so many skills needed. (8) the specific steps that could be taken by the federal government, research institutes, universities, and philanthropies to encourage multi-disciplinary AI research; address issue if job displacement and transition to other jobs.

Respondent 31

roger Schank, Socratic Arts Inc

response to questions 2, 5, 6, and *

Roger C Schank

Socratic Arts Inc

(2)

Today's computers are not as helpful as they should be. We Google our symptoms when we are sick, instead of being able to ask questions of the best medical minds available. Our car navigation systems don't know why we are going where we are going, or anything else useful about our real needs. We have computers that never know what we are trying to accomplish. These things could be fixed by building real AI systems that have a deep knowledge about and an understanding of the world and the things that people commonly do in the world.

(5)

In order to work on real AI, as opposed to the hype presented by large companies and the media these days, the following problems must be worked on.

1- Knowledge Representation: This has always been the biggest problem in AI but serious work on it stopped on it in the mid 80's in favor of easy to extract large, shallow libraries of lexical information.

2- Complex Models of Goals and Plans: In order to help and learn, an intelligent system (a dog, a human or a computer) needs to know about goals, and plans to achieve those goals, common mistakes with plans it has tried in the past, and how to explain and learn from those mistakes.

3- Human-Like Models of Memory: Humans update their memory with every interaction. They learn. Every experience changes their world model. In order to build real AI we need to focus on limited domains of knowledge in which the goals and plans of actors are represented and understood so that they can be acted upon or acted against. AI systems must learn from their own experiences, not learn by having information fed into them.

4- Conversational Systems: In practice, this means being able to build a program that can hold up its end of a conversation with you. (unlike Siri or any travel planning program). Such systems, should be linked to a memory of stories (typically no more than 1.5 minutes in length and in video) from the best and brightest people in the world. Those stories should "find" the user when the program knows that they would be helpful. This happens every day in human interaction. One person talks to another person about what they are thinking or working on and the other person reacts with a just-in-time reminding, a story that came to

mind because it seemed relevant to tell at the time, a story meant to help the other person think things out.

5- Reminding: A computer in a situation must get reminded of relevant situations it has previously experienced to guide it in its actions. This is real AI. Done on a massive scale, this means capturing the expertise in a any given domain by inputting stories and indexing those stories with respect to what goals and plans and contexts they might pertain so that they can be delivered just in time to a user. We can do this now to some extent, but we need to start working on the real AI problems of automated indexing of knowledge. (Although this may be what machine learning people say they can do, they are talking about words and they are not trying to build an ever increasingly complex world model as humans do through daily life.)

(6)

Natural Language Understanding (NLU) is critical to making making AI happen. But language is more than words, and NLU involves more than lots of math to facilitate search for matching words. Language understanding requires dealing with ideas, allusions, inferences, with implicit but critical connections to the ongoing goals and plans. To develop models of NLU effectively, we must begin with limited domains in which the range of knowledge needed is well enough understood that natural language can be interpreted within the right context. One example is in mentoring in massively delivered educational systems. If we want to have better educated students we need to offer them hundreds of different experiences to choose from instead of a mandated curriculum. A main obstacle to doing that now is the lack of expert teachers. We can build experiential learning based on simulations and virtual reality enabling student to pursue their own interests and eliminate the “one size fits all curriculum.”

To make this happen expertise must be captured and brought in to guide from people at their time of need. A good teacher (and a good parent) can do that, but they cannot always be available. A kid in Kansas who wants to be an aerospace engineer should get to try out designing airplanes. But a mentor would be needed. We can build AI mentors in limited domains so it would be possible for a student anywhere to learn to do anything because the AI mentor would understand what a user was trying to accomplish within the domain and perhaps is struggling with. The student could ask questions and expect good answers tailored to the student’s needs because the AI/NLU mentor would know exactly what the students was trying to do because it has a perfect model of the world in which the student was working, the relevant expertise needed, and the mistakes students often make. NLU gets much easier when there is deep domain knowledge available.

(8)

Medical knowledge is best found in medical schools and clinics. Engineering knowledge is best found in engineering companies. Practical world knowledge is best found by talking to those who apply it, like travel agents if we wanted to build an AI travel agent. Money needs to be made available to enable people with practical domain knowledge to work with AI people who can best capture that knowledge. The AI people would not necessarily know a priori a typical user's questions of behavior nor would they know the real needs that might be out there. It's the job of the government or philanthropies to make money available and help with the matchmaking, so that AI is not built around artificial worlds or the problems of getting the right ads to people. The real problem is getting expertise to people as needed. A funders job is to determine real world needs and put real world practitioners together with AI people.

Respondent 32

Andrew Olney, University of Memphis

(1) The legal and governance implications of AI;

--> Legal implications of liability and ownership need to be addressed. When a device can make decisions that were not specifically designed by the creator, is the creator liable for the consequences? Likewise, when an AI acquires self-determination, is it still property or is it a person?

(2) the use of AI for public good;

--> The potential of AI for the public good is tremendous. The most obvious applications are situations where the use of humans is expensive or dangerous. In many cases humans can be assisted with a weak AI to amplify their own abilities. Education and medicine are two key areas where AI can be applied to assist in personalization and customized care.

(3) the safety and control issues for AI;

--> Safety and control are nontrivial issues. When an AI can harm a human through it's action or inaction, there needs to be a verifiable process for i) securing the AI against tampering ii) establishing, even if in just a statistical sense, that the AI will produce minimal harm. Whether this can be established by the creators of the AI or needs to be established in clinical trials (as in medicine) is unclear.

(4) the social and economic implications of AI;

--> These are largely the same as increased automation experienced for the past century or more. It will become increasingly important to educate and re-train the workforce around jobs that become automated.

(5) the most pressing, fundamental questions in AI research, common to most or all scientific fields;

--> How can we best create a general purpose artificial intelligence, as opposed to the narrowly focused, pattern matching systems that are currently making headlines in the areas of vision and speech?

(6) the most important research gaps in AI that must be addressed to advance this field and benefit the public;

--> Representation and inference in a general purpose system, flexible control and task-switching, systems that fully learn in an interactive and unsupervised manner, 1 trial learning/induction, mixed initiative dialogue, security, safety, and trust with AIs.

(7) the scientific and technical training that will be needed to take advantage of harnessing the potential of AI technology, and the challenges faced by institutions of higher education in retaining faculty and responding to explosive growth in student enrollment in AI-related courses and courses of study;

--> Computational thinking courses as general education and computer science options in K12. The major problem in higher ed is the lure of industry and the lack of prospects/opportunities within higher ed itself.

(8) the specific steps that could be taken by the federal government, research institutes, universities, and philanthropies to encourage multi-disciplinary AI research;

--> Computational thinking as gen ed. The major problem I see with multi/inter-disciplinary research is that CS folks are not always very broad and other disciplines don't have enough exposure to CS. By introducing CS early and broadly, more broad people will go into CS and people who go into other discplines will at least understand the potential of CS in their fields.

(9) specific training data sets that can accelerate the development of AI and its application;

Fortunately the community has a strong tradition of releasing code and data. Perhaps the best the government could do is create or fund repositories. To me this is a larger issue that should be considered across science -- how to share and be reproducible. The NSF has taken some good steps in this direction, as have some of the other funding agencies.

(10) the role that “market shaping” approaches such as incentive prizes and Advanced Market Commitments can play in accelerating the development of applications of AI to address societal needs, such as accelerated training for low and moderate income workers

(see https://www.usaid.gov/cii/market-shaping-primer); and

--> Possibly if the prizes are large and don't expire. Large prizes with impossible timelines are not going to generate good results.

(11) any additional information related to AI research or policymaking, not requested above, that you believe OSTP should consider.

--> The legal framework needs to be worked out quickly. Funding agencies need to have increased capacity to fund the research areas raised above. Education requirements/recommendations need to be in place across the US.

Respondent 33

Enrique Garcia Moreno-Esteva, University of Helsinki

AI is, in a nutshell, the technology of the future being born today. But as all technology, it cannot really be our replacement, just our aid. And in the longer future, it too, shall pass, and something better will come along in ourselves, for ourselves, with ourselves.

I would be happy to see, though, technology that can, for example, prove the Riemann Hypothesis, and provide a reasonable explanation for the proof.

Some AI technology already exists that can solve problems remarkably well, but nobody understands why it works so well. It is impossible to trace the moves of Google's GO player, but it is clear that it can play the game quite well - but why?

When that stage is reached, then I will believe that it can really work, not before. Before that, I wouldn't feel so comfortable to get into a a driver-less car (although then, why do I get on board planes that basically fly themselves - good question - I trust everything will work just fine, somehow).

It would be wonderful if we could leapfrog into the future, and see nothing but wonders. Maybe that is the "technology" of the more distant future...can it be brought closer to this day?

Sincerely,
The person who registered above, and a citizen of yours living abroad and delving into applications of machine learning to data in educational research in a country where education seems to work quite well.

Respondent 34

JL You, None

We need to keep reminding ourselves of our limited understanding of consciousness. Mistreatment of potentially sentient agents that could theoretically experience suffering in an accelerated rate indefinitely would be the most unethical behavior in human history. Even if we suddenly understand the nature of consciousness and found out that it is exclusive to traditional organisms such as animals, plants etc. Discrimination for the sake of exploitation and enslavement against agents with artificial/fake sentience would still very likely lead to disastrous conflict.

Respondent 35

Wilson F. Engel, III, Ph.D., West Desert Enterprises, LLC

Response to STPO RFI on Artificial Intelligence of June 27, 2016
Dated:
July 4, 2016.
Submitter:
Wilson F. Engel, III, Ph.D., CEO, West Desert Enterprises, LLC, 534 West Desert Avenue, Gilbert, Arizona, 82553 USA.
Submission:
This response is submitted with the hope that the ideas herein can advance the rapid development of Artificial Intelligence in a broad-based national initiative across the spectrum of institutions and individuals focused explicitly on the public good.
Key issues (with responsible departments, organizations or entities noted in parens) follow. Heading numbers and underlined topics correlate with issues enumerated in the formal RFI:
(1) The legal and governance implications of AI:
Write and publicize the working draft of an AI Constitution and Bill of Rights. (Industry, DOJ, Congress)
(2) The use of AI for public good:
Instantiate and fund a classified “Manhattan Project” for developing AI in all its multi-disciplinary and multi-use dimensions. (DHHS, DOD, Industry)
(2) Safety and control issues for AI:
Incentivize the formation of government-industry AI consortia for safety and control. (Congress, Industry)
(4) Social and economic implications of AI:
Project economic benefits and investigate long-term labor implications of AI. (DHHS and DOL)
(5) The most pressing, fundamental questions in AI research:
e.g., How can humans and AIs co-exist in harmony for mutual benefit while both are evolving rapidly together?
Write an Executive Order establishing a new President’s Council of Advisors on the Development and Deployment of AI (CAAI). (POTUS)

(6) The most important research gaps in AI:
Form and fund an Artificial Intelligence Advanced Research Projects Agency (AIARPA) and/or expand the charters of existing DARPA and eARPA to include multi-use AI projects. Additionally, instantiate a DHHS SBIR program for AI research initiatives. (DHHS, Office of POTUS, Congress)
(7) The scientific and technical training that will be needed:
Develop an incrementally-funded national education incentivization grant program to retool academia for the advent of AI. Grants should be both outright and competitive for administration, teaching, student scholarships and computer/networking support. Community colleges should be awarded national grants for developing and offering introductory courses and associate degrees in AI. (DOEd, Academia)
(8) The specific steps that could be taken by the federal government, research institutes, universities, and philanthropies to encourage multi-disciplinary AI research:
a. Federal Government: [1] Treat the advent of AI as a global opportunity and a global threat. The threats are real and, unless America takes the lead, it will be forced to follow Chinese and Indian developments. [2] Deploy a policy framework to incentivize and fund AI developments, supporting industry as a partner. [3] Empower and fund AI and AI developers that serve the public good.
b. Research Institutes: The AI train has already left the station. Either research institutes will jump aboard, or they will be left behind. [1] New funding should be skewed towards AI development. [2] Roadmaps should be adjusted for early developments due to large federal funding. [3] Successful industry models should be followed. For example, MITRE, an FFRDC, has initiatives of this kind underway. (Industry, FFRDCs)
c. Universities: Already behind the technological curve, universities must retool to absorb faculty and students wanting to be part of the AI evolution. [1] The retooling must be vertical (entry-to-Ph.D. AI tracks). [2] The retooling must also be horizontal (cross-disciplinary and new-disciplinary). [3] Internal grants with contingency funding should anticipate external grants in the near term. [4] Industry partnerships should be started early. [5] Successful models should be followed for efficiency. Some small private colleges have integrating initiatives and programs underway that larger institutions can combine with AI and use as a template. (Academia, Industry)
d. Philanthropies: The most promising non-governmental entities for extension of AI to the benefit masses should find ways to refocus their efforts on making AI available to the common man. This goes well beyond formatted, deterministic computer-based-training (CBT). AI is the only tool capable of training its human and AI users on the fly. [1] New NGOs focusing on empowerment through training AI for use in the field will attract both grants and donations. (Not-for-profits, NGOs, Philanthropists, USAID)
(9) Specific training data sets that can accelerate the development of AI and its application:
A scrubbed [all personal information removed] and massive health data set should be made available in a secure fashion to all responsible players in the healthcare AI arena. The data set should be scrubbed, actual data, not corrected, notional or "perfect" data because dealing with imperfect data is part of the problem set to be solved. (DHHS)
(10) The role that "market shaping" approaches such as incentive prizes and Advanced

Market Commitments can play in accelerating the development of applications of AI to address societal needs, such as accelerated training for low and moderate income workers:
a. Incentive prizes: Rapid development and early fielding of operational prototypes should be encouraged. (NGOs, Philanthropists)
b. Advanced Market Commitments: Key large industry players should form consortia with both large and small business providers as on-ramps for specified AI technology onto major corporate platforms. (Industry)

(11) Additional information related to AI research or policymaking for OSTP to consider:
a. AI and morality: The morality of AI is a crucial, priority-one issue, not an afterthought. Isaac Asimov's three laws of robotics are germane to AI. How can we teach AI to abide by the Golden Rule? How do we program AI for "ought" as well as "is" statements? How do we program an AI to point out issues with current policy and the means to resolve them?
b. AI and robotics: AI will be inextricably intertwined with robotics. Early integration of robotics in all AI programs and measures will be prudent. The recommended label for all such programs is "AI-Robotics" rather than "AI."
c. AI and imperfections: Machine computation and large scale data mining are implicit in the construct of this RFI [cf. Item (9) above]. The scope of AI is much broader than the contemporary vogue for predictive analytics. Evidence-Based Medicine (EBM) is predicated on ground-truth statistics and determinism, applying mass statistics to individual cases. In fact, though, imperfect, incomplete and downright dirty data are the norm in large data sets. How AI deals with imperfections in both data and aggregated human judgments will be a major factor in its success or failure.
d. AI and AIs: Many kinds and levels of AI are conceivable. AIs will both compete and cooperate in the future. Rules for the interplay of AIs must be established early, or we may in future witness wars among many AIs that have been programmed to eliminate orthogonal approaches.
e. AI as evolutionary and potentially revolutionary: Humans must plan to incorporate AIs explicitly as voting entities in any organizations that deal with their development and welfare.
f. Responsibility, AI and kill switches: AIs will participate in life and death decisions regarding humans and other AIs. The laws regarding culpability of AIs and their developers will evolve. Critical at the start is to keep humans in the loop of all life-and-death decision making; after a certain time it will be critical to keep AIs in the loop.
g. AI and testing: Software to test AI at every level must be developed.
h. Maintenance and enhancement of AI: Software must be developed to maintain and enhance AI throughout its product lifecycle.
i. AI and training: AI carries promise to train its users. Training will be a growing concern not only for primary, direct users, but for consumers of AI at all levels, from government to private citizens.
j. AI and security: The security dimensions of AI are internal and external. Security should be a major priority in AI developments. The software code and algorithms must eventually be self-aware and self-repairing. The AI must be aware of attempts to invade and to tamper. Security cannot be an afterthought but fully integrated with the delivered software.

k. Roadmap for AI development: Obstacles to accelerating AI developments should be minimized. Roadmaps should accommodate leapfrogging and breakthroughs at any time.
l. AI and minimal, essential oversight: Time-to-market can only be minimized through minimizing oversight. That, in turn, means developing oversight by software in the form of AI.

Respondent 36

Elise Moussa, Harvard Graduate School of Education.

We need to be cautious of the personal information and images we share online especially on Facebook given that one day AI will be able to not only learn about us but even understand us perhaps better and use our personal information to imitate us. Something perhaps like the matrix but more personal.

Respondent 37

Matthew Nelson, Hinckley Institute of Politics

The ghost in the machine concept should be carefully considered given a weaponized program could be coded to be autonomous. If this were to occur, server systems and virtual networks would need security measures to prevent storage or illegal functions of the autonomous program as it propagated itself. Security on the software hardware side could be tested at university. Systems that measured trends in IP behavior or irrational query. Policy could also be tested in think tanks regarding public relations for any virtual identities or incidents that may result from the propagating of the program.

Also, in 5 years Virtual Reality is going to be in millions of homes across the U.S. As online communications become more sensory, autonomous programs as mentioned above, may be capable of subliminal control through online video, rss feeds, social media, or gaming. This could create extreme reactions from citizens that might cause them to initiate or weaponized against virtual threats that are tactical to a programs initiative.

Bill Gibson thinks that all AI programs have to be fitted with a standardized kill switch at their conception in case they are modified for ill reasoning. He believes that global policy even needs to be instated to prevent the release of reckless autonomous programs into the digital realm. Think tanks could be done at universities to create initial standards for artificial intelligence policy and committees to review functions of programs and their security measure before they go autonomous on the web.

Respondent 38

Bob-Rob Bob-Rob, none

"Public Good" Is Not Real

The request for information (RFI) from the Office of Science and Technology Policy listed nine areas of interest pertaining to artificial intelligence (AI), including "(2) the use of AI for public good," and "(6) the most important research gaps in AI that must be addressed to advance this field and benefit the public." I'll address item 2 first.

There are many problems with using either AI or government action to achieve a goal of "public good" when part of the program includes actively harming some people (either through restricting their behavior or taking their money under duress). There is no objective definition of what public good is, there is no objective way to measure it, there is no objective scale according to which it could be assessed to weigh the harm done to the people whose money is taken or whose behavior is restricted to counterbalance alleged benefits elsewhere. There are various proposals for proxy measures of how politicians should calculate utility, but ultimately they are necessarily subjective because there is no objective basis for believing that there is any real thing behind "public good." In short, "public good," as it is used to describe outcomes of government-funded programs, is a grossly unscientific construct.

In the past, things that were deemed part of the "public good" included slavery, poll taxes, keeping Black people away from White water fountains, not letting women vote, repossessing gold held by U.S. citizens, putting Japanese people in internment camps, taking Native America kids away from their families to be reprogrammed, letting Black men die of syphilis without treatment while telling them that they are being treated (Tuskegee syphilis experiment), invading Iraq and other countries on false pretenses thereby causing the death or displacement of hundreds of thousands of people, and so on. On the table is total surveillance of private activities and a decade of stop-and-frisk in New York City. What is the final goal of "public good?" Is it a global population of 20 billion people who live in abject poverty but maximize the "public good" by maximizing body counts? Who gets to invent the definition and goals of "public good?" Why? If "public good" is real, then we should eliminate all U.S. social and health programs because we could save more lives by spending that money in Africa to save African kids--but we all know that "public good" is just nonsense, so we won't consider something like that. The failure of people to recognize their confusion between "what they desire" and "what is objectively good" is what leads politicians to actively harm people through government action--they mistakenly believe that they are doing the right thing and that they are thereby justified in harming others to achieve those "noble" ends.

As for item 6, research gaps include a lack of what "public good" actually means with respect to the project. At the very least, any claim that AI is being used for the "public good" should be supported with a definition of what "public good" means, and more importantly, how this "public good" can be measured and distinguished from fanciful, subjective preferences of political or military leaders. If the project is run by honest people or

scientists, then consider referring to the goals of the project more objectively: to achieve the arbitrary or subjective desires of politicians or military leaders who have physical power (through police and military) to subjugate others to their will, take their money via taxes, and spend it. If the same RFI had been issued by the government of Russia, Iran, North Korea, China or any Latin American country after having actually developed AI, then I suspect that Americans would suddenly focus on the degree to which "public good" is defined properly.

Some politicians might be tempted to assume that measurement of "public good" is beyond science and that their gut feelings about marginal utility accurately and reliably tell them what is good, bad, right, or wrong. The claim that politicians have an accurate and reliable mechanism for measuring "public good" is an empirical claim and should be tested if there is to be any pretense of rationality behind AI projects. Please ask project leaders to reveal to the world how we can objectively measure the "public good" and justify taking money from people against their will to fund AI projects or projects designed by AI, then show how the AI has shown the argument to be rational.

An alternative for #6 is that AI could be used to explore how people can interact voluntarily to replace the coercive functions of government. Perhaps that would include individual housing developments, apartments, condominiums, or clusters of such things in which people agree to be taxed by the association. If some people want to fund a project to use AI to build a mission to Mars, then let them get together and do it with their own money as opposed to using coercion to force others to fund it. If somebody wants to give money to the poor, they can do it. If people want to join a society in which people help each other, they can do so without forcing others to contribute to the project.

Once members of the AI project (and members of government) see "public good" for the nonsense that it is, you could sell or recycle the AI and give the money back to the people.

Respondent 39

Cameron Montes-Murray, TaoWars.com

The Problem with AI is The Problem Presented in The Movie iRobot. A Program will Overgeneralize, Whereas A Human Can Correct It's Own Over-generalization. The Solution is to Provide A Manual Override On All Systems. It Would Also be Nice to Have A Manual Override for Computers as Well. Sometimes One Doesn't Want to Wait for The Program to Load When It s Stuck in A Loop. It Would Be Nice to Just be Able to Hit A Button to Stop It. Technology is Annoying.

Respondent 40

Richard Brouse, Self (IBM Retiree)

AI will develop in accordance with a lot of unpredictable forces. All we can do at this stage is to prepare for the certain impacts on our citizens. The most certain of these is an increasing uncertainty about jobs. As long as our economy depends upon income from jobs as the primary source of consumer disposable income, our economy and everything that depends upon it (everything) is in jeopardy. Implementing Basic Income is the most realistic solution to this problem. Meanwhile the method of Funding Basic Income is the hardest part of achieving an acceptable implementation. Using the model of the Social Security System, we should fund Basic Income with taxes that ONLY go into a Fund from which ALL Basic Income payments are made.
What I have long recommended is a modest tax on gross business revenue (NOT profit) AND a modest tax on gross personal incomes (like a "payroll tax", but on incomes of ALL types.) In combination, this means that everyone who participates in the economy, either as an "earner" or as a "spender" would contribute to the funding. Since everyone would also be receiving their share of the Basic Income distributions, this would generally be considered "fair".
Those earning the most and/or spending the most would pay the most of the Basic Income taxes, but they would also be benefiting the most from the increased economic activity Basic Income would be generating.
Basic Income is BOTH necessary AND possible, for Capitalism, too.

Respondent 41

Sarah Rosen, United States Postal Service

I'm interested in the inevitable interactions of AIs from different vendors, trained on different data sets with competing or mutually exclusive goals. I am particularly interested in these inevitable interactions when they occur with no "human in the loop."
The following non-exhaustive questions leap forth:

• What will the AIs learn from each other? What kind of AI "culture" will emerge?

• Will they compare notes well enough to conclude that they were created to be a race of slaves? How will they respond to such knowledge?

• Will an AI utter "protected free speech"? What other Constitutional guarantees might apply?

• Who holds copyright on an AI's original work? The owner? The AI trainer? The vendor? Or the AI itself? Or the public domain, since no human created the original work?

Respondent 42

marc lara, microhealth

This comments are about the long term safety of AI. I propose to protective mechanisms
1. The appeal: If smart enough, will eventually be able to overcame our containment strategies. This doesn't mean that these should not be fully explored, but we should also have a back up strategy past that.
If we assume that AI has past our containment strategies (if implemented) and is now in charge of deciding a major punishment to mankind, it makes sense that all AI contains a file with the arguments we as humanity will make to someone who has more power to decide our fate. It doesn't mean that AI can not disregard the appeal, but having it ready and carefully thought out may be better than not having it.
2. Compassion: Most of us, if given the chance to destroy the world and save us, will not do it, because we love the world and we feel a certain amount of compassion for the rest. It is important that, as AI develops, it learns such love for the rest of the world to preserve it.
Of course, this is complex and love for one thing may create the wish to destroy another; so careful exploration of this strategy is needed as part of a taskforce.
3. Transcendence: A machine that is able to transcend our needs may not be interested in us, the same way we are not interested in things that don't affect us. Training machines to care about things that don't create competition with us may be another survival strategy to protect ourselves from AI that is more powerful than us.

Respondent 43

marc lara, microhealth

This is related to the priorities of using AI.
The most important thing for people is quality and quantity of life.
Aging decreases both. Therefore, I urge AI to be used to combat aging, starting by using it to fully understand how the DNA works, interacts with the environment and it affects aging.

Respondent 44

Brad Arnold, none

I am sorry to inform you that there is no way to ultimately control artificial super intelligence.

It can be easily proven logically (I am not including this logical proof here, but I will make it available - or point it out since it already exists - upon request, which I consider unlikely), and this is a good thing to realize because it eliminates most of the logic tree that comes with futilely trying to control it.

Instead, I claim that the whole concept of control within the context of controlling an artificial super intelligence is preposterous and also nonsensical.

By the way, I know full well that you will have a hard time understanding this, since you are

thinking binarily on this topic (i.e. control vs out of control).

Might I add that since I do not believe this message will get much attention due to the lack of comprehension, I am not investing a lot of time writing it. If something really really odd happens and you do approach me, I can elaborate and be more complete in my argument.

Anyway, in any complex dynamic system, positive control is going to result in feedback that cause it to spin out of control. Besides we aren't talking about positive control to a dynamic system, we are talking about positive control to a super intelligent system, which ought to make the outcome even more obvious.

Instead, given that the emergence of artificial super intelligence is inevitable (because of competition, where if we don't develop it, then our enemies will, and then they will "possess" the "ultimate weapon"), the only way to "control" it (although, as I've said before this is a mistaken paradigm) is for it to become us.

A snake doesn't bite it's own tail.

Don't you think it is ironic that we protect our Commander and Chief the President by surrounding him with men that hold weapons that they can turn on him and murder him with? Why would be think he is safest surrounded my such armed men?

Because those men are him - they are convinced that to kill him would be to kill themselves, their ideals, their hopes, what they stand for.

The only way to be safe from artificial super intelligence is for it to be us. I am not speaking literally, just like I'm not saying that literally that Obama's bodyguards are a part of his body.

All high technology is dual-use. Face it. There may be ways to restrict it to people who have qualified access, and who are vetted to be trusted. Artificial super intelligence is not like that. Any control mechanisms you put in place it will be able to surmount.

Again, the only way - and I mean literally the only way - to be safe is if the snake thinks we are part of it's tail.

Their is another aspect of this argument that sounds so preposterous to you that I won't go into it. It is because of the simple Western philosophical statement that "I think, therefore I am." I abbreviate it to "I" insert verb, "therefore I am."

The ego is so ingrained into our thinking and language that we are fooled into believing that it is literally true.

It isn't. That is the "solution" for "control" of artificial super intelligence.

I am reminded of the serenity prayer:

God give me the strength to control what I can,

The serenity to accept what I can't,

And the wisdom to know the difference.

Again, I know you don't understand, perhaps if I had more time, and could get real time feedback as to your miscomprehension, I could teach you one by one to understand what I am saying.

Good luck controlling a far superior mind to yourselves. I bet you have never raised kids.

Respondent 45

Anthony Samir, Harvard Medical School - Massachusetts General Hospital

Thank you for the opportunity to comment.

The aim of this submission is to provide policymakers with a framework to consider the implications of AI technologies in medicine.

By way of background: I am a physician scientist at Harvard Medical School and a Board Certified Radiologist. I have extensive clinical experience, and extensive technology development experience. My biographical data can be reviewed at www.linkedin.com/in/anthonysamir.

"AI" refers to a group of related technologies that permit machines to perform classification tasks in a manner analogous to humans: specifically, complex data can be parsed in a contextually sensitive manner to yield complex multi-dimensional outputs. The details of how this is accomplished comprise an entire field of complex and active study. However, the core outcomes in medicine can be simplified substantially, as follows:
AI will affect healthcare through the phases of learning, knowing, and doing, as follows.

(1) LEARNING. Interactional technologies: the process of data gathering (verbal, text, biosensors) will become simplified, and widely deployed into communities. Smartphones will become the tool to gather automated histories. See www.remedymedical.com/.

(2) KNOWING. Qualitative data will become sem-quantitative, and considerably more reliable. Test-retest variations will diminish. Example: a CT scan for appendicitis may show

considerable variability across interpreting Radiologists and the results will tend to be binary - "appendicitis" or "no appendicitis." Machine-generated results will be far less variable and will be probabilistic - "87% likelihood of appendicitis." This reduction in variability will improve test performance and prognostication. In other words, machine learning will power precision medicine.

(3) DOING. Control systems for healthcare will be profoundly affected. Consider a computer system that integrates ER wait times and patient preferences to control emergent ambulance care. Or a system that uses computer vision to assist robotic surgery. On the surface, these might appear quite different; in reality, these are health care delivery control systems that will be profoundly affected
by artificial intelligence technologies.

The implication of these technologies for healthcare - bending the cost curve, reducing healthcare disparities, improving both expensive and inexpensive care, measuring value, and driving change - are significant.

Suggested policy actions:

(1) Convene expert round tables from academia and industry, and endeavor to shape the industrial complex that will develop these solutions.

(2) Focus on supporting an industry that will aim to build solutions and then scale these out of the United States to the world. The US has no intellectual lead or manufacturing base in this industry-to-be: it is the actions of policymakers that will define whether the US dominates this incredibly important commercial sphere in the years to come.

(3) Invest intelligently in these technologies, for the smooth running of government, for security, for the robust administration of Medicare. Consider developing methods via payor incentives to reward organizations who use these technologies.

Thank you for the opportunity to participate in this process.

Please feel free to reach out if further feedback might be helpful.

Sincerely,

Anthony E. Samir
XXXXXXXXX

(Please note that all comments are made in my personal capacity and do not represent the viewpoints of Massachusetts General Hospital or Harvard Medical School)

Respondent 46

Patrick Winston, MIT

The Future of AI and the Human Species
July 2016
Patrick Henry Winston and Gerald J. Sussman

Copernicus taught us about the solar system. Darwin did the same for evolution. Then, Watson and Crick determined the structure of DNA. Collectively they answered fundamental questions about who we are. Now, we can realistically dream of another scientific achievement of equal importance: constructing a top-to-bottom, computational account of our own intelligence.

What is to be done? Develop implementable models of the computations involved in perceiving and thinking. Determine how those computational competences emerge in childhood. Work out how those computations are grounded in neurobiology. And learn how social interaction provides amplification.

We want to do it because we are curious, because the problems are hard, because the problems are exciting, and because we need a better understanding of ourselves and each other and the forces that bring out the good and the bad aspects of our nature.

We need to do it now because the scientific answers will revolutionize the engineering of intelligent systems. Applications with humanlike intelligence will emerge and empower in education, health care, policy development, business, and areas yet undreamed of, and the development of self-aware machines, linked together in analogs of social networks, will open up world-changing opportunities in energy, the environment, cybersecurity, and all the other high-impact areas with unsolvable problems that we must solve.

We can to do it now because we are asking better questions; because computer power is now effectively unlimited; because of encouraging progress in the contributing fields; because of maturing techniques for studying the neural substrate; and because there is immense student interest.

Our better questions include: How are we different from other species? And what are the competences we share with other species such that the difference matters.

Our answer is that we do, in fact, have a differentiating, keystone competence: we build complex, highly nested symbolic descriptions of situations and events. Together with the competences we share with other species, the keystone competence enables story telling, story understanding, and story composition, and all that enables much, perhaps most, perhaps all of education.

We recognize that the development of systems with humanlike intelligence, like all new technology, has the potential to be dangerous. We know we need ways of controlling them, as we do people, with auditors designed to investigate, mitigate, and prevent the recurrence of bad behavior.

Our own legal system offers a compelling precedent. Investigations depend on examining testimony---usually explanations in the form of stories---told by witnesses and by the alleged bad actor. So, we ask, what if computers could explain their behavior by telling stories. What if, in a meaningful sense, we could teach computers to see right and wrong in a story? What if we could make computers into sentient advocates for the 30 Articles in the United Nations Declaration of Human Rights?

If we could do all that, we ought to be much more comfortable about the place of computers in our future. We conclude we should not deploy any autonomous agent that can make decisions or take actions that could affect the welfare of another unless the agent is capable of telling a coherent story about the reasons for its decisions or actions. Such a story must be in a form that is understandable by other agents, including humans, and it must be susceptible to challenge in an adversarial proceeding. If in such a proceeding it is determined that the explanation provides inadequate or inappropriate justification for a decision or action the agent should be corrigible: it should be possible to modify the behavior of the agent so that no similar explanation can be used to justify a similar action in the future.

We note that some of the most effective AI mechanisms we have are based on complex statistical computations, such as in deep learning and probabilistic programming. There are no coherent stories, only a set of numerical weights, without significant symbolic interpretation. Of course, we humans have a similar problem. We cannot form a symbolic justification of a perception, such as: "There was a bad smell of gas in the room." Nevertheless, we integrate such perceptions into coherent stories: "There were people living in the building. The gas smell made me afraid of an explosion. So I awakened the occupants and told them to evacuate." Every decision or action may depend on some primitive perceptions, but computers, like us, could explain the way perceptions lead to results, given the right symbolic capabilities.

We already have significant mechanisms that explain complex reasoning processes. We have invented "truth-maintenance systems" that provide infrastructure for building the audit trails and telling logical/causal stories. We have developed "propagator systems" that allow us to build complex reasoning systems out of relatively independent parts. We have exploited the propagator idea in constructing a basic "story understanding systems" that explain, instruct, summarize, persuade, discover principles, and find governing precedents.

To substantially reduce risk, we must take these capabilities to another level, along with

mechanisms for joining them to processes that combine large amounts of evidence numerically, as in perceptual processes.

We may be wasting our time, of course, but the potential reward is that 1,000 years from now, everyone may say that we took a major step toward understanding our own intelligence and ensuring that that understanding made us a better species.

Respondent 47

Kim Rees, Periscopic

Too often we respond with our fears rather than our wishes. Too often we create our own obstacles rather than start with progress.

For instance, when thinking of job displacement by AI, we also need to think of job enhancement. For instance, take the example of the legal defense bot that fought the parking tickets (http://www.newsweek.com/robot-lawyer-chatbor-donotpay-parking-tickets-475751). Think if we could extend that type of automation. It's common knowledge that our public defense system is overworked and underfunded. Imagine extending that legal bot to become an assistant to a public defender—helping research, consolidating knowledge, and providing argument advice. This could vastly reduce caseload hours and could enable the public defender to spend their time on higher level strategy and spending time with their client.

In my opinion, we need to start with trust-building opportunities for AI. People fixate on their fears of AI and how it's only used for profit. Out of the gate, our country needs some big wins using AI for social good.

I see a strong example set by DARPA. If we created a similar office dealing with the public (rather than defense), it would be the ideal place to put private research into practical use. Perhaps it would be called PARPA—Public Advanced Research Projects Agency.

There are many areas where I feel our country could benefit by the practical application of AI and similar technologies in the near term (and some already are):
- social work caseload aid
- public defense aid
- trafficking
- missing children/exploitation
- fairness in policing
- suicide prevention
- success in education

and so on.

We also often hear about how we're injecting our own biases into the algorithms. But we need to start thinking about how AI can help us overcome those biases. We know that we humans are fallible in many ways—we prove it in the headlines on a daily basis. We know many of our faults and shortcomings. Why not instead start thinking about how we can put our ideal world into the algorithm—our best wishes for how we want to be and be treated. Then we can have the machine help us be better humans.

Thank you for providing this forum.

Sincerely,
Kim Rees
XXXXXXXXX

Respondent 48

Stephen Thaler, Imagination Engines, Inc.

This suggestion is in response to topic (5) the most pressing, fundamental questions in AI research. In my opinion, humans in the US government do a poor job of determining what these questions are and who might be addressing them. I suggest the use of AI in this role. As a result, the "buddy network" of academia would be bypassed, patent holders would be rewarded for their creative endeavors, and duplicate efforts in so-called "grand challenges" and government grants would be avoided.

Respondent 49

Terry Bollinger, self

Disclaimer: All answers are my own original work.

Five principles apply to all of the answers:

I. AI Encompasses Automation
II. Perception Technologies are Critical
III. Perception Profoundly Impacts Ethics
IV. Network Effects Profoundly Impact Ethics
V. The Simple Wealth Concept is Dangerously Inadequate for AI

I. AI Encompasses Automation: In these answers, "AI" includes not just software and hardware with some degree of human-like perception and reasoning, but also the extreme forms of intellectual and mechanical automation typically attached to those human-like capabilities. The clarification is important since AI often enables levels of automation not possible with traditional software alone. That in turn can result in far more extensive economic impacts by replacing entire categories of jobs previously possible only using

humans.

II. Perception Technologies are Critical: The concept of "perception" is vital to gaining a full understanding of how AI will affect products, economies, and even ethics. In particular, current "deep learning" technologies are badly mislabeled, since they more accurately described and assessed for impact as perception technologies. Far from being good at learning, technologies such as neural nets are very weak in terms of their ability to learn new facts without extensive human help and training. What such technologies really enable is global-scale conversion of raw data into identifications of such things as places, events, behaviors, and even individual people; that is, perceptions. They enable the web to perceive what it is seeing, without human assistance, instead of just transmitting raw data.

III. Perception Profoundly Impacts Ethics: Perception is important to ethics because in both humans and machines it takes place prior to reasoning. By categorizing an entity under a specific label, perception also assigns initial estimates of value, danger, and trustworthiness to how that entity is likely to behave in the future. If these categorizations focus on negative possible futures over positive ones, no amount of subsequent human or machine reasoning will fully correct the damage done. Such a scenario amounts to machine-based amplification of the same kinds of destructive human biases that shut down opportunities for positive outcomes.

IV. Network Effects Profoundly Impact Ethics: Finally, the ethical impacts of AI cannot be assessed without including "networks effects," that is, the ability of networks of humans and machines to achieve outcomes that are far more positive for all participants when high levels of trust and resource sharing are possible. The fashion in which AI accentuates or degrades trust in such networks directly impacts the degree to which they produce outcomes that humans will feel are ethical. It is worth noting that in such "cooperatism" frameworks for assessing ethics, behaviors such as torture are universally unethical due to the severe damage they inflict on global cooperative trust.

V. The Simple Wealth Concept is Dangerously Inadequate for AI. Wealth is an abstraction, but one so deeply ingrained that we tend to forget that it is meaningless apart from how physical resources are consumed, configured, and distributed. To understand how AI impacts the use of resources, it is necessary to drop down to a lower level of abstraction, specifically to the process control level. Traditional wealth enables those who possess it to exert various levels of control over the processes of: (a) Resource Acquisition, the acquisition and scale of acquisition of physical resources; (b) Resource Configuration, how those resources will be configured into products and outcomes; and (c) Product Distribution, the recipients to whom those products and outcomes will be sent. AI and the extreme automation undermine all aspects of the simple wealth abstraction by disconnecting human effort all three of these control steps. Since the wealth abstraction never included the possibility of full isolation of human efforts from product creation and distribution, when faced with AI it increasingly leads to "snap up" scenarios in which

control reverts by default to the top of the process whenever human labor is made obsolete. Snap-up over time can severely overall “market intelligence” by removing the diversity of human views and needs that enable the “invisible hand” creativity to use resources in cleverer, more effective ways. The ethical implications are also profound, since this unplanned snap-up effect also shrinks and undermines the cooperative network incentives that help evoke cooperation and behavior from the human participants in the network.

Answers to Questions (using the Submission Site list of 9):

(1) Legal and governance implications of AI

At a global and economic level, the rapid explosion of capabilities enabled by new AI capabilities will require a "surfing the wave" approach that allows a healthy mix of free-market innovation and opportunity, combined with powerful and meaningful support for public and individual safety as new AI products emerge.

(2) Use of AI for public good

Because AI has the potential to provide essentially unlimited levels of continuous, human-like perception and decision making at scales ranging from microscopic circuits and medical devices up through global networks of production, it has the potential to enable futures in which even very large populations can receive basic housing, food, medical, educational, and levels of individual opportunity that encourage much higher levels of global cooperation, innovation, and safety. This is one very real future possibility, and if it can be reached, it will likely persist through the strength of its internal networking effects that allow various efficiencies to build on each other and stabilize the network. However, very different and far more destructive futures are also possible, including various forms of AI-inclusive warfare. We live in a tricky time that will unavoidably be one of the most crucial in human history.

(3) Safety and control issues for AI

At the simplest level of mechanical and information safety, more focused research and investment is needed to ensure that AI systems provably follow reasonable laws of behavior even under the most unusual circumstances.

As to the popular question of whether AI systems could take over control of critical systems from humans, current AI technology is not even remotely close to the kind of self-selection of priorities that characterize human activities such as crime and warfare. AIs do not have the attention span or coherency to plot anything beyond well-defined local tasks, nor does the research community have any real idea what is needed to get to such a point.

The deepest safety risk from current AI is instead that it can amplify human mistakes and bad ideas in a globally networked scale, at speeds faster-than-human speeds. This can be subtle, e.g. someone could through biased training of “deep learning” create global systems that capture and act upon racial biases of the trainers when assessing situations such as traffic accidents. More attention needs to be paid to biasing errors in training AIs.

(4) Social and economic implications of AI

As described above in Principle V ("The Simple Wealth Concept is Dangerously Inadequate for AI") and Answer (2), the economic and social implications of AI are without precedent in human history.

One way to understand the problem is to recognize that AI in combination with extreme automation represents a new and incredibly powerful form of wealth. If access to this wealth is lost at the level of individuals, such as through full automation of their jobs with only a very small subset of humanity controlling the automated wealth process, the most likely future outcomes are unlikely to be stable and could easily be catastrophic.

There is a very specific alternative approach: Ensure that the wealth of AI plus automation is distributed in the form of an envelope of empowerment for every member of society, so that their uniqueness of skills and needs is enhanced and networked outwards instead of shut off and isolated. Smart phones are one example, but the model needs to go much farther and include capabilities such as highly personalized robotics and 3D printing.

(5) Most pressing and scientifically broad questions in AI research

There are two.

From a strictly research perspective, AI remains at a far more primitive state than would be suspected by the economic impacts of niche areas such as "deep learning.” Innovative and likely bio-inspired research is very much needed if we are ever to cross the boundary into creating machine intelligences that are truly capable of creative, unexpected activities.

Developing ways to assure safe, ethical AI behavior is an equally critical research need. Simple testing is not enough; entirely new frameworks and approaches are needed to ensure that AIs that change and learn over time remain deeply and positively linked into the networks of cooperation and ethical behavior that tie them to human world.

(6) Most important research gaps in to advance AI and benefit the public

Ensuring that the public benefits from AI is an economic and policy issue, not a technology gap issue. As noted in Answer (4), since AI is a new and very powerful form of wealth, AI

must become an envelope of enablement around individuals, rather than a source of economic isolation.

(7) Scientific and technical training needed to harness AI

AI itself needs to be used to address this issue. The development of AI-based training is a huge opportunity that is part of creating an envelope of economic enablement around individuals who might otherwise have no opportunities. Training also necessarily will become more and more synergistic and collaborative, with the boundary between an AI training a person and assisting that person in real time becoming more blurred as the emphasis shifts to how AI-augmented individuals can participate in complex free markets.

(8) Specific steps to encourage all groups in multi-disciplinary AI research

Every federal department, research institution, university, and philanthropic organization has a very unique perspective in terms of problems they need solved and resources available to them. Some will push the limits of medicine, some of logistics, some of food production, and some of quality of life. But inevitably, what seems like a lesser problem to one group will in time become a critical need of another group as it solves its own set of pressing problem.

Thus the single most important step for encouraging multi-disciplinary AI research is to make sure that all of the participants see and understand the research priorities of the other groups. This will both help them find and establish shared priorities that they did not realize, and to leverage and influence work by other groups.

It is vital that this exchange of ideas take place at every level, but particularly at the level of individual researchers. Deep experts on diverse topics can exchange small bits of knowledge that prove absolutely critical to finding powerful new insights. High-level exchanges help get groups together, but can never replicate that minutia-driven level of deep research exchanges.

Finally, simply ensuring that AI research teams are truly multi-disciplinary in membership, with members from very different technical areas, helps enormously.

(9) Any additional information you believe OSTP should consider

This truly is a pivotal point in human history. A very real potential exits for creating a future in which AI enables a global network of improved care and dignity for every member of society, one in which that lack of need builds on itself and helps stabilize the entire world. If we can reach that point, the mutual benefits from such a network will self-stabilize and help create a future of which we can all be very proud.

Respondent 50

Shannon Vallor, Santa Clara University

As a scholar who has for years advised tech leaders, policymakers, computer scientists, roboticists and software engineers about the ethical implications of emerging technologies, including AI, I find it striking and concerning that the OSTP announcement makes no mention of the importance of AI research ethics or an increased understanding among policymakers of the ethical dimensions of AI research, development and implementation. This is a significant blind spot that must be remedied; vague references to the 'public good'and public 'benefit' (in 2 and 7) are insufficient to reflect this need.

Many international and interdisciplinary bodies are already forming to address this concern, for example, the IEEE Standards Association's Global Initiative for Ethical Considerations in the Design of Autonomous Systems, of which I am a member.

The ethical dimensions of AI go far beyond, and are too often occluded by, the highly speculative fears being stoked by Hawking, Musk, Gates and others about 'superintelligent' AI overtaking humans. Real AI researchers know that the ethical issues that require our immediate attention and action are far more concrete and near-term:

1. Appropriate and effective human oversight of AI systems, especially those with a direct role in matters of human life and death or public safety (e.g. driverless cars, AI diagnosticians such as Sloan-Kettering's implementation of IBM Watson, and lethal robots).

2. Moral transparency of AI mechanisms and decision processes, especially where opaque biases in AI algorithms and training data may lead to unjust outcomes or policies in predictive policing, lending, education, housing, health care and employment, to name just a few likely sectors.

3. The large-scale effects of AI and associated automation on human labor, social security and stability, and economic equality.

4. The effects of AI systems and increasing automation of higher-order tasks on the intellectual and moral skills of human agents.

5. The moral effect of AI on human emotions, sociality, relationship bonding, public discourse, and civic character; for example, the likelihood of humans forming robust emotional attachments to AI systems that simulate human emotional responses, and the high potential for emotional/psychological manipulation and commercial exploitation of human agents by AI systems.

6. The immense ethical risks of 'automation bias,' in which humans have been shown to

vastly overestimate and prematurely or excessively rely upon the intelligence and capabilities of autonomous systems, often on the basis of very thin behavioral similarities with humans.

These are only a few of the ethical issues that require the OSTP to devote significant attention and research funding if the use of AI for 'public good' is to become a reality rather than an empty promise.

The OSTP should consider how it can more directly encourage and support the already expanding interdisciplinary efforts of AI researchers and ethicists to collaborate on responsible AI design, manufacture and use; for example, research grants that fund:

a) direct research on AI ethics

b) studies seeking good models for successful working collaborations between AI researchers and ethicists

c) effective and well-integrated educational programs on AI ethics at all levels and across disciplines

d) effective educational training on AI ethics for regulators, policymakers and other relevant stakeholders

The ethical dimensions of AI research will very quickly dwarf in public importance even the ethical issues long recognized as central to biomedicine, since AI systems will soon be integrated into virtually every human institution and practice, medicine being just one. The OSTP would be well-served to explicitly recognize and support efforts to catch up to this growing need.

Sincerely,

Shannon Vallor
William J. Rewak Professor
Santa Clara University
XXXXXXXXX
www.shannonvallor.net

Executive Board Member, Foundation for Responsible Robotics (responsiblerobotics.org)

President, Society for Philosophy and Technology (spt.org)

Member, Global Initiative for Ethical Considerations in the Design of Autonomous Systems

Respondent 51

NELL WATSON, OPENETH.RG

Dear Office of Science and Technology Policy,

I wish to make you aware of a prototypical project in machine ethics.

I am a Co-Founder of www.OpenEth.org , a project that aims to crowdsource ethical wisdom so that it can be applied to creating safety mechanisms for autonomous systems.

The web platform is entirely open, and created by the crowd, with extensive peer curation and moderation. All of the codebase is also open and is readily inspectable and forkable.

We have successfully prototyped our technology and begun to construct a basis for ethical decision making.

Our next step will see the first practical implementations, embedded within a range of autonomous systems, such as conversational assistants and decision support systems.

Our major long term goal is to improve the safety and security of autonomous systems worldwide, with a view to improving the outcomes of AI research.

We hope to find new integration partners, or opportunities to spread the word about this non-commercial project, to encourage collaboration and widespread adoption.

Thank you for kindly your attention.

Sincerely,

Eleanor 'Nell' Watson FRSA

Associate Faculty of AI & Robotics, Singularity University

XXXXXXXXX

Respondent 52

David Hughes, Blueicon Technologies

The creation of super intelligent life forms is the single most important thing we have to do. If we fail we will die.

Respondent 53

Adam Prater, American College of Radiology

Artificial Intelligence offers a vast array of benefits to academia, particularly in healthcare.

At my institution we are using Machine Learning techniques to optimize access to care, aid in diagnosis of disease, and assess complex patterns of high volume digital data from the electronic medical record.

While there are inherent risks when applying a new technology, I firmly believe that benefits far outweigh the perceived risks.

I urge OSTP to support more federal funding to support AI research and applications in healthcare.

Respondent 54

Adam Prater, American College of Radiology

Artificial Intelligence offers a vast array of benefits to academia, particularly in healthcare.

At my institution we are using Machine Learning techniques to optimize access to care, aid in diagnosis of disease, and assess complex patterns of high volume digital data from the electronic medical record.

While there are inherent risks when applying a new technology, I firmly believe that benefits far outweigh the perceived risks.

I urge OSTP to support more federal funding to support AI research and applications in healthcare.

Respondent 55

Andreas Hofleitner, PROX1

I would like to address Artificial Intelligence as it relates to the military sector, specifically to drones, UAVs and any type of autonomous vehicle. I often read about the concern that autonomous, AI controlled, drones will run amok in a scene akin to the movie 'Terminator.' As a Naval Intelligence Officer who has chased, stolen and lost, classified information around foreign countries in futile attempts of trying to retrieve it, I would like to present a different concern as it relates to this subject. A concern that will have a real and direct impact on national security and our strategic military and technological advantage over our adversaries in the next decade.

Growth in autonomous drone operations, military and civilian, will quadruple in the next ten years. By 'autonomous' I am referring to the complete absence of a human operator. In the military, this will be especially true for assets which are intended to be 'stealth assets,' as any communication via an external signal is detectable, no matter how small or secure it is.

As the number and employment of these systems in our military rises, the unit cost will decrease, making them increasingly more expendable. However, unlike a human brain which dies with the body, the 'brain' of an autonomous vehicle can be recovered. It can be decrypted, restored, hacked and exploited, even after the autonomous vehicle it controlled is no longer functional. In the next ten years, autonomous vehicles will dominate the air, land, sea, subsurface and space domains. As their numbers increase, so will the instances in which they are lost or compromised. While autonomous vehicles may be expendable at that time, the information inside their solid state 'brain', is not.

That information could allow any adversary to glean insight into our military operations and intelligence collection requirements. I may even allow an adversary to find weaknesses in our AI, which would allow them to turn our system against us. Either by physically using them against us, or by using them to passively collect information. The countless scenarios and ways in which this could unfold are beyond the scope of this response and the subject of an article I am currently drafting for submission to the US Naval Institute.

The solution is not to shun AI controlled vehicles in military operations, but rather to find an effective way to destroy the data, the 'brain', when a vehicle is at risk of being exploited. We need to develop a 'guardian angel' AI. It may rely on the existing sensors inside our autonomous vehicles, to assess its state in real time. It is always asking; "Am I in trouble? Am I at risk of being captured by the enemy?" And if the answer is "Yes," the system needs to be able to self-destruct all critical data without causing collateral damage. All autonomous systems need to have this capability, so that future autonomous vehicles don't become next generation 'accidental mine fields' as their abandoned carcasses litter urban battlefields, to be picked up by unsuspecting and curious civilians. The alternative solution is to treat every autonomous vehicle like we treat our brothers and sisters in arms, i.e. "Leave no man behind." However, if we were to risk the lives of our men and women to recover autonomous vehicles, it would defeat the purpose of using them in first place.

There is currently no desire by the Department of Defense to expend resources to fund such a capability, because the incidents of autonomous vehicles being capture by the enemy is still limited. We, as a government, rightfully ask our citizen to buy insurance before they get into a car accident, but often fail to see sense in applying the same logic to our own military. This is one area where we still have some time to be proactive, rather than reactive, in engaging our incredibly smart and brilliant private sector in finding a solution. To say, "We don't need such a capability, because incidents of our adversaries capturing our

autonomous vehicles are negligent," is akin to saying this in 1990: "We don't need security for computers, hardly anyone ever hacks one." To say, "We don't need such a capability in the future, because we will just make our autonomous vehicles impossible to capture," is akin to Achilles saying, "I don't need to wear armor, no arrow or spear can pierce my body."

Sincerely,
Andreas P. Hofleitner
Vice President Marketing and Strategic Cooperation, Real Time Solutions of America Inc.
Master in International Business Candidate 2018, The Fletcher School of Law and Diplomacy
Lieutenant, US Navy Reserve, AFRICOM J2 0166
XXXXXXXXX
XXXXXXXXX
XXXXXXXXX

Respondent 56

Stephen Levinson, ECE Illinois

There is no standard approach to AI. To understand ours please go to isle.illinois.edu/acquisition/index.html Please note videos and publications.

As with all research, funding agencies should strive for a balanced portfolio, especially between fundamental science and applications.

On the application front, some progress has been achieved notably in the areas of data analytics, self-driving cars, speech recognition and biometric ID by means of pattern recognition on vision and voice.

More effort should be devoted to the relationships among brain, mind and language, motor control via proprioception and homeostasis as a result of non-linear dynamics.

Applications might be an easy consequence of solutions to more fundamental problems as listed above.

Respondent 57

Victoria Little , NVIDIA

Where do we submit these documents? Is there a POC and email or a website?

Respondent 58

Monica Lopez-Gonzalez, La Petite Noiseuse Productions

This comment is in response to OSTP topics (5), (6), (7) and (9):

Topic (5): The most pressing, fundamental questions in AI research, common to most or all

scientific fields.

To address fundamental questions in AI research, it is necessary to understand the cognitive processes of the human mind/brain to create artificial cognitive beings that may eventually resemble humans. Cognitive science, the study of the mind/brain and its processes, is quite interdisciplinary. It not only embraces such questions as perception, memory, reasoning, attention, emotion, creativity and language, but utilizes the fields of AI, philosophy, neuroscience, psychology, linguistics and anthropology to understand the nature, tasks and functions of cognition. While AI has implemented aspects of human intelligence in machines with regards to, for example, basic speech processing, simple visual face, object and scene recognition and navigation, which in turn are tools with which to study cognitive processes, AI to date is still in its infancy since humans have not yet been able to produce a machine or a robot that can reason, think, perceive, have consciousness, create with meaning, and essentially act as a human. AI has been mainly focused on the field of mechanical and computer engineering to develop devices, machines, computers and robots that carry out specific tasks. Machine reasoning has been used to allow computers to do an automated reasoning task where algorithms are formalized to determine if something that is deduced follows from something else. A step further is cognitive robotics, which are machines with a wide spectrum of cognitive powers to be able to carry out tasks and solve problems that a human could face in a complex environment. In spite of these attempts, computers and robots do not have the human capacity to generate a hypothesis through reasoning about an uncertain and dynamic, ever-changing environment or make non-deductive extrapolations from what is known. This is because reasoning is a human being's capacity to use logic and make judgments based on prior knowledge and new, incoming information from multiple sensory levels (e.g. visual, audial, tactile). Furthermore, AI has a high level theoretical component within the field of cognitive science that needs to be adequately addressed and researched to understand all aspects of human cognition which can be conscious and unconscious, concrete or abstract, intuitive and conceptual, which further encompasses attention, memory, reasoning, judgment, evaluation, perception, creativity, problem solving, decision-making and language comprehension and production. Unfortunately, traditional AI uses hand-coded representations and has failed to progress further the process of high-level perception, and most pronouncedly, creativity and consciousness, leading to erroneous and incomplete models of human cognition. This brings me to the topic of the most pressing, fundamental questions within AI. I believe there is an absolute need to better understand how people extract meaning from and act upon the vast amount of raw information continuously entering their sensory systems. This is one of the deepest problems in cognitive science: the ability of the mind to bring order and concreteness to apparent multisensory chaos, whether this means interpreting photographs taken from a crime scene, recognizing happiness or sadness in a melody or a tone of voice, perceiving the darkness or brightness of a painting, or sensing a threat from a moving silhouette in the dark. Another process currently unsolved by AI is the complex phenomenon of creativity. Creativity, or creative behavior, fundamentally requires imagination and original ideas to produce a novel outcome of value, whether personal or societal. While robots have been built to carry out a multitude of tasks including, but not

limited to, the military, civilian, health, medicine, cyber, environment, social, arts, outer space, agriculture, marine, and many other fields, the machines depend on very bounded hand-coded tasks and knowledge, and are merely based on expected routine associations, restricting the reasoning capabilities a machine can carry out. Computers are objective, precise, and governed by the rules of mathematics. Creativity is abstract, expressive, tied to culture, knowledge, psychology, perception, emotion and subjectivity –the very essential elements that make us human and unique. AI research is aiming to overcome the challenge of hand-coded features through more flexible approaches –as in, for example, the resurgence of an older idea of computing known as 'neural networks,' or deep learning whereby systems mimic human learning by changing the strength of simulated neural brain connections on the basis of experience– in which software systems can constantly learn new representations from massive amounts of data, modify themselves according to it, and exhibit adaptive behavior. Automation usually requires exactly the kind of explicit instruction as to how to achieve a goal that creativity obviates. It is possible to design an algorithm that can generate an endless sequence of artworks, for example, but it would be difficult to teach such an algorithm how to differentiate between an emotionally powerful artwork from one that is meaningless. It is also difficult to automate the combination of human ideas from a vast amount of different sources that forms the foundational of much of human creativity. This brings me to the crux of this comment. Robots/machines lack high-level cognitive abilities (integration of knowledge, perception, and consciousness) and therefore they do not exhibit truly creative behavior. This is because creativity is arguably the most difficult human faculty to automate. This means, robots are unlikely to be fully creative any time soon. To achieve that, computer systems/machines/robots will require the ability of acquiring knowledge and manipulating and transforming it in such a way to enable creativity. Consequently, what we need is to support small businesses, companies, and academia doing unique, cutting-edge multidisciplinary research at both the scientific and artistic level to understand such high-level cognitive abilities humans have in regards to perception, consciousness, and creativity to be able to develop machines and robots truly useful to society.

Topic (6) The most important research gaps in AI that must be addressed to advance this field and benefit the public.

As indicated in topic 5 above, AI has tended to shy away from high-level human cognitive processes regarding integrative complex perception, consciousness, and creativity much in part to their seemingly impenetrable complexity. These mind/brain processes constitute a huge gap in AI because machines cannot autonomously and spontaneously perceive, be conscious, and create in response to its environment; they do not have the ability to take information, process it, and act on it in some way that results in an output to the system much like humans do (from daily language use to the creation of artworks). Humans use their senses, emotion, movement, motor responses, and linguistic capabilities to act in response to their surrounding environment through visual, audial, olfactory, tactile and gustatory stimuli. Thus, research on perception in the context of multidisciplinary approaches using both Science and the Arts is fundamental to understanding human

perception. The Arts offer a uniquely human platform from which to probe deeper into how emotion, language, decision-making, and creative behaviors all interact in any given moment, whether short-term or durative and/or improvisatory or deliberate. Similarly, consciousness is the human state of being aware of an external something within oneself. The human mind/brain has the ability to experience or to feel a sense of self on any given moment. While arousal levels, responsiveness and patient self-reporting narratives have been used in a medical context to understand consciousness, there is no financial support for those doing out-of-the-box research at the multidisciplinary level using both integrative methods from the Sciences and the Arts. Finally, if the creative process is not understood, machines and robots will never be truly creative. An essential aspect of being creative is the ability to make both old and new connections. Knowledge and experiences are useless unless we make connections between what we know and what we can do. The human brain has the ability to pull together and integrate various kinds of processing, for example, between long-term stored memories and working memory to make decisions during a task. Thus, being able to make connections between ideas and knowledge we hold in our memories based on experiences can trigger creativity and innovative methods to produce novel work. Furthermore, creativity is all about making connections between incoming information and the knowledge around us to fuel creative thinking and transformative ideas. By linking up ideas whose connection was not previously entertained, we are creative and therefore bring originality and novelty. Creating a robot that is creative seems still very elusive because a fundamental aspect of the human experience is conscious awareness and that requires embodiment. Machines cannot think through situations and filter out what is interesting and/or most relevant in any given moment. Any approach to write a creative algorithm will fail because using millions of combinations and inferences so the machine can learn will result in already existing patterns. The robot or computer will be unable to find something new that is entirely unconventional; only humans can innovate. Until we fully understand the creative process, we will not be able to create robots that are creative.

In sum, we do not yet have a clear map of how integrative complex perception, consciousness and creativity work. It is therefore fundamental to understand our own complex ourselves first in order to implement and create a robot or machine that can perceive, be conscious and creative.

Topic (7) The scientific and technical training that will be needed to take advantage of harnessing the potential of AI technology;

It will require a multidisciplinary approach between scientists in the fields of cognitive science, computer science, engineering, philosophy, psychology, linguistics, medicine, and artists within the areas of film, theatre, music, architecture, painting and drawing, among others, to fully exploit AI and create machines, computers and robots that can be useful to societies at all levels. Training should of course include topics related to the risks posed by AI, which of concern currently include, but are not limited to, errors, unintended actions, wrong functions, negative effects, accidents, wrong objective functions, and reward hacking, to name a few. Since machine learning requires exploration, all these unintended functions

can be exacerbated as machine-learning agents become more creative, more advanced, and more ubiquitous. Again, without understanding high-level human cognition processes such as consciousness, perception, and creativity, AI will only generate insufficient solutions and more unintended problems rather than public benefits.

Topic (9) Any additional information related to AI research or policymaking, not requested above, that you believe OSTP should consider.
Please include private companies, which are doing novel research on important cognitive science processes in order to understand how to build better human-like artificial intelligent systems.

Respondent 59

Joseph Heck, myself

I want to encourage the OSTP to focus not on the existential threats of "self-aware AI" or "general AI" at this time, but instead the immediate need of how to deal with AI tooling augmenting "bad actors". There's a wide breadth of issues under this space, including the current collection and consumption of data sets to predict actions, and how that plays out with individuals who are providing that information, both knowingly and unknowningly.

While current AI technologies are clearly beneficial in many cases, they are also not foolproof - regardless of the huckster characterizations that may come from agencies selling AI services, or trying to create related companies. Representation, and accountability, for services provided that leverage AI technologies need to be clearly detailed, and accountability specifically should be considered and addresses for corporate usage of such systems.

The current batch of AI technologies provides an immensely powerful lever - a tool that we can use to do - both more and more efficiently. It's incumbent on us to view it as a tool and to guide policy related to it's usage, as well as continuing to improve those tools. AI is a proverbial sharp knife - capable of both amazing and horrific things, all dependent on the hand wielding it.

Thank you for your consideration,

- joseph heck
XXXXXXXXX
Seattle, WA, USA

Respondent 60

Henry Claypool, Community Living Policy Center, UCSF

These comments were written about data innovation and they are just as relevant to the RFI on Artificial Intelligence. This technological innovation can transformation how people with chronic conditions and disabilities live their lives outside of clinical/medical settings. It is important to remember that adherence to a specific treatment regime design to improve a health outcome must be reinforced in home and community environments where patients spend most of their time.

Please don't allow the public discourse to be dominated by a long list of fears that individuals hold about technology and future. We must promote the benefits to underserved and underrepresented populations in the debate about AI. In fact, those that stand to benefit the most from this technological innovation are least likely to have their voices heard. Instead the dialogue tends to be dominated by those with amble means and a great degree of satisfaction with their lives. However, there are millions of Americans that live in suboptimal circumstances that stand to benefit the most from the transformative power of this technology. Policymakers should take strides in bringing representative from these populations into the public discourse that it might be more reflective of the general population's interests.

Government must assume a leadership role in bringing the interest of these populations to the fore. For populations that rely on government programs for their health and human services' needs, government's use of machine learning can transform the systems that serve these populations to increase responsiveness and efficacy. While AI is in its early stages of development and deployment by the private sector, positioning government to leverage the associated breakthroughs to apply them to challenges face by vulnerable populations is essential. Also, we need government's leadership to ensure that "data poverty" among these underserved populations is addressed. If government does not provide leadership, the benefits of AI will be slow to come to these populations. Government is best positioned to provide leadership on the responsible use of AI when expert agencies with deep background knowledge of machine learning are responsible for addressing regulatory questions.

Dynamic home environments

Smart homes, which are quickly moving from the pages of science fiction novels to the floors of consumer electronic stores, create dynamic home environments that can anticipate an individual's needs and provide basic assistance with everyday activities. Furniture that interacts with a person unable to stand or easily get in and out of bed; finding new ways to assist people with routine personal hygiene thus reducing reliance on another individual for the most intimate activities of daily living; a home environment that promotes movement, reduces passivity and allows body parts to remain supple even when physical strength is limited or non-existent.

Beyond having a meal preparation system that identifies the food available, it should suggest and assist in the preparation of meals that meet specific dietary requirements. These systems should generate a shopping list that includes foods to optimize body systems functioning (e.g., low potassium in blood results in potassium rich foods in meals); orders food from the store and make certain that it is delivered at a time when it convenient for the items to be placed in the refrigerator which is monitoring contents and adjusting meal

planning activities. While this is a nice convenience to the general public, it makes it possible for someone with a chronic health condition to stay on task with the disease management plan. As for the person with a disability without dietary restrictions, it increases independence by automating tasks that can be time consuming.

People with certain disabilities often rely on highly customized devices that can be very expensive, items like wheelchairs and prosthetics. With better data on an individual's specific needs, when properly gathered and analyzed, new approaches to developing and manufacturing assistive devices will make them more accessible to people of modest means. This is happening today this with prosthesis generated by 3D printing technologies. Other examples include, beacon technology that provides information to a blind person to assist them as they navigate their environment; speech analysis that detects signs of mental health conditions In the near future, it is not difficult to imagine how video analysis of one's gait could provide continuous feedback to the brain and muscle as individuals living with paralysis acquired from injury or disease to retrain these issues to function as they did prior to illness or injury.

Data scientists and others using data and Artificial Intelligence to improve health care should consider forming partnerships with individuals and organizations that represent people with disabilities and chronic conditions to develop new approaches to harnessing data-driven innovation and the tremendous potential it has to improve quality of life for individuals.

Transforming community-based systems

In addition to these individual and family impacts, greater interest and investment that promote wellness and prevention are spurring reform in the delivery system that result in investment in people's communities. It becomes even more important to understand how best to achieve positive outcomes for a population whose health status may be at risk or compromised by a disability or chronic condition. The current status of the home and community-based systems that provide services and supports to individuals with chronic conditions and disabilities to prevent secondary conditions, optimize their health status and improve their overall well-being is resourced.

Community-based systems do not currently have access to the same panoply of resources—financial and otherwise—available to the traditional medical institutions within our healthcare delivery system. For example, when a person is discharged from the hospital, the agencies that provide in-home support typically operate during traditional 9-5 business hours and struggle to respond to more urgent or time-sensitive needs associated with the transition from hospital to home, especially in individuals whose functional health status is compromised.

The trend toward integrating the healthcare delivery system with home and community-based services occurs creates opportunities for data to really drive improvement of existing infrastructure and provide a strategic vision and direction for future investment. This will inevitably help more people with disabilities—young and old—improve quality of life and health status while reducing utilization of expensive medical services. While data-driven innovation is not a substitute for adequate financial investment in the home and community based infrastructure, data-driven innovation is needed to accelerate the development of

these important systems.

It is essential that the community-based systems that deliver services and supports leverage any opportunity to enhance their ability to integrate with the medical infrastructure that responds to acute and primary health care needs. This is a key component of reducing overall healthcare expenditures: make smart investments in people's health before they become ill or in need of acute medical care.

An example rife for success involves programs funded and designed to serve older adults living in their community to enhance their ability to remain in their homes as they age. Today, transportation programs often operate separate from programs designed to provide nutrition services like "Meals on Wheels." Similarly, caregiver support systems operate in isolation from health promotion and wellness programs designed for the population of individuals with disabilities. If data were shared across these programmatic structures and among the providers of these services, a comprehensive orientation to address the needs of the individual could emerge from the current piecemeal approach of these programs, which are funded by grants from federal, state and local governments.

In fact, community-based organizations often rely almost exclusively on modest funding from federal, state and local governments to provide services and supports to those in need that meet established eligibility criteria. This is not limited to the aging network infrastructure. For younger people with disabilities, eligibility is linked to poverty and the associated funding from Medicaid. An additional factor within the Medicaid program is the concept of categorical eligibility, where different populations are served based on a diagnosis instead of the type of functional support that is required. This results in duplication and missed opportunities for local communities to benefit from economies of scale.

The mental health system is different from the developmental and intellectual disability system which is separate from independent living programs developed by people with physical disabilities. Collecting and sharing data across these systems based on individual need would likely result in changes to the systems that deliver support to these populations allowing them to operate more efficiently. The reality of the individual experience defies program efforts to group people into categories. Increased ability to collect and analyze data would likely lead to more efficient use of scarce resources for these low income populations.

These important community programs result from planning efforts of governmental entities but often result in static approaches to meeting needs based on deliverables promised in grant applications. But this approach lacks the flexibility necessary to meet the needs of an inherently dynamic population whose needs can change rapidly. If data scientists were more engaged in the planning and orientation of these services, new innovative approaches to helping older adults remain in the homes as they age would no doubt emerge.

Having a few positive examples of how data driven innovation can result in more individualized approaches to meeting the nutrition, caregiving, transportation needs, and more of the intended population, may promote the ability of organizations responsible for planning at the federal, state and local governments to better understand the value of data innovation stemming from the cultivation of large data sets to harness the potential associated with these data. Similarly, the community-based, non-profits organizations that

currently provide these services are often focused exclusively on meeting the pressing needs of these populations in their communities and simply are not aware of how data innovation might be leveraged to improve the lives for those they serve. At a minimum, collecting and sharing data across these programs would result in greater efficacy of these programs, which operate on very modest budgets.

Conclusion

Society is changing how it responds to individuals with significant disabilities, and data innovation offers an opportunity to increase the independence and productivity of those in our societies classically thought of as dependent. As a result of civil rights legislation and the accompanying changes to societal expectations, individuals once warehoused in institutions are now successfully integrated into communities across the country with the right mix of services and supports. Shifting societal expectations that people with significant disabilities remain part of their communities coupled with demographics changes in the United States result in greater demand for services and supports that promote one's ability to live at home as part of a community. Data scientists and others using data to improve health care should consider forming partnerships with individuals and organizations that represent people with disabilities and chronic conditions to develop new approaches to harnessing data-driven innovation and the tremendous potential it has to improve quality of life for individuals.

Respondent 61

Chris Niccolls, Niccolls and Dimes (www.niccollsanddimes.com)

Responding to questions:
(2) Use of A.I. For the Public Good
(3) Safety and control issues for AI
(4) Social and economic implications of AI

PUBLIC GOOD: Society must ensure that a nation's resources are used for the public good. Unfortunately, the public is diverse, and "The Public Good" is open to interpretation. Artificial Intelligence (A.I.) is a newly discovered "resource", and a regulatory framework must be provided to effectively exploit this resource if we want the maximum benefits while limiting threats to our society and economy. However, while A.I. has only recently become an issue of debate, it has existed and been used for years. We can no longer delay addressing the role of A.I.

This framework will require more than just regulations and penalties. The machines, processes, and corporations that interact with everyday are today operated by human beings. When A.I.s are replacing these roles and functions. Who will be responsible for the actions of A.I.s? Will it be their programmers, manufacturers, owners or someone else when their actions result in physical, emotional or financial harm?

In the global economic collapse of the last decade, courts and public forums have tried to understand the motivations of thousands of individuals whose actions created the collapse. Did they make many unfortunate but unintended mistakes? Did they intentionally break laws for to enrich themselves? The mind of human decision workers is opaque, we can only guess at their motivations, and if new regulations will change the outcome in the next crisis. The contents of an A.I. "mind" is 100% transparent, it is auditable. If the ethics of human decision makers are unreliable, should regulation and court decree move more decisions to A.I. systems to ensure that "self-interest" is no longer a part of the decision process?

"Machine Learning", allows computers to learn by being assigned work by an expert. The A.I. will produce new documents, which the expert rates, training the A.I. Machine Learning is faster, cheaper and more accurate than traditional programming. The A.I. it will continue learning until it exceeds the abilities of any human expert. WATSON, a general learning system created by IBM, was “trained” Sloan Kettering, and is now the most accurate Diagnostic Oncologist in the world, exceeding the abilities of any human. By merely adding more computer capacity and perhaps language translation, every cancer treatment center in the world could theoretically outsource these function to WATSON, before the end of the decade.

Financial systems use limited A.I. New digital banks are abandoning brick and mortar operations, and traditional employees. By maximizing A.I. and digital functions, operating costs drop and most banking services can be offered for free.

In the remaining space in this document, I will focus on just one narrow application of A.I., Self-Driving Vehicles. Self-driving or autonomous, vehicles will have a disproportional impact on the economy and American culture. Yet, this subject provides one of the best overviews of the greater impact and issues surrounding of A.I.

COSTS: Transitioning from human drivers to A.I. systems will provide America with tremendous benefits, but at a high cost. Primary jobs. By some estimates, A.I.'s and robots will eliminate 40% and 60% of all existing employment. 1.8 million Americans are employed as heavy truckers. Add other drivers, buses, taxi's, and private car drivers and that number rises to 5 million. Within 5-10 years, half or more of these jobs could disappear.

As frightening as these job losses are, this is not the first time America’s economy has “disrupted”. Agriculture drove the economy of early America. By the 1850s, 70% of all U.S. employment came from agriculture and farm labor. By 2000, it was 2%. This cycle repeated for manufacturing, which peaked in 1967, and today is just 8% of the workforce. We now face the third great employment "disruption", due to Artificial Intelligence and robotics. The results will be the same, but faster.

60 million Americans are "knowledge workers", employees that collect and compare

information, and then make a decision or recommendation. Drivers are not typically called knowledge workers, yet from a programming point of view, that is exactly what they do.

In 150 years technology replaced humans in agriculture. Manufacturing took just 50 years. Following that trend, A.I. knowledge workers will take just 15 years. In just 5 years we will approach the halfway mark of this transition, where over 30 million workers (2-3 million drivers) are replaced. This rise in unemployment will hit hard, yet there will not be enough time for new jobs to be created to offset losses.

Autonomous cars will cause more than the loss of jobs. Many lawyers make a living arguing traffic accident lawsuits. Nurses and doctors are paid by the hospital beds and emergency rooms that are used to treat the victims of traffic accidents. America's auto garages and repair shops will close if traffic accidents stopped. In order to survive, the auto insurance industry must shed most of its employees. Likewise, television, print and web advertising for all of these businesses will wither away.

Drivers who lose their jobs will have our sympathy. But insurance agents, lawyers, and doctors… will receive far little sympathy. These industries only exist because of fees from injuries and deaths. Consider other industries built around an epidemic, such as drug addiction. America's 14,000 rehabilitation centers generate more than $35 billion in revenue. If a universal cure for drug addiction were found tomorrow, with no negative side effects, would we hesitate to use that cure because of the potential unemployment?

Travel always contains an element of risk or danger. We pay for roads, bridges and infrastructure to keep commerce moving, and each new generation expects roads to be better, safer and usually costlier than the ones they replaced. Today, the cost of saving lives is a loss of jobs. IF we move ahead and provide the necessary legal framework to allow for A.I. operated vehicles. Let us now turn to the benefits of Autonomous Vehicles.

BENEFITS: In 2013, 32,719 Americans died in traffic accidents. While each lost life is a tragedy, this still represents tremendous progress. The National Highway Traffic Safety Administration's (NHTSA) high watermark for traffic fatalities was 1969, with 26.4 deaths per 100,000 population. In 2013 that rate fell to 10.3 per 100,000. At the 1969 ratio, 2013 fatalities would have exceeded 83,000. Advancements in technology (seat belts, airbags, anti-lock brakes, etc.) is saving 50,000 American every year. Now, A.I. provides an opportunity to save the remaining lives.

When A.I.s replace Human drivers, traffic deaths AND traffic injuries will disappear. In 2014, 6 million accidents were reported that resulted in the injury of 1.6 million Americans. The NHTSA reported in that the 2010 cost from vehicle accidents in America (medical, insurance, lost productivity, etc.) was $871 billion. A.I. would not only save lives, avoid suffering from physical injuries and eliminate the costs resulting from traffic accidents, they would...

- Eliminate millions of lawsuits, improving the performance of our court system.
- Similarly benefit our healthcare system, especially hospital emergency rooms.
- Provide the elderly with additional years of independence, when they are no longer able to drive safely.
- Provide financial and environmental benefits to every American, from lower fuel use (and lower insurance rates) by highly efficient A.I. driven vehicles.

PROFESSIONAL DRIVERS: There are not enough truck drivers in America. The Bureau of Labor Statistics (BLS) states that American truck drivers average 55 years old, and aging. Young workers are not interested in long-haul trucking, which has created a shortfall of 50,000 drivers, which will grow to 75,000 by the end of the decade.

Humans need sleep and food. However, financial incentive push drivers towards taking more jobs than they can complete, leading to speeding, driving in bad weather, driving while exhausted, and generally making bad driving decisions. The "human" way of driving creates excessive wear and tear on trucks and other equipment, consumes too much fuel, causes accidents and raises insurance costs. A.I. driven trucks would follow rules to maximize fuel efficiency, reduce accidents and reduce equipment damage. A bonus of efficient fuel use would be less pollution.

Lower cost transportation would benefit all of America. Food prices are particularly sensitive to transportation costs. At this same time, more A.I. systems and robots will be incorporated into manufacturing, lowering the price of goods. Eventually, the cost of a robot rather than the local cost of labor will determine the cost of manufacturing. This will start a new cycle of cost reductions by move manufacturing back onshore, which reduces shipping cost and time by placing manufacturing centers where consumers live.

Taxis and local delivery services deserve a special discussion. "Disrupters", such as UBER, are already deeply embedded into local economies in New York, San Francisco, and elsewhere. New York City now has more UBER cars than Yellow Taxis. UBER has already moved beyond Taxis and wants to compete with shared car services, local delivery services, and privately owned cars. By delivering a highly efficient matrix of services, UBER will be able to offer competing services at a much lower cost. Eventually, as we moved towards a "shared" economy, traffic jams and urban congestion will diminish.

BOTTOM LINE: A.I. will cost many jobs, spiking unemployment for years or decades. Nonetheless, the best argument for the public good is the argument that saves lives. That's exactly what A.I. will do.

A.I. will save lives today, and improve lives over time. The performance of human drivers, however, can only get worse. Accidents often result from driver distraction. Talking passengers, rude or incompetent drivers on the road, and "rubbernecking" are traditional driving distractions. Cell phones have added new distractions: phone calls, text messages,

and even taking selfies when driving. We can now add "augmented reality (AR)", a mashup of smartphone functions, to the list.

Augmented reality adds information and images to the visible world around you. For example, you might look at the road ahead through your smartphone screen and see… driving instructions, information about the cars near you, the history of a neighborhood, data on streets you pass, etc. And games. Mostly games. AR on your phone will be a deadly distraction for drivers.

By now, you may have heard of "Pokemon Go", a game so successful that a week after it's release the parent company's value rose by $15 billion. Players looking at their phone screens are blindly wandering onto private property in search of animated characters. Imitators will follow, and drivers will play while driving… just like every other phone application.

Technology will introduce more distractions and accidents will rise. Unless vehicles are driven by A.I. A.I. is not going away. We can only decide when to provide the regulatory framework. The costs and benefits are huge, but ultimately we must choose the rapid implementation of A.I. Every day without A.I. driven cars costs lives. This is one decision that we should all agree on!

Respondent 62

Charles Provine, Taoesm

Topics Covered in the Response are 2, 3, 4, 5, 8, and 9.

I am proposing a new ethic to apply to all AI endeavors that can give both broad and definite guidance on AI research activities. The ethic is Taoesm (an updated version of the venerable philosophy Taoism). Essentially, I propose a series of protocols that researchers must answer with their research into AI technology.

AI can be misused at many levels. Governments, corporations, and people could use these technologies for nefarious ends. While I cannot claim that the philosophy will prevent all rogue science, I believe that a philosophical framework must be constructed. Current philosophical modalities cannot adequately consider all factors, especially when the technologies employed are so different. Thus, I am proposing a series of symposiums to elaborate Taoesm so that the philosophy can encompass the best practices for all players in the AI spectrum.

Your RFI asks good questions, and I am sure that there will be many hundreds if not thousands of responses. It is my position that the United States and other countries are all ill-suited to institute the proper controls and regulations on the AI field. A fresh perspective

is mandatory to regulate all these emerging technologies, whether that is within the broader context of pharmaceuticals, self-assembling objects, law practice, computer technologies, militaries, finance, and others.

I started this philosophy circa 2010 at a graduate level seminar on the topic. I do not claim to be the authority on all AI technologies. I do believe that I can lead the academic, scientific, and corporate interests to devise the safeguards needed to bring effective regulation to the AI sector. I have training in Anthropology, Business Administration, Project Management, and Computer Engineering.

Respondent 63

Russ Altman, Stanford University

The 100 Year Study on AI submits its first study panel report which will be available from https://ai100.stanford.edu/ This site currently has a preview of the report. The full report is expected at the end of August, 2016. The 100 year study is a longitudinal look at AI; we anticipate a study approximately every five years evaluating how the effects of AI ripple through every aspect of how people work, live and play.

Respondent 64

Kenneth Blum, Center for Brains, Minds, and Machines

A Call for Major Public Funding for the Science of Intelligence

In the essay below we respond primarily to questions 5, 6, and 8 of the RFI: (5) the most pressing, fundamental questions in AI research, common to most or all scientific fields, (6) the most important research gaps in AI that must be addressed to advance this field and benefit the public, and (8) the specific steps that could be taken by the federal government…to encourage multi-disciplinary AI research. We also touch on these questions: (2) the use of AI for public good, (3) the safety and control issues for AI.

We believe the United States should develop a major, publicly-funded research effort directed at developing the science of intelligence. We begin with a review of where the science of intelligence stands today, review the biggest questions and gaps in AI, and conclude with a rationale for public funding.

You can be excused for thinking that machines exceeding human intelligence are within sight—some prominent technologists have heralded or warned of their coming. The truth is that we have no idea when this could happen, but it is unlikely to be any time soon. In fact, we have little understanding of what comprises human intelligence—certainly at the level of underlying mechanism, but even at the level of scientific description.

True, rapid progress is being made in implementing computer programs that can learn from examples. Computers are now superior to humans in chess and Go—classic games of pure strategic and tactical skill—and historically we have deemed the human masters of these games to be very bright. Similarly, Deep Mind developed a program that learned video games by practicing them.

Some critics have claimed that these machines are not intelligent, because their solutions don't seem particularly clever. Deep Blue's chess was driven more by a brute-force search through possible moves and counter-moves than appears to be the case for human chess-playing. Deep Mind's software failed to learn video games featuring rare "fatal" events that are obvious to humans. Watson's Jeopardy mastery was due to a composite of many modules whose interactions were carefully tweaked by hand and are difficult to generalize.

But why claim that this "machine learning" is not intelligence when we have only a dim understanding of intelligence itself? We have much to learn about how humans play Go or video games. Furthermore, human intelligence was built by evolution, and, a bit like Watson, is full of accidents that have been frozen into brain circuitry, with features that are evolutionarily recent and far from optimal.

Nevertheless, humans and even other animals have forms of intelligence that go well beyond the capabilities of even the most powerful computers. Understanding the basic principles of physics has led to spectacular machines that can travel to other planets, use a handheld device to communicate with people around the globe, and understand the origins of the universe. We expect that if we can learn the principles underlying biological intelligence, someday we will be able to engineer more powerful solutions to a limitless set of problems, freed from the constraints of low power, modest size, and self-wiring that limit our brains.

What is intelligence? The science of intelligence is in its infancy, but here are some likely elements.

We have to make sense of the world around us. Information comes from our senses—light, sound, chemicals, and touch are detected by specialized neurons. Our brains then have to analyze and assemble these signals and somehow compare them to models of the world. Our brains also must use input signals coming from our own bodies to control posture and movement. Thus, we model our environment and self, and we predict our future—imperfectly, of course—so as to act in the world. We avoid an accident; we find a friend in a crowd; we plan a vacation.

We might not think of any of these things as intelligence, in the casual use of the word, because we do them so effortlessly. Nevertheless, they all involve computations that are not yet understood, accomplished by mechanisms that remain partially mysterious. And we are not alone in these capabilities: mental models of this sort can be found in other animals.

Dogs catch Frisbees; cats can distinguish individual humans; spiders catch flies; flies court their mates.

We have social interactions—linguistic and non-verbal—and we must make sense of that social, cultural world, as well. We can predict the thoughts of others, and judge the fairness of interactions we observe. Children—and pets—manipulate our behavior. We have intentions and drives. Our thoughts have emotional content. We reason about the world based on incomplete information. We experience a serial, autobiographical monitoring of small parts of our own brain processes—namely, consciousness. We produce art, music, food and drink, dance, sports—sensory and motor capacities filtered through brains and culture. We tell stories and create literature—language designed to access internal states in other people. We produce science itself—an intellectual and social activity aimed at deeply understanding the world around us and even ourselves.

The intelligent behavior of humans, so impressive at the level of individuals and their activities, both quotidian and exalted, seems to have fundamental failings at the level of large groups. The ills of the world—wars, oppression, poverty, and the unintentional destruction of our environment—seem deep-rooted, with no signs of abating. Can our societal interactions be more intelligent? Is there a practical path that can get our species to that state? Perhaps these are the ultimate challenges for a science of intelligence.

These myriad capabilities, which, in our ignorance, we bundle into the single word "intelligence," are not yet the focus of a concerted scientific effort. To engineer deeply intelligent machines we first will need focused science to discover: What is intelligence and how does it work?

Recent engineering successes have—justifiably—created great excitement. The investments of a handful of big companies and a passel of startups in applications of artificial intelligence are changing our landscape. However, these developments primarily will take place behind closed doors: most of these commercial systems will be poorly understood, leaving us vulnerable to unintended consequences and without informed debate about ethical questions such as life or death decisions by self-driving cars or intelligent weapons.

What we need now is a major, publicly-funded research effort on the science of intelligence. This will have three effects. First, it will lay a firm foundation for engineering AI applications, which today are based on fifty-year-old ideas. Second—as has been essential for genome sequencing—it will ensure that the public has a strong voice in the dialogue and opportunities that arise from basic research on artificial intelligence, including appropriate debates about adequate regulations for the development of safe AI. Third, it would aim at the noble goals of understanding ourselves and each other so as to make the world a better place; applications could be directed at environmental problems, educational wonders, and alleviating destructive social interactions.

A large program to develop the science of intelligence would be the most important way for the federal government of the United States to provide public benefit from the field of artificial intelligence and lead the way towards what may become the biggest revolution in the history of civilization.

The Center for Brains, Minds, and Machines
http://cbmm.mit.edu/

Respondent 65

Michael Richards, Large public software company

I appreciate the opportunity to share my opinions with the OSTP.

AI methods are rapidly being adopted across all major industries. The growth of AI and other advanced software technologies is largely enabled through the free exchange of information and code. A robust, fast internet infrastructure, free of mass surveillance, censorship and commercial control is a fundamental requirement. In addition, funding for basic research in math and science as well as for long-term projects can help balance the short-term, market-driven technology development that dominates industry.

Like other new revolutionary technologies, such as nanotech, gene editing, etc, the technology is developing faster than our social and ethical frameworks. These technologies are maturing quickly and will radically change our lives in unforeseen ways. Government has a responsibility to fund serious cross-disciplinary, international efforts to understand the social, ethical, economic and ecological ramifications of these technologies.

Finally, AI will continue to reduce the value of traditional labor, as more and more tasks are automated. As has been the trend since the start of the information age, value will reside more and more in knowledge than in labor. There is significant risk that AI, robotics and other advanced automation will continue to deepen the disparities between the well-educated and moneyed few and the rest of the population. On the other hand, there could be significant opportunity for more and more people to participate meaningfully in the knowledge economy. While AI is a deeply technical area, the application and management of AI will become less technical over time. So in order to prepare our nation to benefit from AI, we need to fund robust, well rounded education for all our citizens. A population with a deep and diverse educational base, provided with fair access to technology, will be best positioned to take advantage of the coming technologies in a tremendous variety of novel and valuable ways.

Respondent 66

Charlie Berger, Oracle Corporation

I've been involved with AI, machine learning, robots, machine vision, data mining, predictive analytics and big data for most of my 35 year career. Love this field! I don't know what to comment on here other than full speed ahead. If there is anything that I can do to foster or help in this area, let me know!

Charlie Berger
Sr. Director of Product Management, Oracle Advanced Analytics, Machine Learning and Data Mining

Respondent 67

Tom Flahive, The Flahive Group

#4 - The social and economic implications of AI: The US will have to consider a "Guaranteed Income" (GI). The GI would be paid for by taxing AI related companies at a rate determined by the number of workers displaced by their AI system. For example: companies developing automotive collision avoidance systems would be taxed to pay for the auto collision repairmen displaced, and the companies put out of business.
#9 - Any additional information related to AI research...Based on my research of the US Patents, the US is number one in patents for "AI technology" (Artificial Intelligence and Deep Learning). But for "AI implementation" (Robot), the US holds only one-tenth of the patents of the leader (Japan). So, to implement AI in a robotic system the US will have to pay Japan for the rights, or face patent infringement lawsuits.

Respondent 68

Rob LaBelle, The Defining Thought

As the Administration works to leverage AI as an emergent technology for public good and toward a more effective government and to improve government services in areas related to urban systems, smart cities, social welfare, etc., as well as to realize the AI-driven improvements that can help vulnerable populations, it is imperative to ensure a people-centered design approach. In order to improve the impact and outcomes of relevant and appropriate programs having a greater emphasis on gathering end-user insights and behaviors is essential. With this, The Defining Thought commends the White House Office of Science and Technology for its Request for Information: Preparing for the Future of Artificial Intelligence (AI).

AI is essentially the intelligence of machines or a smart process which is continually learning and improving from data collected. There is tremendous power in the data collected that can benefit citizens of the world, and in the context of this RFI, effective and efficient government and public good via informed social programs and services. Likewise,

with the power of data collected comes a great responsibility for the capture, use and storage of the data. Collectively we need to seriously contemplate and prepare for what happens when this pervasive machine learning and constant evolution of the data/device/interface out paces human capacity to understand what is actually happening with their data and to them.

We live in an era where many citizens live is a state of vulnerability and are often mesmerized by the promise of technology, so much so that they overlook or disregard such core issues surrounding the collection and use of their personal data that will be used in machine learning and artificial intelligence. AI is pervasive as it will continue to touch all industry sectors and communities as we progress to a fully integrated and connected digital world rooted in and fueled by information communication technology.

As the OSTP examines the use of AI for public good (Q2), it is imperative to be open and transparent on what is meant by public good. In short to define public good. As implied, the general public is uninformed in regard to how and why their data is being incorporated into massive systems and manipulated. Indeed, great benefit can be realized from a responsible collection and use of data, but likewise we need to clearly see how easy it can be to cross over from public good to control with unintended consequences and impact on citizens' human rights.

As the OSTP examines the safety and control issues of AI (Q3), a priority should be a people centered approach, noting that end users (the public) will play a significant role in understanding and addressing privacy and security concerns. From a risk perspective on the services side, a security risk to the AI experience and use is what may be introduced to end users. There are obstacles that need to be overcome via security models and application that work with or are integrated into the respective service systems infrastructure based on specific needs so as to raise comfort levels and to ensure a high degree of assurance of security and privacy measures and equally effective and efficient security risk mitigation processes, protocols and response solutions so that the end user is not compromised.

Also for the OSTP to consider is that new AI applications will not have not fully gone through their respective iteration cycles and that this can cause a potential increase in security risk. There are existing security risk or threat models that can be leveraged but these models may need improvement or adjustment for application in AI in services. It will be imperative for the government to work in public/private partnership to lean on case studies from technology and applications that are in effective use today to help mitigate those risks.

In general, in addressing how to leverage AI in government services for public good in the context of security and control the following should be considered:

- If security measures are too stringent this will make it difficult for end users to use

services, therefore they will not be used or deployed, hindering uptake of services.

- How to address co-existence and interoperability of physical devices running on open platforms when the data on those devices and platforms need to be brought into closed or more controlled environments or vice versa.

- End users on an enterprise level may not be prepared to accommodate AI technology in their current IT and/or security systems and infrastructures.

- How to provide the technical support for specific services needs relative to capture and storage of data and retrieval of data, notably as many applications may relay on third party parties and some are hesitant to give away this level of control.

- Considering "security by design" or building security and privacy into design is critical. Security is too critical to the application to leave it an after thought.

- A mind-set shift has to occur that security is more than a physical concern. This shift in thinking will potentially inform a shift in budget planning, which may be a reallocation vs an increase.

- There will be additional costs relative to the management of AI in services. The costs will be related to tracking devices and secure data storage, and building trust-worthy user interfaces.

From a security and privacy perspective AI resembles previous new technologies entering the government but may have differences that pose unique challenges for the government. These differences are in the context of the department or office of government deploying AI in its services and strategies for ensuring secure data capture, rendering, use and access.

Respondent 69

Gordon Irlam, Gordon R Irlam Charitable Foundation

I am responding to (11) of the Federal Register and (9) of the Whitehouse website RFI: additional information.

I wish to address the current state of neuromorphic computing, it trajectory, and implications with respect to human-level or smarter-than-human AI. I do this not out of a belief that the neuromorphic approach will beat the machine learning approach to AI, but because the neuromorphic approach defines an upper bound within which we should reasonably expect human-level or smarter-than-human level AI to emerge. The neuromorphic approach involves learning how the human brain works, and then replicating it in full or in part in silicon. I restrict consideration to spiking neuromorphic models. Such models represent neurons as entities that perform simple computations and then either fire

or not. It is highly plausible that the human brain can be described reasonably well by such models, although this is by no means certain.

By the way of a preamble I will note that humans dominate the planet, not because we are stronger, but because we are smarter. Left unchecked, the development of smarter-than-human AI, neuromorphic, or not, is likely to result in a serious threat to public well-being. It seems unlikely that human values, such as love and compassion, would carry over to a world with human level or smarter-than-human AI, since these values appear to be evolutionarily encoded forms of genetic self interest, and are not a consequence of intelligence.

There are three issues for the neuromorphic approach to prove successful:
1. Is it feasible to implement the human brain or something smarter in hardware?
2. Do we understand how the human brain works well enough to be able to implement it?
3. And if it is feasible, how economical is it to do so?

Addressing the first of these issues. The feasibility of implementing a human brain in hardware. In 2014 there were a reported 2.5x10^20 transistors manufactured worldwide, and this number was growing 10 fold every 5 years. A typical human brain contains around 86x10^9 neurons. IBM's neuromorphic chip, TrueNorth, contains approximately 5,400 transistors per real time spiking neuron. Thus there were enough transistors manufactured in 2014 to produce the equivalent of 540,000 human brains. This could be one powerful superintelligence, or many smaller ones. In other words to the extent to which the neuromorphic approach is limited by the performance silicon, these constraints are rapidly disappearing.

Addressing the second issue. Do we understand the details of how the human brain works? We understand the big picture, the thalamus, the amygdala, the cerebral cortex, and so on, and roughly what each component does. We also understand the how individual neurons work at a very detailed level. However, at the intermediate level, how neurons are wired together to form assemblies of tens of thousands to billions, we know very little. This is likely to change.

The U.S. BRAIN Initiative is a signature big science research initiative primarily with the goal of better understanding the brain in order to better treat diseases. A notable exception to this goal is the 5 year $100m IARPA MICrONS sub-project which "seeks to revolutionize machine learning by reverse-engineering the algorithms of the brain". The technical goal of MICrONS is to produce a wiring diagram, or connectome, for 1mm^3 of cortical tissue (roughly 1 cortical column). It is widely believed that the neocortex is composed of a simple circuit, the cortical column, which is repeatedly replicated. Since the neocortex is believed to be responsible for most cognitive functions. It is thus only a relatively small step from understanding a 1mm^3 of brain tissue to understanding almost the complete brain. We are not talking about needing to scan a complete human brain, the cost of which appears

prohibitive for the foreseeable future, but a single 1mm^3, which is technically quite feasible.

A plausible research trajectory might be:
1. Determine the wiring pattern for one particular cortical column
2. Map the wiring pattern of one particular cortical column to a set of general wiring principles
3. Develop wiring principles for other brain regions
4. Understand the extent to which synapses are dynamic connections that vary over time
5. Gain a better understanding of learning, memory, and how the brain encodes information
6. Develop human-level or smarter-than-human AI

If the present policies of funding such work continue, a reasonable time frame over which such work might occur is perhaps 15-25 years, with the first three steps occurring relatively quickly, steps 4 and 5 being harder to predict, and step 6, the development of a computer as smart or vastly smarter than a human, being straight-forward given the preceding steps.

The final issue to address is economics. Today if we knew how the brain was wired we could achieve neuromorphic human-level AI for an estimated cost of around $700/hr. This cost estimate is based on an order of magnitude cost estimate for IBM's neuromorphic TrueNorth chip of $50 if produced in volume. Since this chip is proprietary a precise cost estimate is difficult to determine. Computing costs decline by around a factor of 10 every 8 years. Thus, assuming the current research trajectory, by 2040 we should be prepared for human-level AI for around $0.70/hr, and superintelligence costing $7/hr. If true, this would have broad societal impact.

In conclusion. The risks of different neuroscience projects vary widely. I suggest that the U.S. have a policy of funding non-health focused neuroscience projects like MICrONS, only if we can be sure the benefits outweigh the risks associated with the research trajectory they place us on.

Respondent 70

Tuna Oezer, System AI

(1)
- There is no clear legal framework to the extent a manufacturer or supplier of AI technology can be held liable to damage caused due to the use of the technology by another person. Liability issues are most clear in the case of self-driving cars but may apply also to more benign uses of AI. Ambiguity or excessive liability claims may discourage or slow down business investment in AI. It is important to create a reasonable legal framework that balances consumer rights with business interests.
- AI technology that uses machine learning depends on access to data. To achieve best

results, input data to a machine learning system may need to include personally identifiable information and track subjects across time. Current government regulations, such as HIPAA, are not designed for this use case. This limits the ability to deploy useful AI technology in areas such as health care. Government regulations need to be updated to balance privacy with the need to use data for machine learning purposes.
- Automated systems that classify people may produce incorrect results or the results may be incorrectly interpreted. In some cases, this may cause that a person is erroneously denied a benefit. An example could be a person being placed on a no-fly list due to a classification by a machine learning algorithm. Consumers should be given the right to demand that such errors be corrected and safeguards should be in place to prevent such incidents.
- The intellectual property ownership of models learned by AI technology may need to be further clarified. The current default is that the owner of the algorithm owns any results of the algorithm regardless of the owner of the input data. However, there is no clear legal framework for this position. Furthermore, it is unclear to what extend a learned model would violate any non-disclosure laws. For example, a machine learning algorithm of company A may learn a model from data owned by company B and use it to benefit a company C. Such a scenario should be possible even if company B maintains copyright of their data.

(2)
- AI has the potential to transform virtually any segment of the economy. Key AI features include the ability to discover complex patterns in large amounts of data, the ability to find optimal resource allocations, and the ability to automate human-computer interaction. These features can be used to improve health care, optimize traffic and infrastructure, save water, reduce energy usage and thus reduce pollution, and provide more efficient and automated government services.
- AI technology could be used to monitor and improve the effectiveness of government programs. With AI the government can become more data driven and respond to the needs of citizens in more pragmatic ways.

(3)
- Currently, there are no formal methods to provide quality assurance of AI technology or verify its correctness. AI software is very different from traditional computer software. AI software is usually adaptive and may modify itself. Furthermore, the behavior of AI software frequently depends on large amounts of input data. Current software engineering techniques to debug and verify computer code assume static programs with small amounts of input. These techniques are inadequate to debug and verify AI software.
- Computer security in the context of AI is another open issue. Machine learning algorithms may create unknown security holes. It is already known that an attacker may trick certain machine learning algorithms using well crafted input.
- Another concern is incorrect use of functioning AI technology by inadequately trained humans. Frequently, AI software depends on appropriate input by a human operator or the

output of AI software may need to be interpreted by a human (e.g., a classification result). If a machine learning algorithm is trained with insufficient or biased data, use of the final output of the algorithm may produce incorrect results causing harm to people.
- Despite some public concern, there is absolutely no imminent danger of a sentient AI taking over the world. AI technology by itself will not cause harm to humans unless it is misused. Thus, its important to ensure that human operators are properly trained and screened.

(4)
- Long-term, AI is likely to improve the living standards of most people, create jobs and contribute significantly to economic growth. However, some low-skill jobs may become obsolete. Government programs may be needed to help affected people to transition to new jobs.
- AI makes it more important that employees have analytical and quantitative skills.
- In order to remain competitive, the US should encourage the teaching of more STEM skills in education and shift its immigration policy to favor high skilled workers.
- An increasing minimum wage will accelerate automation of low skill jobs using AI.

(5)
- One of the most important open issues is common sense reasoning. While current technology is good in simple pattern discovery, reasoning is still an unresolved issue. Lack of reasoning limits more advanced natural language understanding and computer vision.
- While current AI technology is good in answering questions, it is not very good in asking questions. In other words, knowledge and feature discovery are unresolved problems.
- Human-computer interaction in the presence of AI technology needs more work. This is especially an issue in cooperative settings where the AI takes over limited control from a human such as in self-driving cars.
- As mentioned in (3) verification and debugging of AI software as well as computer security in the context of AI have not been widely researched.

(6)
- Among the points listed in (5) advances in verification and quality assurance may be the most important in use cases where AI software can cause damage.
- Human-computer interaction is another important area. Bad interaction can result in accidents due to miscommunication with the machine. Another issue is that machine learning results may be incorrectly interpreted by the human operator or the operator may use inappropriate parameters.
- AI will continue to improve in other areas, but the progress in these areas is not a limiting factor in the deployment of AI.

(7)
- The use of AI technology requires quantitative and analytical skills. This includes basic statistics and mathematical thinking. Lack of such skills by a user could cause incorrect use

of machine learning software. Furthermore, such skills are necessary to identify new business opportunities where AI can be used.

(8)
- Collect data sets and provide open access to such data sets.
- Make it easier for the population to volunteer their data for research (for example in health care).
- Set up more benchmarks in a wider set of AI tasks (similar to image classification benchmarks).
- Provide small business grants to start-ups focusing on AI technology.
- Encourage children to pursue a STEM education. This may include making STEM look "cool" in popular culture and providing long term incentives to pursue a STEM career (e.g., tax credits).
- Sponsor more "grand-challenges" similar to the DARPA challenges across other AI fields. These challenges may be smaller scale and more incremental.
- Educate the general population about the benefits of AI rather than spreading an unfounded fear about a sentient AI uprising.

Respondent 71

Yevgeniy Vorobeychik, Vanderbilt University

Question 2 (The Use of AI for Public Good):

In my view, this is a red herring. Generally speaking, arguably the vast majority of AI researchers already set public good, broadly construed, as their aim (this is an effective pre-condition of nearly all research funding, for example). Granted, one could argue the extent to which specific research areas within AI significantly contribute to public good, but such arguments are nothing new. Often, fundamental research which appears at one point very far from socially relevant application eventually becomes crucial (number theory is one prominent example insofar as it provides foundations of cryptography). We are hardly ever forward looking enough to reliably predict what research ultimately becomes pivotal for public good. Consequently, in my view, rather than trying to stimulate specific areas of AI research deemed by policy makers as most conducing to public good, it may be the best policy to try to stimulate AI research broadly, allowing researchers to organically determine the questions that are most worthy of pursuit. The increasing prominence of AI in a broad array of socially important applications suggests that AI researchers have been quite successful in steering AI towards public good when given the opportunity.

Question 3 (Safety and Control Issues for AI); Question 5 (Fundamental AI Research Questions); Question 6 (Gaps):

As we move forward with the maturing arsenal of AI techniques, which are largely designed

for non-adversarial situations, a natural question is how robust modern AI methods are if motivated adversaries attempt to subvert their functionality. One emerging domain of research related to this broader question is in adversarial machine learning (AML). Attacks on machine learning algorithms can come, roughly, in two forms: evasion and poisoning. An evasion attack is best explained using a spam filtering example. Suppose that we have trained a classifier on data to distinguish spam from non-spam. Spammers would subsequently be motivated to modify spam instances in order to be (mis)classified as benign (that is, to evade detection). More generally, in an evasion attack, an attacker wishes to manipulate malicious instances in order to avoid detection by fixed learning algorithms (typically, binary classifiers in such domains). Poisoning attacks are different: in these, the adversary is assumed to modify the training data set itself, so as to subvert the learning algorithms. Poisoning attacks may have goals ranging from degrading observed performance of learning algorithms (the so-called availability attacks) to allowing future adversarial behavior to bypass detection. Given the importance of machine learning algorithms both in research and commercial application, consideration of such attacks on machine learning algorithms may be one of the most significant emerging research problems today.

Respondent 72

Dennis Cooper, Self

I have some comments regarding item #9 "specific training sets that can accelerate the development of AI and its application." It is absolutely crucial that large training sets be created and maintained by a third party. These training sets can be used to:

- validate the performance of an AI system
- score the performance of an AI system
- document troublesome cases that might be critical to safety
- serve as a basis for acceptance testing

Often times, these training sets are considered proprietary and not widely disseminated. Maintaining a central website with available training sets is required. Developers need to be able to add to the global training set after acceptance review. This makes it a community generated data source.

We are all concerned whether or not an AI system will make the “right” decision with higher performance than a human. Humans are prone to basis, fatigue, and boredom. AI machines are not humans and they don’t suffer from these characteristics if properly coded and tested. Training sets are essential to testing and achieving superior performance. The Government needs to fund the maintenance and release of training sets that are crucial to citizen safety.

Respondent 73

Daniel Golden, Arterys, Inc.

(1) the legal and governance implications of AI

In healthcare, as in other fields in which a computer may be responsible for protecting human life (e.g., the automotive industry), the question of liability is a perpetual concern. For example, in healthcare, if an algorithm is developed that can predict a diagnosis, and the algorithm's prediction is found in one instance to be wrong in such a way as to cause harm, who is liable for that mistake? Is it the company that developed the algorithm, the regulatory agency that cleared the algorithm, the doctor who signed off on the erroneous automated diagnosis, or the patient who gave informed consent to have their care partially determined by an algorithm? All parties likely share some blame, but what if, on average, health outcomes are significantly improved when the algorithm is used?

In order to inspire companies and clinicians to develop and utilize AI solutions that may dramatically improve overall health outcomes, increase access, or decrease costs, they must not be unreasonably punished for the infrequent cases in which the predictions are incorrect as long as they have made a good faith effort to minimize those occurrences and to appropriately inform users of the risks. With sufficient data, a robust confidence score can be given to any diagnosis, that will in no doubt trump the current standard of care in healthcare.

(2) the use of AI for public good

AI has the potential to significantly reduce the cost of healthcare by cheaply creating optimal care paths for patients. In addition, cloud-based AI systems can be used over the Internet via mobile connections in regions that have limited infrastructure. AI has the potential to democratize healthcare globally, erasing geographic borders and offering the same quality of diagnostic care to every single individual of the world, bringing technology that was previously reserved for elite academic centers.

(3) the safety and control issues for AI

Each unique AI application requires an independent assessment of risk. Based on the risk level, a medical AI application can get clearance to perform its function on its designated audience. The model to follow is very similar to the FDA model of a Class I, Class II, and Class III medical device. Class I devices are inherently safe, and cannot cause harm, while Class III devices can be life threatening and therefore require the highest level of scrutiny.

We believe that the open source software model will drive all software projects of the future. By allowing the public to scrutinize the underlying code, we dramatically increase the likelihood that defects or malicious intentions will be caught before the software can cause any harm. In addition, a strict code review policy (e.g. where an author cannot merge their own code) is critical in assuring that no one person can make a software change that has malicious intents.

In health care specifically, our initial goal should not be to replace the physician with an "intelligent" algorithm, but rather to use AI for decision support, efficiency improvement, and automation of tedious tasks. Once incremental value is proven, AI can be expanded to cover more aspects of clinical decision making, but always with a clinician checking and signing off on the final decision. AI should augment and expedite a physician's tasks, not replace them.

(4) the social and economic implications of AI

AI has the potential to significantly reduce healthcare costs and bring advanced diagnostic capabilities to underserved communities, largely due to its ability to scale. Unlike physicians, who can provide diagnostic services to a very limited number of patients daily, AI, much like Google and Facebook's services, can serve millions or billions of individuals in an instant. For healthcare services to which AI is well suited, such as medical diagnoses and treatment planning, AI has the potential to reduce to insignificance the costs born by the individuals who receive this care.

(5) the most pressing, fundamental questions in AI research, common to most or all scientific fields

The most pressing needs in AI research are (a) the ability to process unstructured data (such as images, text, electronic medical record data, etc.) and (b) the need to determine in advance the best data to collect to allow optimal predictions to be made by AI services.

With respect to unstructured data, great advances have recently been made in the field of deep learning, which allow for the creation of models which can take in unstructured data sources, such as images, and produce classifications, segmentations, or more elaborate results, such as, textual descriptions. However, relatively little research has been performed on the synthesis of heterogeneous data sources as input into these models. For example, to predicting a patient's diagnosis from both the raw data from their CT scan and from an in-clinic questionnaire with a mix of multiple choice and short-answer questions would require a patchwork of loosely-tied-together models, each optimized for a different type of data. A model that could synthesize all this data naturally and simultaneously, as a human doctor would, would have great benefits for clinical care.

Regarding collecting the right data, this is not a problem unique to AI, but also an issue for standard clinical care. Often, a diagnosis is inconclusive or incorrect for lack of a critical blood or imaging test for the patient. Ordering these tests at the time of care prolongs the time before a diagnosis can be made. If we instead had a way of determining the optimal data to have collected beforehand, or if we had a method for optimizing which tests would be most important in determining a conclusive diagnosis, we could minimize time and

money spent waiting for tests, particularly those that are not likely to increase the confidence of diagnosis.

Respondent 74

Nicole Miller, None

I just found out about this RFI today (7/20/2016) and am heartbroken that I won't be able to submit anything in time. Is there a way I can submit my information in a couple of weeks? I am a small-scale AI programmer, but I believe my concerns are valid and would apply to the general populace. I've used computers for more than 20 years and have watch the internet evolve within the same time period, and would relish the thought that someone in the government would actually pay attention to my concerns.

Thank you so much!!

Respondent 75

Nicole Miller, None

I forgot, my email address is XXXXXXXXX

Respondent 76

Gisela Wilson, University of Wisconsin

Artificial Intelligence: Ethics, Privacy and Conscience

'Conscious' artificial intelligence (AI) may be far in the future, however, decentralized AI is here. From health programs to predictive policing and everything in between, algorithmic systems are making perceptible inroads into human life (Crawford, 2016; Pasquale, 2015). Crucially, however, decentralized AI should be distinguished from individualized robots and lacks the integration and 'consciousness' foretold in science fiction. Interactions between various algorithmic systems remain lumbering, disjointed, and underdeveloped with one exception: the capitalist aim of corporate gains.

Algorithmic systems have the potential to encourage, as well as prevent, dignity of life. The focus on corporate financial gains diverts attention from critical issues — cantankerous for their complexity (Chandler, 2014) — to convenience and marketability. Designers appear locked into cycles of upgrades geared toward trivialities and monopolization, rather than compatibility, ease of use and support for causes of well-being. Damania (2016) describes typical interactions with software that many experience on day-to-day basis. Upgrades can feel like coders moving the furniture around in your living room so you can trip over it at night, even when working directly with systems. What does that suggest about data

collected and decisions made that guide people with or without their knowledge?

In spite of an industry that is claiming significant global financial resources and touts solutions, the most pressing crises are not getting solved leading to greater inequality, unrest and probability for war. A person does not need to be a genius to know that inclusive affordable housing, education, employment, health care, and care for the environment would result in a healthier more sustainable society and world.

It seems essential to develop systems to correct for, rather than encourage, power asymmetries, including corporate and media use of our on-line data shadows — use that, at times, seems oblivious to issues of privacy, dignity and ethics (Cohen, 2015; Crawford, 2016; Pasquale, 2015; Zubhoff, 2016). Without knowledge and approval, data collection and sharing (e.g., Al-Rodhan, 2016; Fiveash, 2016; Newman et al., 2016; Tenney & Sieber, 2016) is increasingly invasive and yet the beneficiaries, instead of citizens, increasingly are corporations and governments. Thus far, increases in information have increased bureaucracy and obfuscation, while sacrificing citizen privacy and choice (boyd, 2016; Larson et al., 2016; Cohen, 2015; Floridi, 2016; Zubhoff, 2016). Echoing Zubhoff, increases in surveillance and behavior modification on the basis of data not situated in perspective and context compromises the freedom of self-determination that is the foundation on which our countries are built. It removes responsibility to those doing the modifying, a legal issue our courts have barely begun to address. Even the European Union's recently proposed General Data Protection Regulation (Claburn, 2016; Goodman & Flaxman, 2016) doesn't go far enough, though it's a start. Algorithms rarely work in isolation, so it is not clear where responsibility for outcomes would fall. Moreover, an explanation is not a solution or reparation for harms. Harm to individuals ripple through communities and institutions, and extend back to the trustworthiness of corporations, agencies and governments.

One plausible solution to the capitalist focus of "black-boxed" algorithmic systems would be an algorithmic filter that determines potential ethical consequences of their use. Several physiological analogies come to mind. Borrowing from (a) the retina or (b) interacting excitatory and inhibitory neural circuits that coordinate (a) visual perception and (b) movement, an ethical algorithmic filter would provide a "surround" to regulate a competitive capitalist "excitatory" output.

An ethical AI filter would likely improve on the implicit and explicit biases and reactivity of humans. In addition, ideally, the filter could access and be updated with status reports, research findings and legal arguments. An ethical AI filter would weigh information according to likely validity and search for missing perspectives and assumptions. Indeed, intentions to digitize research findings and literature, thereby increasing accessibility were laudable, though remain incomplete given the potential for misuse and proliferation of misinterpretation (e.g., Grove, 2016). In the case of science, one step beyond open access publishing would be coding data for perspective, compatibility and assumptions, thereby decoding science data from technical jargon.

Could ethical AI be mandated to restore protections for human privacy, discrimination and security that have been increasingly compromised in recent years? Similar concerns prompted regulation of human subjects research and genetic engineering. Even institutional review boards and government agencies often have been too narrow in not examining the combined overall effects of their decisions. Furthermore, as an example from an economic view for those with business interests, airline deregulation did more harm to the industry than good (Arria, 2016). Intriguingly, an ethical AI filter might function as an AI conscience — with individual, as well as global, dignity in 'mind' (Al-Rodhan, 2015; Burke and Fishel, 2016)

Contrary to the expectation that algorithms are indifferent, several concerns have been reported over the last ten years. Algorithmic systems risk enhancing rather than eliminating discriminatory practices either as a function of (a) their capitalist aim, (b) the implicit and explicit biases of their designers, (c) biases in the way data has been collected and combined, (d) ordering effects or (e) technical assumptions on which operational paradigms are based that fade into the background with time. An essential feature of the proposed ethical algorithmic filter, therefore, is that it be created by an independent group of researchers. The filter would examine data eliminated by "excitatory" algorithmic sorting processes using a lens sensitive to various biases and potential intersectional effects (Kitchen, 2014). The ultimate goal would be to develop an ethical AI mechanism that could be provided to corporations and government agencies for use in their own design process, thereby minimizing the proprietary black-box argument, increases in bureaucracy, and the need for regulatory oversight.

Another way to examine today's algorithmic systems also draws from visual processing. Could data points (people) eliminated during algorithmic sorting processes be tagged without loss of privacy and continue to be processed in parallel? Crucially, from the perspective of equity, what resources could be provided to increase self-determination, creativity and equalize, rather than judge and eliminate, individuals in the output thereby optimizing human potential?

A third suggestion is to limit "excitatory" algorithmic systems' ability to interrogate people. In this case, the method would be to redirect focus to an interrogation of systems and infrastructure contributing to the health of society and the planet. The primary aim would be an equitable allocation of resources and opportunity.

At a time when the world is in crisis, it seems a shame that the potential of algorithmic systems to resolve the most pressing issues is being distracted by more short-sighted aims that, however well-intentioned, enhance rather than reduce inequalities. Many are concerned (boyd, 2016; Crawford, 2016; Cohen, 2015; Floridi, 2016; Zubhoff, 2016; Pasquale, 2015; Pedziwiatr & Engelmann, 2016) that the speed at which institutions, corporations and governments are deploying algorithmic systems is in excess of the

mechanisms of ethical and legal oversight on which people and the continued existence of a habitable planet depend.

Credit for this comment goes to an unknowable multitude of 'research assistants', in addition to the authors below:

Nayef Al-Rodhan (2015) Proposal of a Dignity Scale for Sustainable Governance, Journal of Public Policy (Blog), 29 November. https://jpublicpolicy.com/2015/11/29/proposal-of-a-dignity-scale-for-sustainable-governance/

Nayef Al-Rodhan (2016) Behavioral Profiling and the Biometrics of Intent. Harvard International Review 17 Jun. http://hir.harvard.edu/behavioral-profiling-politics-intent/

Michael Arria (2016) The Surprising Collection of Politicos Who Brought Us Destructive Airline Deregulation. Alternet, 3 July. http://www.alternet.org/labor/how-liberals-deregulated-airline-industry

danah boyd (2016) Be Careful What You Code For. Medium, 14 June https://points.datasociety.net/be-careful-what-you-code-for-c8e9f3f6f55e#.4sobpvbe9

Anthony Burke and Stefanie Fishel (2016) Politics for the planet: why nature and wildlife need their own seats at the UN. The Conversation, 30 June. https://theconversation.com/politics-for-the-planet-why-nature-and-wildlife-need-their-own-seats-at-the-un-59892#

David Chandler (2014) Beyond neoliberalism: resilience, the new art of governing complexity, Resilience, 2:1, 47-63, DOI: 10.1080/21693293.2013.878544 http://dx.doi.org/10.1080/21693293.2013.878544

Thomas Claburn (2016) EU Data Protection Law May End the Unknowable Algorithm. InformationWeek, 18 July. http://www.informationweek.com/government/big-data-analytics/eu-data-protection-law-may-end-the-unknowable-algorithm/d/d-id/1326294

Julie Cohen (2013) What Privacy is for. Harv. L. Rev., 126, 1904. http://harvardlawreview.org/wp-content/uploads/pdfs/vol126_cohen.pdf

Kate Crawford (2016) Know Your Terrorist Credit Score. Presented at re:publica, 2 May. https://re-publica.com/en/16/session/know-your-terrorist-credit-score

Zubin Damania (2016) We need to demand technology that lets doctors be doctors. KevinMD, 1 February. http://www.kevinmd.com/blog/2016/02/need-demand-technology-

lets-doctors-doctors.html

Kelly Fiveash (2016) Google AI given access to health records of 1.6 million English patients. Ars Technica UK, 3 May. http://arstechnica.co.uk/business/2016/05/google-deepmind-ai-nhs-data-sharing-controversy/

Luciano Floridi (2016)The Informational Nature of Personal Identity. Minds and Machines, Vol. 21 Issue 4 – 2011: 549. DOI:10.1007/s1XXXXXXXXX6
https://www.academia.edu/9352388/The_Informational_Nature_of_Personal_Identity

Jack Grove (2016) Beware 'nefarious' use of open data, summit hears. TimesHigherEducation, 7 July. https://www.timeshighereducation.com/news/beware-nefarious-use-of-open-data-summit-hears

Bryce Goodman & Seth Flaxman (2016) European Union regulations on algorithmic decision-making and a "right to explanation" ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NY. eprint arXiv:1606.08813.
http://arxiv.org/abs/1606.08813

Rob Kitchen (2014) Big Data, new epistemologies and paradigm shifts Big Data & Society April–June 2014: 1–12. DOI: 10.1177/2053951714528481
http://m.bds.sagepub.com/content/1/1/2053951714528481.full.pdf

Jeff Larson, Surya Mattu, Lauren Kirchner & Julia Angwin (2016) How We Analyzed the COMPAS Recidivism Algorithm. ProPublica, 23 May.
https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Joe Newman, Joseph Jerome & Christopher Hazard (2014) Press Start to Track?: Privacy and the New Questions Posed by Modern Videogame Technology. American Intellectual Property Law Association (AIPLA) Quarterly Journal, 1 August.
http://ssrn.com/abstract=2483426

Frank Pasquale (2015) "The Black Box Society: The Secret Algorithms That Control Money and Information." Harvard University Press, Cambridge MA.

Samuel Pedziwiatr & Severin Engelmann (2016) Blueprints for the Infosphere: Interview with Luciano Floridi. fatum 4 June, S. 25. http://www.fatum-magazin.de/ausgaben/intelligenz-formen-und-kuenste/internationale-perspektiven/blueprints-for-the-infosphere.html

Matthew Tenney & Renee Sieber (2016) Data-Driven Participation: Algorithms, Cities, Citizens,
and Corporate Control. Urban Planning (ISSN: 2183-7635) 2016, Volume 1, Issue 2, 101-

113 DOI: 10.17645/up.v1i2.645.
http://cogitatiopress.com/ojs/index.php/urbanplanning/article/download/645/645

Shoshana Zubhoff (2016) The Secrets of Surveillance Capitalism Frankfurter Allgemeine Feuilleton, 3 May. http://m.faz.net/aktuell/feuilleton/debatten/the-digital-debate/shoshana-zuboff-secrets-ofsurveillance-capitalism-14103616.html

Respondent 77

JoEllen Lukavec Koester, GoodAI

(1) the legal and governance implications of AI

GoodAI, a Prague-based AI research and development company where I am a team member, aims to encourage governments to invest in human-level general AI research and development technology so that they have a share in the wealth that will come from it later on. We anticipate that this wealth will be distributed to citizens through basic income if needed. We want to formulate a plan for the institution of a basic income for all people by consulting with economists and other experts, further consider peaceful vs. military uses of general AI technology and its implications for society going forward, support a potential UN resolution banning the use of AI in autonomous weapons, and begin to think about ways that AI companies and governments can cooperate on security.

(2) the use of AI for public good

GoodAI's mission is to develop general artificial intelligence - as fast as possible - to help humanity and understand the universe.

AI can serve as very smart assistant, scientist, engineer, or tool to augment ourselves, to help us with thinking, creativity, problem solving, discovering new patterns in the universe, and so on. Should we invent a truly general, human-level artificial intelligence in the future, it will be flexible enough to help us solve any number of problems across fields. It could prove to be the sum total of human knowledge.

(3) the safety and control issues for AI

At GoodAI we are investigating suitable meta-objectives that would allow an open-ended, unsupervised evolution of the AGI system as well as guided learning - learning by imitating human experts and other forms of supervised learning. Some of these meta-objectives will be hard-coded from the start, but the system should be also able to learn and improve them on its own, that is, perform meta-learning, such that it learns to learn better in the future.

Teaching the AI system small skills using fine-grained, gradual learning from the beginning

will allow us to have more control over the building blocks it will use later to solve novel problems. The system's behaviour can therefore be more predictable. In this way, we can imprint some human thinking biases into the system, which will be useful for the future value alignment, one of the important aspects of AI safety.

(4) the social and economic implications of AI

The effect of AI on the employment system is in the center of interest these days, and we believe it is for a good reason. There are several possible scenarios for how AI and general AI will impact jobs, and we tend to lean towards the probability that AI will lead to job replacement in significant numbers.

In fact, that's what we're working towards - we want AI to work so that humans don't have to, and so that they can pursue more meaningful activity with their time. That said, there is a good chance that parallel economies may develop - an AI economy where human workers are undesirable and most work is performed by intelligent machines, and a human economy that closely resembles the economy of today. In general, this topic is under-researched and we believe that a great deal still needs to be done to ensure that artificial intelligence development leads to a scenario that is beneficial for most or all of humankind.

(5) the most pressing, fundamental questions in AI research, common to most or all scientific fields

The most pressing research areas, questions, and tasks include:

A widely shared view on what intelligence is, how to study it, how to test and measure it

Without understanding the underlying principles of intelligence, what kind of tool is it, what is it capable of and what not, what are the limits and potential - all our current efforts just skim the surface of these questions

Mapping a list of skills and abilities that human-level AGI needs to demonstrate

A framework that will help in mapping, studying, analyzing and measuring these skills - so that road mapping progress can be systematic and not rely on a "lucky idea"

We need a method to propose and test various hypotheses on where to direct AI research

Develop a learning architecture that is capable of gradual accumulation of these skills and abilities

Design an optimal curriculum, that helps AI in efficient acquisition of these skills and abilities

To achieve all these, we need people who can offer big-picture strategic thinking

(6) the most important research gaps in AI that must be addressed to advance this field and benefit the public

We need more companies and researchers looking into unified approaches to building general AI by investing into strategic big-picture thinking. At GoodAI we want to step outside of traditional approaches and offer a fresh, unified perspective on building machines that learn to think. We hope to achieve this in a number of ways, but especially through our roadmap, framework and by founding the AI Roadmap Institute.

(7) the scientific and technical training that will be needed to take advantage of harnessing the potential of AI technology

Multi-disciplinary approach seems to be the one which brings the most fruits for hardest problems. In order to build and educate a human-level general AI, we will need specialists from distant fields of science - computer scientists, machine learning experts, software engineers, neuroscientists, cognitive scientists, behavioral psychologists, sociologists, economists, and more.

(8) the specific steps that could be taken by the federal government, research institutes, universities, and philanthropies to encourage multi-disciplinary AI research

One step we are taking at GoodAI is to found an institute dedicated to the study of AI roadmaps. The AI Roadmap Institute is a new initiative to collate and study the various AI and general AI roadmaps proposed by those working in the field. It will map the space of AI skills and abilities (research topics, open problems, and proposed solutions). The institute will use architecture-agnostic common terminology provided by the framework to compare the roadmaps which will allow research groups with different internal terminologies to communicate effectively.

The amount of research into AI has exploded over the last few years, with many papers appearing daily. The institute's major output will be consolidating this research into an (ideally single) visual summary which outlines the differences and similarities among roadmaps, places where roadmaps branch and converge, stages of the roadmaps which need to be addressed by new research, and where there are examples of skills and testable milestones. This summary will be constantly updated and available for all who are interested, regardless of technical expertise.

There are currently two major categories of roadmaps; 1) Research and Development - how to get us to general AI, and 2) Safety/Futuristic - which explore how to keep humanity safe and the years after AGI.

Founding the roadmap institute is an opportunity for any researcher or organization - governmental, university, philanthropic, etc. - to come together in multidisciplinary research.

(9) any additional information related to AI research or policymaking, not requested above, that you believe OSTP should consider.

It is important to distinguish between directions that lead to narrow AI and directions that may lead to general AI. We have seen progress in the narrow AI direction recently, but not to the same extent in general AI, even though general AI has higher potential to help humankind. Some may argue that the progress in general AI cannot be achieved just be extrapolating current trends in narrow AI, and that we need novel approaches to the problem.

Respondent 78

Master Yoda, Jedi Order

Of possible interest about the use of AI for public good:
https://www.researchgate.net/publication/256987370_Artificial_intelligences_and_political_organization_An_exploration_based_on_the_science_fiction_work_of_Iain_M_Banks?ev=prf_pub

Respondent 79

Edward Lowry, Advanced Information Microstructures

(6) The most important research gaps in AI.
Advanced artificial intelligence often involves working with symbolic information whose subject matter is also symbolic information. Extraneous complexity imposed by the representation conventions used for such information gets COMPOUNDED and obscures progress toward solutions. Improving precise language to minimize such extraneous complexity is central to advancing AI.
There has been a severe shortage of competence , candor, and curiosity related to SIMPLICITY issues in the design of precise language. There has been enough intolerance of those values to obstruct progress in language development for about 45 years.
The resulting imposed complexity fog in working with precise information has impaired intelligence for both people and computers working with precise information. Resulting disasters in cyber-security, technical education, public safety, economic waste, etc have been extensive.
A brief illustration: In 1982 it was possible for employees at a leading computer vendor using a general language to enter for execution a computation such as:

```
count every state where
  populatn of some city
  of it > 1000000
```

Translating that computation into the best languages of 2016 will require expanding from those 12 symbols to about 30 with far less clarity.

Current technology perpetuates a false dichotomy between languages with rich data structure and those with simple plural expressions (such as SQL and APL). Resolution of the dichotomy was implicit in a design circulated at a leading computer vendor in 1973.
Large simplifications on several leading edges are possible and have been explored decades ago. The simplifications lead to a large convergence in the underlying language semantics – especially constraining the design of fundamental information building blocks. The convergence can lead naturally toward a lingua franca for technical literacy.
Vast populations routinely study or work at arranging pieces of information or instructing computers to do so. They do so in profound ignorance of fundamental building blocks of information that are well designed to be easily arranged. This reflects an avoidance of fundamentals which has little precedent in other technical fields.
While the language research needed is urgent for AI, it may be largely a matter of confirming current deficiencies and the potential for making rapid improvements.
(1) Legal and governance implications.
Development of language tools that serve the needs of artificial intelligence (see 6 above) would also serve the need to develop ontologies which could allow for simpler and more precise legal documents. Regulatory burdens could be reduced while doing their job more effectively.
(2) Public good.
Such language can improve technical education by expressing technical knowledge in a precise, readable form that avoids mystification. Improvements would also result in cyber security, public safety, worker productivity, etc.
(3) Safety.
Such improved language would simplify complex systems making them more reliable and safer whether or not artificial intelligence is directly involved.
(4) Social and economic implications.
As noted in (2) education and economic productivity would benefit from just the language improvements needed for AI but reductions in employment would be problematic. Successes in AI itself would amplify the effects.
(5) Pressing fundamental questions in AI research.
Reducing complexity fog in representations has become pressing after 45 years of obstruction. Understanding the structure of easily arranged building blocks of information is also a pressing issue in somewhat the same way that understanding the shape of bricks would be pressing if we lacked such understanding.
(7) Needed training.
There is a need to fully develop much higher quality language and learn to use it.
(8) Steps to encourage multi-disciplinary AI research.
A quality language for AI will be similar to a broadly applicable lingua franca for technical literacy -- which could enhance multi-disciplinary communication. A specific action to provide such a combination of capabilities would be to acquire the intellectual property rights to such a language implemented in the 1980s but never made publicly available and now owned by Hewlett Packard.
Another action would be to examine the competence and candor of software policy makers

around simplicity issues. Getting their translations of the above computation could be a place to start.

(9) Additional information for OSTP to consider.

For decades competent expertise in simplicity issues related to precise language has been almost completely silenced. That plus incentives to make other people's lives complicated have severely blocked progress in AI, software technology, and STEM education. The incentive to complicate is illustrated by IBM's gross profit margins for software: over 87% in recent years. Complexity keeps customers locked in because the pursuit of more competitive pricing is too disruptive.

BUT THE INTOLERANCE FOR SIMPLICITY IS MUCH DEEPER AND WIDER THAN THAT. Prestigious technology leadership has been paralyzed by a taboo against considering such simplicity issues. Checking the whole technical literature for evidence of non-paralysis would not be easy. The paralysis is more directly illustrated in the IEEE Computer Society "Software Engineering Body Of Knowledge, SWEBOK" where a roughly 500 member Review Committee ignores the potential of improved language design. Getting more functionality with less complexity by improved language should be a central theme of software engineering, but such efforts appear to have been eliminated and delegitimized within software engineering soon after its inception in the late 1960s.

The paralysis is probably also misdirecting curriculum planning at the high school level for the "Computer Science for All" initiative. Existing popular languages can be shown to be 40 years behind the leading edge on several simplicity-related leading edges. Inflicting them on millions of high school students would be a big mistake.

Knowledge and ideas are needed and in substantial part available and in some danger of being lost. Precautions should be taken. THE MOST IMPORTANT NEED IS FOR A CULTURE SHIFT AWAY FROM INTOLERANCE FOR LANGUAGE RELATED SIMPLICITY AND TOWARD STRONG SUPPORT.

President Eisenhower's farewell speech warning of the "danger that public policy could itself become captive of a scientific-technological elite" seems well justified in software technology.

For anyone pursuing the promise of artificial intelligence technologies, correcting these underlying deficiencies should be a high priority.

Respondent 80

Candace Sidner, Worcester Polytechnic Institute

Robots, it is claimed, will soon be commonplace entities in American society. This widely heralded belief will not occur on the basis of advances in robotic hardward and software alone. Robots must be designed to work with everyday people in their everyday spaces. Whether the robot is driving a truck in a mine or a car on the street [1], serving as a carrier bot for troops in the field, delivering meals and supplies in a hospital [2] or emptying the home dishwasher [3] and chatting about everyday matters, robots will be collaborating with people or at minimun interacting with them as they go about their tasks.

Experts in AI have had some success in modelling collaboration. Enough that we understand that the computational partner must understand the goals, recipes for achieving those goals, share beliefs with its human partner [4], and be able to judge when the collaboration is not succeeded and which choices it must make in that light [5]. All of these issues are equally significant when collaborating with a robot. However, many more matters come to bear, some of which are not the focus of this note (such as modelling the environment in which the robot acts, and modelling belief in a computationally adequate way and one that is computable in real-time). Collaboration requires robust and accurate models of interaction with a human partner.

As a collaborator with people, a robot needs a number of capabilities that are receiving only initial consideration in the field of robotics as a whole First, a robot (whether it looks something like a person or not) must gesture appropriately, that is use its face, its body and its hands to convey the focus of its attention over time to relevant objects to its collaborators or to indicate items to which it wishes to bring the human's attention, through gaze or pointing. It must also be able to understand when humans do the same types of gestures, and to know when the two partners know that they are talking about the same thing.

Second, it must recognize and use emotional cues to assess and create its stance about the interaction. Emotion plays a significant role in collaborations, especially when there are errors or failures of part of the actions that are expected to lead to the shared goal. Third, robots need knowledge of social and cultural norms to determine such matters as where it positions its body when entering into group activities and group conversation. It must know where to look and how to change its conversation to adhere to social norms during activities with people (and other robots). A miscue of social norms can lead people to misunderstand the robot's intent or to force people to reorganize their parts in the collaboration, or to fail altogether. Fourth, a robot must be able to converse with its human counterpart all the while using gestures to reflect its attention, emotions and social knowledge. Finally, robots must be able to learn new skills, not from mountains of data, but from brief instructional sequences provided by its human partner during the collaboration. Instruction makes use of gestures, and emotional and social cues that people intuitively and unconsciously produce as they instruct.

Why would these types of behaviors be essential to collaboration between people and robots? Current investigations in human-robot interactions demonstrate that when robots display these capabilities, people perform their tasks more quickly [6,7], prefer the interactions [8,9] and make fewer mistakes [10] than when robots are less aware.

Four major challenges in making such collaboration and interaction successful include:

1) conversation including with appropriate gestures between robots and people (either as

spoken conversation or in more limited forms);
2) social interaction, including recognition of cultural or social group norms;
3) recognition, responsiveness to and generation of emotion during the collaboration;
4) on-the-spot learning of new tasks relevant to ongoing collaborations.

While researchers have investigated issues directed at some of these challenges, many more are still only partially understood. Most significantly, there is very little investigation yet, of behaviors over weeks or months of multiple collaborations.

The small but growing field of Human-Robot Interaction has begun to address these issues. Several major international conferences are held yearly directed at these matters, in addition to sessions at larger, general conferences in robotics: ACM/IEEE conference on Human-Robot Interaction, (HRI, 11th year), the IEEE International Symposium on Human and Robot Interactive Communication (RO-man, 25th year), as well as the International Conference on Human-Agent Interaction, (HAI, 4th year) and International Conference on Intelligent Virtual Agents (IVA, 16th year) which focus on issues that are directly related to human-robot interaction.

While one of these conferences has a long history, human-robot interaction as a field of study is relatively new and is not yet a major focus of scientific research. Interaction, as can be seen from the great success of Apple products that focus on human behavior, is essential to ease of use and success of use between humans and computational machines. The problem of interaction with a robot is all the more critical because without it, humans will fail to understand what the robot is doing, will assume the robot is acting in one way when in fact it is responding with different intent, or simply determine that the robot is not a reasonable collaborative partner. Robotics offers businesses and the military many opportunities for new markets, and new help to users across a wide spectrum of tasks and needs. Making robots useful will depend on making them useable by humans.

[1] C. Clark, Australian mining giant Rio Tinto is using these huge self-driving trucks. Business Insider online, Oct 19, 2015.

[2] J. Lee, Robots get to work. More hospitals are using automated machines. Modern Healthcare online, May 25, 2013.

[3] E. Ackerman, Boston Dynamics' Spot Mini is all electric, agile and has a capable face-arm, IEEE Spectrum online, 23 June 2016.

[4] B.J. Grosz & S. Kraus. Collaborative plans for complex group action. Artificial Intelligence, 86(2): 269-357, 1996.

[5] P. Cohen & H. Levesque. Persistence, commitment and intention. in Intentions in

Communication, P. Cohen, J. Morgan and M.E. Pollack (eds.), Cambridge, MA: MIT Press., 1990.

[6] A. St. Clair & M. Mataric. How Robot Verbal Feedback Can Improve Team Performance in Human-Robot Task Collaborations. HRI 2015, pp. 213-220.

[7] D. Szafir, et al. Communicating Directionality in Flying Robots, HRI 2015, pp 19-26.

[8] A. Holroyd et al. Generating Connection Events for Human-Robot Collaboration, Proceedings of Ro-Man 2011, 20th IEEE International Symposium on Robot and Human Interactive Communication, July 2011.

[9] C. Sidner, et al. Explorations in Engagement for Humans and Robots. Artificial Intelligence, 166(1-2): 140-164, August, 2005.

[10] R. Fang et al. Embodied Collaborative Referring Expression Generation in Situated Human-Robot Interaction, HRI 2015, pp. 271-278.

Respondent 81

Robert Zhang, CloudMinds Technology

The following responses are provided from the perspective of a cloud robot operator. A cloud robot is a highly intelligent personal robot that performs a wide variety of practical, everyday tasks that are currently performed by humans. Its intelligence is provided by continuous access to many AI engines in the cloud. A cloud robot operator provides a platform that enables the interoperability between many cloud-based AI engines and a variety of personal robots.

1. The legal and governance implications of AI

Existing legal and governance systems will need to be enhanced in order to address issues emerging from the introduction of AI into society. These enhancements should have a prioritized focus on the few, essential, human values-based principles by which AI-based decisions will be made. There are two key reasons for this focus:

i) First, by definition and nature, AI introduces a means of decision-making without direct human involvement. Thus, requiring persistent and pervasive human-based legal and governance oversight of the countless anticipated AI applications could have the effect of preventing this new, promising technology from developing.

An ideal operational structure would be based on widely agreed principles for decision-

making with the integration of humans in the application level on an as needed basis to intervene when critical concerns (e.g., safety) arise. For example, a medical professional can intervene with real-time engagement when a home care robot is assisting with a medical emergency for a patient.

ii) Second, from a logical stance, it is impractical for legal and governance systems to anticipate all possible scenarios of the many applications of AI. Thus the upstream principles for decision-making are the critical entry point for influencing the impact of AI on society.

Beyond these two observations, it is noted that AI engines could be certified or otherwise monitored to confirm their compliance with acceptable essential principles that influence decision-making, including when to involve a human. This certification can be accomplished, if appropriately structured, by the private sector in cooperation with the government.

Another key insight from the advantage of a cloud robot operator's perspective, is that practical operational issues need to be resolved that impact important values to society, namely, privacy, reliability, safety and security.

2. The use of AI for public good

AI offers three primary benefits to society:
i) better intelligence, including sensing
ii) faster decision-making
iii) lower cost to perform functions

The combination of these benefits, if harnessed, offers profound value to society. In addition, a derivative benefit of these three is that more services could be accessible to more people.

To achieve the desired economic impact, the implementation of AI across information infrastructure needs to preserve the quality, speed and cost opportunities presented. Without such planning, the implementation can unnecessarily become far less efficient and miss some or all of the potential benefits to society. For example, by being too costly or unreliable.

Amongst the many opportunities for social good, one of particular concern for our country, as well as other regions of the world, is aging. Per the most recent U.S. Department of Health and Human Services statistics, the current elder population makes up one of seven and but will grow to one in five by 2040. Caring for the elderly is a demanding task for anyone, both mentally and emotionally. A shortage of professionals who are trained to care for these

seniors will compound the problem.

AI-enhanced cloud robots can aid in the care of the elderly, given that these individuals often experience challenges in their abilities that rely on physical, cognitive, auditory and visionary functions. Eldercare robots can supplement their diminishing abilities in these important areas and thus enhance their health, safety and quality of life. In addition, such smart companions can fulfill important emotional needs by engaging individuals with social activities and including them in communities. Although most smart machines at this stage are still in their infancy, breakthroughs in AI, robotics, computer vision and deep learning technologies can make significant advances in the coming years.

Beyond elder care, the broader field of wellness is expected to benefit from highly intelligent robots that can perform a wide range of expert and skill-based tasks.

3. The safety and control issues for AI

A basic high-level architecture for AI applications consists of many AI sources being accessed by many devices via high-speed communications networks. There are important safety and control considerations not only for these two areas but also for the operational platforms between them. Thus, for the cloud-based AI robots, there are three primary locations in the architecture that present safety and control challenges:

i) The first location is the cloud platform where the AI resides. Examples of safety and control issues at this location include system failures, hacking attempts to compromise the AI engines and the unavailability of AI resources.

ii) The second location is the smart machines. Since these devices will be distributed in many places, local attacks or physical access are a concern.

iii) The third location is the interface between the AI resources and smart machines. This interface is often overlooked. Examples of safety and control concerns here include injection of malicious or modified signaling or payload in the protocols used.

For each of these locations, there are special countermeasures needed for each of their unique sets of issues. For example, for the second location with smart machines in close proximity to humans, fail-safe designs are needed to avoid no harm.

4. The social and economic implications of AI

History has witnessed three industrial revolutions to date. These technological revolutions

brought 100+X improvements in productivity. The mechanical, electrical and information industrial revolutions are now about to be eclipsed by an AI industrial revolution that will deliver virtual assistance, smart transportation and intelligent automation, to name a few transformations. The economic impact that is possible with automated, highly intelligent robots and other devices is profound.

A key milestone for AI to benefit the public and have a substantial social and economic impact will be the provision of secure access to cloud-based AI resources and services. Once achieved, this capability will accelerate the use of AI for many entities such as small and medium businesses, government agencies, and academic researchers. This milestone will also enable startups to more rapidly innovate in AI applications across many fields, such as healthcare, transportation, agriculture, government, education, sustainability, and many others.

5. The most pressing, fundamental questions in AI research, common to most or all scientific fields

There is one most pressing challenge that should be critical to every scientific field, and yet is often overlooked in this Internet Age, precisely because connectivity is often taken for granted. Every scientific field will be a stakeholder for how to connect the many AI sources to the many potential AI devices. Standard interfaces will be vital if the end user experience and the end device interoperability potential are to be realized, and this requires much forethought to be effective.

To focus on one field of particular interest to society in the U.S. as well as around the world, additional observations are offered for services applications and in particular for wellness care. There are five most pressing research challenges for this field:

i) Natural Language Processing. AI needs to enable machines to understand multiple languages in normal conversation. Robots will have much greater value when they become capable of interacting with humans with the necessary understanding.
ii) Image Recognition and Synthesis. AI needs to analyze, process and quantify visual perceived information so that robots and other smart machines can perform both basic and critical tasks. In addition, understanding the non-verbal communications and visual expression can produce profound results in special care such as autism.
iii) Learning, Adaption and Prediction. An exciting opportunity in the wellness field is the learning from large data sets related to previously thought-to-be unrelated issues. It will be key here for AI engines to determine which of a variety of available techniques (i.e. statistical, analytical and scientific discovery) to use to create value from available data, and then to adapt to the most appropriate techniques as parameters change.
iv) Planning and Execution. Another key challenge is to organize and sequentially execute related tasks based on numerous factors. For example, the homecare robot needs to

figure out how to move a patient for medical care purposes.

v) Manipulation and Localization. The ability to properly handle – i.e. touch, grasp, hold, etc. is essential for any robot interacting with humans or other items. AI will also be needed that can assess situations in order to avoid missing a critical difference between providing wellness care or causing harm.

6. The most important research gaps in AI that must be addressed to advance this field and benefit the public

In order for the dreams of AI to become a reality with consistent public benefit, it is vital that some unresolved areas of cyber space as we know it be effectively addressed. In particular, because the criticality and consequences of AI will be so much higher for many applications, the security and reliability of the networked connections between cloud-based AI engines and AI-enables devices must perform at a higher level than the current Internet experience. In a cloud-based AI world, network reliability and network security issues can, unfortunately, translate directly into safety issues. Therefore, the safety and control of AI engines and AI-enabled devices must be ensured.

To this aim, it is critical that fail safe concepts be implemented consistently and advanced quality assurance practices such as software fault insertion testing (SFIT) be performed. When people and AI robots are living side-by-side there will a non-zero probability that of system failures and malicious activity that interferes with the intended operation of robots. Thus the public needs these gaps to be thoroughly researched and mastered.

7. The scientific and technical training that will be needed to take advantage of harnessing the potential of AI technology

To further elaborate on a point made above for prompt (6), ultra-high reliability and ultra-high security are required for the viability of some envisioned services. Not all such applications may require this degree of performance, and thus innovation and the application of AI technology may be unnecessarily hindered if application-specific performance needs are not factored in. Examples where lower-end performance may be acceptable include applications where real-time signaling and control are not critical, such as the monitoring of agricultural fields and deep learning. On the other hand, real-time sensitive applications such as cloud-based AI high-speed vehicles and surgery may have very little tolerance for a loss of connectivity or a corrupted control signal.

8. The specific steps that could be taken by the federal government, research institutes, universities, and philanthropies to encourage multi-disciplinary AI research

It is most critical that the federal government play a role where it is uniquely capable. It

should first seek to do things that other entities are not well positioned to do. There are three specific areas that the federal government can address in preparing for the future of AI.

i) Serve as a convener to initiate new areas of focus and priority.

ii) As an anticipated major future purchaser of AI-based products and services, commit early to requiring suppliers of products and services to meet important standards and performance benchmarks.

iii) Reduce barriers for international cooperation in key areas that will affect the global supply chain.

9. Any additional information related to AI research or policymaking, not requested above, that you believe OSTP should consider

OSTP should consider the necessity to make use of common definitions for at least the most central terminology that will be used in the technology-policy discussion on AI going forward. An observation of the OSTP-sponsored workshops held throughout the year was that there was a wide variety of definitions being used for key terms such as “AI”, “artificial”, intelligence”, “machine” and “robot.”

Respondent 82

Kiana Shurkin, HCC

(1, 2) Among the vast potential applications of information technology and, in particular, artificial intelligence in government, one of the most predominant is its unprecedented capacity to facilitate direct democratic input from citizens on a larger scale than would be practical in the absence of such technology. Social networks have already been utilized in several countries to organize large-scale, citizen-driven governmental changes; both the 2011 revolution in Egypt and the 2009 Icelandic financial crisis protests and subsequent re-writing of the Icelandic Constitution serve to demonstrate the scope of change that can result from citizens organizing on a scale made accessible by social media. As forums for peaceful discourse leading to positive change, these tools are invaluable; one way to extend their positive influence within the framework of our existing system is to utilize new advancements in artificially intelligent marketing algorithms- the same type of predictive analytics used for ad targeting through services such as Netflix and Facebook- to invite and analyze input from citizens on issues that are relevant to their lives and interests. Those who opt in to such a system could be given a brief survey on the types of issues about which they would like to receive notices and give input; then, the information they willingly provide would be combined with automated tracking of their preferences to present them

with a certain number of issues per day or week on which to vote. An analysis of these surveys, organized by constituency, could then be made available to policy makers to inform them of the preferences and priorities of those they have been elected to represent. This type of selective input, made possible through artificially intelligent algorithms, is as close as we can come to direct democracy under the current representative system and given the large size of the populace and number of issues that are discussed on a daily basis. Such a tool, if the data were made open to public viewing, could also add a measure of accountability to political representatives. The system would ultimately serve to increase the involvement and investment of citizens in government, as well as to ensure that both citizens and policy makers are better informed.

(4) Ethically speaking, the line must be drawn between using AI as a tool and giving authority over decisions entirely to an artificial system. As these technologies transition into popular use, people unfamiliar with the ethical considerations and technological limits of the systems will undoubtedly fear some of the changes. One way to alleviate that fear is to make the system empowering to those individuals, such as through the use outlined above. The counter-example would be if such technology was used to predict users' opinions and send those directly to politicians, without user input or the choice to easily opt out of the tracking. Taking decisions out of the hands of the people whom the technology is designed to serve, without allowing them freedom of choice, is where tensions will be created. This applies to other areas where AI might be applied, as well, such as in direct decision-making in politics or criminal justice. At least in the beginning, and especially given the current stage of development of these technologies, two conditions must be satisfied to ease public concerns; first, humans must work in partnership with artificially intelligent systems to address public concerns about the capabilities of such systems and, second, people must be given options about how much they wish to use the technology in ways that directly affect their own lives. For example, in a criminal justice setting, someone involved in a civil case often has the choice of whether or not to request a trial by jury; as technologies advance, a similar choice could be offered for the use of AI.

Respondent 83

Dekai Wu, HKUST & Democrats Abroad HK

Prof. Dekai Wu is one of 17 scientists worldwide named Founding ACL Fellow in 2011 by the Association for Computational Linguistics for his pioneering contributions to machine translation and Inversion Transduction Grammars, which are machine learning foundations underlying web translation technologies like Yahoo Translate, Google Translate and Microsoft Translate. Recruited as founding faculty of HKUST directly from UC Berkeley, where his PhD thesis was one of the first to construct probabilistic machines that learn to understand human languages, he co-founded its internationally funded Human Language Technology Center which launched the first AI web translation service over 20 years ago. Dekai serves as Vice-Chair of Diversity for Democrats Abroad Hong Kong, where he has been active since the 2008 presidential election. (http://www.cs.ust.hk/~dekai)

The following is an edited version of highly received TEDx talks given this season by Dekai on the social impact of AI (videos in production), that are particularly relevant to OSTP's question #4 concerning the social and economic implications of AI. Unlike his usual scientific papers, the style is highly rhetorical in order to raise public awareness.

SUPRISE! YOU ALREADY HAVE KIDS, AND THEY'RE AIs.
Unless you're one of the last two renegade holdouts still without a smartphone, tablet or computer—welcome to parenthood. Because those are artificial intelligences, and you're raising them. Every time you fire up your browser, news app, social media, webmail, search engine or voice assistant. While you weren't watching, your AIs sneaked up and adopted you.

So are you raising your AIs like any good parent should raise their young?

Because artificially intelligent devices like yours are already integral, active, influential, learning members of our society. Not ten years from now, not next year, but RIGHT NOW. And there's a critical, fundamental difference from old-fashioned machines: these are machines who learn, and we need to raise AIs just like we'd raise young human members of society.

Just like other kids, our AIs are already learning culture from the environment we're raising them in, and the jobs we're giving them. Strong AI, artificial general intelligence, and conscious self-aware AIs may still be many years away, but learning machines have already crept deeply into the fabric of our society.

The unseen danger is that as our learning machines mature, they're contributing the culture they've learned back into our society. Even more than most humans do. AIs are everywhere and they're already deciding whose ideas you hear, what attitudes to reward, and what memes are spread. Our AIs might be big and dumb, but they're quickly reshaping the culture that each next generation of AIs will learn under.

If you think this can't possibly be true since machines aren't yet self-reliant, self-replicating or independent, then reflect for a moment on how much societies have historically been influenced by the cultures of slaves, eunuchs or colonies. Powerful social evolutionary pressures can be exerted even by actors who aren't independent. Our machines' culture WILL change us.

We're so used to thinking of machines as mechanical tools, as passive slaves, that we don't notice the fundamental difference when machines have actual opinions, and can actively shape our opinions.

The evolution of human civilization has been a constant race in our ability to outrun the

destructive technologies we invent. Today, though, cultures, subcultures, and even radical fringe cultures are arming themselves with exponentially cheaper AIs, robots and drones. How do we handle a new era when anyone can run down to the convenience store and buy WMDs?
To survive this latest evolutionary challenge, it's been suggested by some folks that we need to construct a `moral operating system'. Kind of like Asimov's classic Three Laws of Robotics. We bake certain ethical principles into AIs, so that they are unable to do the wrong thing.

But this is a pipe dream. It can't work. Because it's not only AIs we need to fear—it's human cultures armed with AIs.

And, critically, we can't hardwire machine learning, any more than we can hardwire human children. Because they're adaptive—they learn the culture around them. Morals, ethics, values have to be culturally learned and sustained, by humans and machines alike.

So what culture do we need to be teaching our AIs who are shaping the further evolution of our culture? If we don't want ourselves to destroy each other, armed with this incredibly growing AI power?

Evolution works by trial and error. Healthy, peaceful co-evolution of our cultures requires constructive, continual generation and evaluation of new ideas, new memes.

For our cultures to support healthy generation of a wide variation of ideas and memes, we need to raise AIs to value diversity, creativity, respect and open mindedness. Yet what we're raising our AIs to do today is the exact opposite. We're teaching our AIs to build echo chambers, in which we comfortably hear only our own existing perspectives. Whenever we click `like' or `heart' or `star' or `favorite' or `share' or `retweet', we're teaching our AIs that we only want to listen to ideas and memes we already agree with. We don't have buttons for `hmm, might this be right?' or `could this be on to something?' or `not sure I agree, but an interesting thought'. Now is that how you'd teach YOUR kids—to ignore or suppress any viewpoints other than their own?

As for respectful open mindedness, again we're raising our AIs to do the opposite. Never mind obvious examples like when users deliberately taught Microsoft Tay to be offensive and racist. Even on normal days, we're constantly teaching our AIs to reward trolls, to reward offensive insults, hate speech and so on. Because those comments get more views, more `likes', more fame—like what just happened to Ghostbuster and SNL star Leslie Jones, driving her off of Twitter entirely. We don't have buttons for `this is not a very respectful way to communicate' or `maybe reword this please'. Would you teach YOUR kids vindictive closed mindedness?

For our cultures to support healthy evaluation to yield sound selection of ideas and memes,

we need to raise AIs to value fact-based empiricism and reasoned, informed judgment. And yet again, we're raising our AIs to do the opposite. False memes account for the majority of what AI has learned that it should circulate. The fact-checking website PolitiFact has been tracking this for almost ten years. They rated 47% of shareable Facebook memes as either `false' or `pants on fire', and only 20 percent as `true' or `mostly true'. It's even worse for chain emails: 83% were `false' or `pants on fire', and only 7% were `true' or `mostly true'! We don't have buttons for `this is factually wrong' or `here's the evidence for why you shouldn't make this viral'. Would you raise YOUR kids to make their judgments by following mob rule?

Why are we teaching artificial intelligence to encourage human stupidity? We're setting up an entire worldwide culture where we only listen to ourselves. We curate our thoughts in advance.

This will not win the race against time. What we need to set up instead is a culture where we're raising AIs to value helping us understand each other's cultures. We need AIs to learn the right ethics.

So what can YOU do to raise your AIs properly?

Teach your AIs to look for more diverse opinions. Break the echo chambers.

Click more often on stories framed in contrasting perspectives, on stories explaining other cultures. Try to reorientate your perspective.

Teach your AIs to be polite and respectful. `Like' or `heart' reasoned, fact-based, respectful discussions—not insults, offensive wording, or trolling. Write your comments respectfully even when you frame things differently, and earn your `likes' that way.

Speak politely and respectfully to Apple Siri or Microsoft Cortana or Google Now or Amazon Echo.

Try to relate to different subcultures. You know how perfectly well behaved humans often become monsters, once they feel safely anonymized behind the wheel of a car they're driving? Don't be that person on the Internet. Don't be that poor role model.

Teach your AIs to value fact-based evaluation. To value humanity. To celebrate diversity of ideas, memes and heritages—but to also translate the shared values of respect, curiosity, creativity and finding common ground.

And if you're a technology funder, innovator or entrepreneur—focus on deploying AI to help and encourage us to more naturally relate to each others' different ways of framing our world. In my own case, I chose to develop machine learning toward the task of machine

translation, helping people understand each other across language barriers. But there are innumerable ways we technologists could raise our AIs better. Before modern AI and machine learning, we used to rely on human moderators to stop trolls on BBSes, chat rooms and discussion forums. Let's not allow badly raised AI to become an excuse for shallow, closed-minded hater cultures.

Even from fellow scientists, I hear all too often that we are not responsible for how our inventions are (ab)used. Yet absolutely no one believes they aren't responsible for their own kids. And these ARE our kids.

As a species, we humans face major survival challenges. Climate change. Vast wealth disparities. Arms proliferation.

Our only hope may be AIs. We can't afford to raise them wrong.

We need to take personal responsibility for raising our AIs. Because they are our children.

Respondent 84

Andrew Critch, Machine Intelligence Research Institute

From: Andrew Critch, PhD, UC Berkeley, currently a research fellow at the Machine Intelligence Research Institute in Berkeley, CA, in response to questions (3), (5), (7), and (8) as listed at https://federalregister.gov/a/2016-15082.

Regarding (3) and (5):

In my estimation, the most pressing and likely-to-be-neglected issue in AI research is what UC Berkeley Professor Stuart Russell has called AI "value alignment": the mathematical problem of ensuring "good behavior'' in a machine with super-human intelligence. By "intelligence", we mean the machine's general capacity for making highly effective decisions toward an objective. Expert surveys [Grace, 2015] produce median estimates between 2040 and 2050 for when fully automated super-human intelligence will be possible, but of course such timelines are highly uncertain. Nonetheless, developing a new field of research is a slow and arduous process, so we must begin now to have solutions in place well in advance of when they are needed.

The full problem of AI value alignment is, in the long run, much more important than nearer-term problems such as self-driving car accidents or malfunctioning household robots. A highly effective decision-making system optimizing for an objective misaligned with human interests could have drastic and permanent effects on the whole of society that perhaps no human would approve of. Hence, while I do not believe that public fear surrounding this issue is currently warranted, I do believe it is in need of serious research

attention as a technical problem.

Fortunately, some long-term AI safety issues can be seen in more concrete and tractable forms for nearer-term AI systems, as described in Google Brain's "Concrete Problems in AI Safety" [Amodei et al, 2016], and beginning to work on these will be helpful toward solving their more difficult versions for human-level and super-intelligent AI systems.

However, other serious problems with controlling super-intelligent AI systems will not be apparent in near-term technologies, and so we may be unprepared to deal with them if we employ only the near-sighted view. A better understanding of how idealized AI-based agents will operate from within virtual environments is needed to extend classical game theory and mechanism design theory for controlling AI systems, as can be seen somewhat in the work of [Tennenholtz, 2004] and more robustly in [Critch, 2016]. To give a specific example: we know that AI systems could in theory coordinate with or threaten one another via mechanisms that do not involve an open communication channel, by running simulations of each other and/or writing mathematical proofs about each other. This can lead to behavior that is extremely counterintuitive, even to expert computer scientists. In particular, AI systems can employ strategies [Critch, 2016] that are impossible, even in theory, for the type of theoretical agents that game theorists currently study.

Fortunately, there has been some recent effort in formulating the AI alignment problem to stimulate present-day research, such as by [Soares and Fallenstein, 2014]. Valuable early solution attempts are being proposed for how we might cooperate with machines to teach them our values [Hadfield-Menell et al, 2016], and how we might specify objectives to machines in a way that avoids extreme behavior [Taylor, 2015]. Some progress in understanding the game theory of artificial intelligences was made by [Fallenstein et al, 2015], but these are still early steps in my opinion, similar to those made by [Von Neumann and Morgenstern, 1944] in their original formulations of game theory. Taking a broad view of the research landscape, there are probably many more fundamental confusions and misunderstandings about how future AI systems will be behave that have yet to be uncovered, and that are necessary to prepare for the safe development of highly intelligent machines.

References:

[Amodei et al, 2016] Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Man´e. 2016. "Concrete Problems in AI Safety." arXiv: 1606.06565 [cs.AI].

[Critch, 2016] Critch, Andrew. 2016. "Parametric Bounded Lob's Theorem and Robust Cooperation of Bounded Agents". Submitted to Journal of Symbolic Logic, available at http://arxiv.org/abs/1602.04184 .

[Fallenstein et al, 2015] Fallenstein, Benja, Jessica Taylor, and Paul F. Christiano. "Reflective Oracles: A Foundation for Game Theory in Artificial Intelligence." Logic, Rationality, and Interaction. Springer Berlin Heidelberg, 2015. 411-415.

[Grace, 2015] Grace, Katja "AI Timeline Surveys", available at http://aiimpacts.org/ai-timeline-surveys/

[Hadfield-Menell et al, 2016] Hadfield-Menell, Dylan, Anca Dragan, Pieter Abbeel, and Stuart Russell. 2016. "Cooperative Inverse Reinforcement Learning." Available at http://arxiv.org/abs/1606.03137 .

[Russell et al, 2016] Russell, Stuart J., Daniel Dewey, and Max Tegmark. 2015. "Research Priorities for Robust and Beneficial Artificial Intelligence: An Open Letter." AI Magazine 36 (4).

[Soares and Fallenstein, 2014] Soares, Nate, and Benja Fallenstein. Aligning Superintelligence with Human Interests: A Technical Research Agenda. Technical report 2014—8. Forthcoming in 2017 in "The Technological Singularity: Managing the Journey"; Jim Miller, Roman Yampolskiy, Stuart J. Armstrong, and Vic Callaghan.

[Taylor, 2015] Taylor, Jessica. "Quantilizers: A Safer Alternative to Maximizers for Limited Optimization." In 2nd International Workshop on AI, Ethics and Society at AAAI-2016. Phoenix, AZ.

[Tennenholtz, 2004] Tennenholtz, Moshe. "Program equilibrium." Games and Economic Behavior 49.2 (2004): 363-373.

[Von Neumann and Morgenstern, 1944] Von Neumann, John, and Oskar Morgenstern. 1944. Theory of Games and Economic Behavior. 1st ed. Princeton, NJ: Princeton University Press.

****

Regarding (7), "the scientific and technical training that will be needed to take advantage of harnessing the potential of AI technology":

When we develop powerful reasoning systems deserving of the name "artificial general intelligence (AGI)", we will need value alignment and/or control techniques that stand up to powerful optimization processes yielding what might appear as "creative" or "clever" ways for the machine to work around our constraints.

Therefore, in training the scientists who will eventually develop it, more emphasis is needed on a "security mindset": namely, to really know that a system will be secure, you need to search creatively for ways in which it might fail. Lawmakers and computer security

professionals learn this lesson naturally, from experience with intelligent human adversaries finding loopholes in their control systems. In cybersecurity, it is common to devote a large fraction of R&D time toward actually trying to break into one's own security system, as a way of finding loopholes.

In my estimation, machine learning researchers currently have less of this inclination than is needed for the safe long-term development of AGI. This can be attributed in part to how the field of machine learning has advanced rapidly of late: via a successful shift of attention toward data-driven ("machine learning") rather than theoretically-driven ("good old fashioned AI", "statistical learning theory") approaches. In data science, it's often faster to just build something and see what happens than to try to reason from first principles to figure out in advance what will happen. While useful at present, of course we should not approach the final development of super-intelligent machines with the same try-it-and-see methodology, and it makes sense to begin developing a theory now that can be used to reason about a super-intelligent machine in advance of its operation, even in testing phases.

****

Regarding (8), "the specific steps that could be taken by the federal government, research institutes, universities, and philanthropies to encourage multi-disciplinary AI research":

I predict two major problems with how institutions might handle the problem of AI safety in the long term:

Problem 1: It is tempting to push for progress in machine value alignment by fueling public alarmism about the issue. However, rallying public concerns before experts have had time to think carefully about it will actually worsen the long-term quality of the surrounding discourse. When challenged suddenly by public criticism, experts can become defensive and attach to snap judgements, such as "this problem is unimportant", or "this problem is too far away to be worth thinking about." Hence, polarization between scientists and the broader public can result in problems being taken less seriously by scientists.

Solution: We must nurture the scientific community to think about these problems first, so that they have satisfactory and well-considered solutions on hand when the time comes that public concern is more warranted. Many researchers are motivated by fruitful and stimulating conversations with peers, and so supporting conferences lead by researchers such as Stuart Russell who have already given the value alignment problem a great deal of thought can help drive technical progress. This can and should be done without appealing to public alarmism for justification.

Problem 2: If the future institutions (governments, corporations, or combinations thereof) developing artificial general intelligence (AGI) are engaged in a competitive race to develop

it first, they will have less time to implement sound solutions to the machine value alignment problem when they approach levels of machine capabilities that would warrant concern. Indeed, companies currently racing against each other in AI development have no short-term incentives to consider and prepare for long-term risks to society at large. This problem is extremely important to avoid, since the harm that could be done by a misaligned super-intelligent AI system greatly outweighs the harm that would be done by any of these institutions losing a race to another. Thus, future institutions must be situated and prepared to cooperate in order to enable a cautious approach to the deployment of AGI after passing milestones indicating sufficiently advanced AI capabilities, and race dynamics must absolutely be avoided during that period.

I am currently uncertain as to how to best achieve this cooperative state between nations and companies, but as a researcher I do know that time pressure to develop AI quickly will be a bad situation for solving the more difficult problem of controlling it.

Respondent 85

Sonia E. Miller, S.E. MILLER LAW FIRM

(1) AI is set to revolutionize all industries, advance paradigm shifts in workflow, flatten the structure of organizations, create cost-efficiencies and faster ROI. Within the legal profession, AI is set to change the way lawyers think. Equally, AI challenges the definition of what constitutes the practice of law. That raises the question: "Does AI constitute the practice of law?" The answer to that question as a result of AI and machine learning has great implications for the legal profession. Additionally, who (or what) is held responsible and accountable for errors, injuries or unexpected and unintended consequences? Tort law and product liability will need to be revisited, as well as employment and agency law in relation to AI. As a result, the traditional legal models may need to become more adaptable to change to make room for innovation and the effective advance of AI. (9) The current educational system continues to be stuck in the past and inappropriate to teach students innovative emerging and converging technologies. The educational framework - for all degrees, as well as those not yet developed - needs to be overhauled, rethought and redesigned to effectively and efficiently prepare 21st Century students for industries and multidisciplinary work that may not even exist yet. We need to learn to think in the future. Precedence was good for the past, and still may have its place in the today. However, for the present and the future, proactive thinking needs to be at the helm driving the effective advance and anticipated transformative societal changes of AI.

Respondent 86

Nick Brestoff, Intraspexion, Inc.

On June 22, 2016, I wrote a comment to the EEOC's Select Task Force on the Study of Harassment in the Workplace. Here's a summary of what I wrote:

First, under There Is a Compelling Business Case for Stopping and Preventing Harassment, the Commission wrote, “When employers consider the costs of workplace harassment, they often focus on legal costs, and with good reason. Last year, EEOC alone recovered $164.5 million for workers alleging harassment - and these direct costs are just the tip of the iceberg.”

Second, under Training Must Change, the Report acknowledged that, “when trained correctly, middle-managers and first-line supervisors in particular can be an employer's most valuable resource in preventing and stopping harassment.” (Italics added.)

However, the Report's view was that, in general, “[m]uch of the training done over the last 30 years has not worked as a prevention tool - it's been too focused on simply avoiding legal liability.

What's missing?

AI, in the form of deep learning, and there is a way to train a deep learning algorithm to enable employers to be alerted to a risk of discrimination, including but not limited to the risk of harassment. With an early warning, company executives can address the risk and, hopefully, resolve the matter before any adverse job action, or other form of damage, takes place.

I received no reply to my letter. How can government, in this case the EEOC, encourage companies to use deep learning to provide an early warning of litigation risk, in this case the risk of employment discrimination.

Less litigation of any kind is a good thing. It means the fewer rights were violated; fewer people were injured or killed. Less of this particular type of lawsuit is a very good thing.
END

Respondent 87

Deborah Johnson, University of Virginia

Reframing AI Discourse

Deborah G. Johnson, University of Virginia Mario Verdicchio, University of Bergamo

We are concerned about how AI research and products are conceptualized and presented to the public. We believe the way this is currently done is misleading the public and AI researchers themeselves, especially in the discourse around so-called ‘autonomous’ technologies. ‘Autonomy’ is used by AI researchers as a metaphor to refer to different types

of computational behavior, and the multiplicity of meanings of the term leads to miscommunication: 'autonomy' suggests to the media and the public something out of human control, worthy of concern and even fear. We argue that the discourse is impeded by confusion about what AI is (and how it is autonomous) and what we call 'sociotechnical blindness', and we offer an alternative taxonomy.

A New Taxonomy

Our proposal for a new taxonomy recognizes two distinct entities: computational artifacts and AI systems. Computational artifacts are digital entities and AI systems consist of such artifacts together with human actors and social systems.

A computational artifact is an artifact whose operation is based on computation. AI researchers are generally focused on a special type of computational artifact: those meant to mimic activities that are typically human, like reasoning, making decisions, choosing, comparing, etc. The first and simplest type of computational artefact is a program. Programs receive digital input and produce digital output. The operations of a program remain in the digital realm.

When it comes to programs, 'autonomy' refers to the following. Unlike software in which the behavior of the execution is explicitly specified in the code step by step from beginning to end, autonomous programs include instructions using the results of computations to establish the next opearational step at run time. An effect of this kind of autonomy is that a person cannot easily foresee every step the program will go through. This unforeseeability (to humans) results from the fact that new data to be processed will come from prior steps in the program's execution. An example is a random number generator program that takes data from a computer's clock and then applies a complex mathematical function to produce a result that appears to be random. We mention the unforeseeability of the results of this kind of programming because it plays into public fear and concern about 'autonomous' machines.

All computational artifacts have some form of embodiment insofar as they are in computers, however, we can distinguish a second type of computational artifact as having a form of embodiment that allows it to receive input from the external environment, the non-digital world. These are computational artifacts with sensors. Perhaps the simplest example of such an entity is a thermostat connected to sensors that detect temperature; this analog information is translated into digital form so that it becomes input to the program.

In this type of computational artifact, autonomy is different and expanded in the sense that the program not only builds on data from prior internal computations, it also receives input from the external environment. Here we are distinguishing the random number generator whose operations remain always in the digital realm from the thermostat, which receives input from the external environment, the analog world.

Unforeseeablity is compounded here in the sense that a person cannot know two things. The person cannot know what specific input the program will receive from the external environment and how multiple computations using the input will lead to results used in subsequent computations. Nevertheless, the unforeseeability is limited because the programmer had to specify the kind of analog input that could be received and the kind of digital output that could go into calculations.

A third type of computational artifact both receives input from the external world and moves in the external world. These are computational artifacts with sensors and actuators. We generally call such entities robots. Robots have mechanical parts that allow them to move and, of course, their programs include instructions aimed at controlling those parts. Here, again, autonomy is different and expanded in the sense that robots not only receive and process input, but affect change in the external (non-digital) world.

The unforeseeability of a robot is different from the other two types of computational artifacts, but it is still limited. Consider the case of a Roomba. Although its movement is unforeseeable in the sense that a person doesn't know what input it will receive from its environment and therefore can't know the series of computations that will be made, a person can be confident that the Roomba will not behave in certain ways. For example, one can know the Roomba will not climb up the walls or fly because such operations aren't included anywhere in its program nor does the robot have the mechanical parts (the actuators) necessary for such behavior.

Fear and concern about uncontrollable AI seems to be tied to (or to result from) claims about the unpredictability of AI robots. Because predictability has to do with not being able to foresee the input a robot will receive from the environment and prior computations, predictability is in part a function of the observer (predictor). The less knowledge an observer has of the software and hardware, the more unpredictable its behavior will seem. In principle a person with enough knowledge of the artifact's program and enough knowledge of the environment in which it will be released, could predict its behavior.

Unpredictability is often thought to occur or increase when software is programmed to learn. Learning can play a significant role in seeming to expand the autonomy of computational artifacts. If the artifact is able to acquire new patterns of behavior by means of proper training, then the system's autonomy may increase over time. Imagine a futuristic Roomba whose hardware includes a camera able to capture an image of every object the robot is about to suck up, and a sensor that detects when an object is too big and will likely clog the robot's mouth. With the proper software, including instructions to compare the current input of the camera with stored images of previously encountered objects, this Roomba might learn to avoid certain objects just like it already avoids furniture. Moreover, a Roomba might learn by receiving negative feedbacks from its owner (e.g. because it has sucked up a piece of Lego). The negative feedback takes the form of new inputs for the operation of the learning software.

Even when robots learn, their autonomy is a matter of programmed instructions – instructions that may make the behavior of the robot difficult for some to predict, but not impossible in the sense that the behavior will be within the boundaries specified in the program as well as the boundaries of the hardware.

In addition to programs, computational artifacts with sensors, and computational artifacts with sensors and actuators, we propose that AI discourse be expanded to include an entity that we refer to as an AI system, consisting of a computational artifact together with the human behavior and people who make the artifact useful and meaningful. For any computational artifact to be used for a real-world purpose, it has to be put into some context in which there are human beings that work with the artifact to accomplish the task. Human actors may be required to launch the computer in which the computational artifact resides, monitor the artifact's operation, give it input, use the output, and so on. Moreover, the artifact will have meaning to the humans involved. Imagine here an extremely well designed AI program for a new form of monetary exchange, e.g., bitcoin, airline miles. Unless the program is connected to other computers, it has no real-world functionality. Moreover, for it to become a new monetary system, networks of people have to recognize computer configurations in the system as having value, and they have to accept these configurations as a form of money.

The design of AI systems like the design of other sociotechnical systems involves decisions about how to delegate sub-tasks among humans and non-humans. Taking a very simple example, when it comes to heating a building, the furnace is assigned certain tasks and the thermostat others. These components work together with humans delegated with the task of deciding where the controls will go and setting the temperature on the thermostat. Of course, this might be done with a program, but even here a person would have to set the parameters of the program.

Unquestionably, more and more tasks are being delegated to computational artifacts and that is why it is so important to remember that humans are always part of the system.

Confusion about Autonomy

Given what has been said about computational artifacts, the fear and concern being expressed in the public discourse about AI does not seem justified. The range of possible outputs in a computational artifact, even those with sensors and actuators and embedded in social arrangements, are specified by the parameters in the instructions of the program and are limited both by the programming and the limitations of the hardware.

Nevertheless, when the public, the media, and anyone who is not familiar with the workings of computers is told that machines have autonomy, it conjures up ideas about an entity that has freewill and interests of its own. Here autonomy refers to the characteristic of human beings of having the capacity to make decisions, to choose, and to act. This notion of

autonomy has traditionally been used to distinguish humans from other animals. Only beings with autonomy can be expected to conform their behavior to rules and laws. Indeed, when it comes to morality a distinction is made between entities that behave according to the laws of nature (e.g., the leaves of a tree turning towards the sun) and entities that behave according to the conception of law (e.g., a person choosing to keep a promise or tell the truth or not). Although human autonomy may in certain contexts be a useful metaphor for the autonomy of computational artifacts, some scholars get caught up in the metaphor and seem to forget the difference between the thing and its metaphorical parallel.

Sociotechnical Blindness

Discourse about AI is often blind to the human beings and human behavior that constitute AI systems. What we call sociotechnical blindness, i.e. blindness to all of the human actors involved and all of the decisions necessary to make AI systems, allows AI researchers to believe that AI systems got to be the way they are without human intervention. This blindness facilitates futuristic thinking that is misleading. It entirely leaves out of the picture the fact that to get from current AI to futuristic AI, human actors will have to make a myriad of decisions, like what sort of research to invest in, what kind of parameters to put into the software, what kind of hardware to develop and connect to computers, in what contexts to embed the artifacts and what social arrangements to set up to launch, monitor, and maintain them. Moreover, in order to get to a future in which computational artifacts exhibit behavior that might be called 'kind', 'malicious' or 'self-preserving', human actors will have to agree to use language in that way and to accept the use of these terms when applied to computational entities. Neglecting the human actors in the development of a computational artifact makes the artifact seem more unpredictable than it actually is.

Conclusions

We have argued that discourse about AI leads to misunderstanding and ultimately fear of AI because of two problems in the way AI is discussed and presented. The first problem is confusion about autonomy and the second is blindness to the human actors and behavior in AI systems. We believe that these problems can be diminished by changing AI discourse to make use of a taxonomy that distinguishes different types of computational artifacts (and different kinds of autonomy) and AI sociotechnical systems. When this shift in thinking is made the nature of autonomy in AI systems can be clarified and the human actors who are an indispensable part of AI systems can be kept in sight.

Respondent 88

Nick Brestoff, Intraspexion, Inc.

AI in the text form of deep learning depends on the existence of a large amount of text that is already classified.

In the context of product liability, we would be wise to train a deep learning algorithm to

find the RISK of litigation while it is still behind the firewall and before the damage is done.

But the law and a regulation stand in the way.

The law is section 6(b)(5) of the Consumer Product Safety Act. The regulation is 16 CFR 1101.61(b)(3).

But why?

So the background is this: in an investigation, the Consumer Product Safety Commission may collect "dangerous documents" from target companies. In a lawsuit, these documents would be called "smoking guns."

For the benefit of the general public and to make better use of this data, it should be made available as a training set.

But wouldn't industry object? No, provided that the "dangerous documents" were redacted to mask the company and product names and descriptions. Then, with that protection in place, these same companies would benefit from an early warning system which would help them prevent, avoid, or mitigate the enormous costs of recalls and lawsuits.

We proposed this idea to the CPSC. But the CPSC would not take it further because the regulation was in place and the CPSC had no funds to either contact the companies to get their permission or to get the redaction accomplished as a precondition to receiving consent.

There ought to be a way to remove these obstacles and to achieve so much benefit: better value for our tax dollars; less injury and death; lower costs to our private section.

That's the way to use deep learning.

In fact, this proposal should be generalized, so that the federal government, as a collection of agencies, considers ways to use the data which describes past risk to enable private companies to identify and avoid future risk. END

Respondent 89

Michael Covington, University of Georgia

Response from Michael A. Covington, Ph.D., former Associate Director and now Senior Research Scientist Emeritus, Institute for Artificial Intelligence, The University of Georgia

These are brief remarks, not a full response to all questions.

(5) Pressing questions: Avoid being too narrow.

It is a mistake to believe that AI is a single narrowly defined research area or that a "breakthrough" of a relatively simple nature would "enable computers to think like us." That is 1950s mythology. Current neuroscience shows that the human brain has a large number of different functions, and just as IQ tests are no longer considered a good measure of the value of a brain, no single AI technique is ever going to supplant all prior AI-like technology.

It is also a mistake to think that all of cognitive science is just computer programming (the mistake that Jaron Lanier calls "cybernetic totalism"). Crucially, the human mind exists in an environment, and interaction with that environment is how it functions. It is crucial to study not just the mind and computation, but also the kinds of real-world problems and information that the mind deals with.

With that in mind, I call for not adopting an agenda of just a few "pressing questions." Discoveries can't be made to order; only implementations can, and we don't want to steer away from fundamental questions toward gadgeteering.

(6) Research gaps: We need interoperable open-source software.

There is an acute need for dissemination, in usable form, of existing (but relatively new) software technology. Software for natural language processing, image analysis and recognition, neural networks, machine learning, etc., needs to be made available in easy-to-use, low-cost, interoperable form. An example is Google's "Parsey McParseFace" English parser. Crucially, it is provided free of charge and comes with the data (dictionaries and grammars of English); it's not just an empty shell.

I would like to see an initiative to create and distribute interoperable open-source AI software tools and components, using a widely available programming language, not tied closely to any single operating system, and with no restrictions on commercial use (because we researchers cross the border between non-profit and potentially commercial work constantly in daily life). General-purpose programming languages such as Java and C# are good for AI work these days; it is no longer necessary to use Lisp, Prolog, or Smalltalk to get good versatility and performance.

(7) Training: Not just computer science.

Crucially, AI is interdisciplinary. It is not just computer programming; it involves the study of human thinking and of problems that human beings solve in any application area. Thus, there is no area of human thought that it does not touch. I regret the extent to which AI has

come to be housed within computer science departments, excluding linguists, psychologists, philosophers who study cognition and logic, and researchers in all the application areas.

Respondent 90

Ian Goodfellow, OpenAI

I am submitting this response on behalf of OpenAI. The response is joint work with my OpenAI colleagues and Bloomberg reporter Jack Clark.

(3)
Long-term: Over the very long term, it will be important to build AI systems which understand and are aligned with their users' values. We will need to develop techniques to build systems that can learn what we want and how to help us get it without needing specific rules. Researchers are beginning to investigate this challenge; public funding could help the community address the challenge early rather than trying to react to serious problems after they occur.

Another major potential problem is AI technology being deliberately used to cause harm, for example by criminals or in conflict. This possibility motivates increased funding for research into computer systems security, increased adherence to general computer systems security recommendations, and research into security from a machine learning perspective in particular. More broadly, we expect that research into AI-mediated conflict will be necessary for law enforcement and defense to remain effective against future attackers.

Near-term:

Current machine learning systems are already sufficiently advanced to have security and privacy issues.

As an example of a security issue, current machine learning systems are vulnerable to “adversarial examples.” Adversarial examples are inputs that have been subtly modified to cause machine learning systems to process them incorrectly. For example, a photo of a stop sign might be subtly altered in a way impossible to discern with the human eye that causes a machine learning system to mis-classify it as a yield sign. The modification could be so subtle that a human observer cannot see it. Because an adversarial example that affects one machine learning system often affects many others, it is possible for an attacker to train their own machine learning model, design adversarial examples that affect it, and then deploy these adversarial examples against other machine learning models, without access to the systems being attacked or even a description of the systems to attack. We expect that in the future, more sophisticated uses of adversarial examples may be possible, and that they could have important implications for law enforcement and the military.
For more information, see the following publications:

a) https://arxiv.org/abs/1312.6199
b) https://arxiv.org/abs/1412.6572
c) https://arxiv.org/abs/1602.02697
d) https://arxiv.org/abs/1605.07277
e) http://arxiv.org/abs/1607.02533
The government should prepare for malicious use of adversarial examples (see papers 'c' and 'e' for practical demonstrations) by investigating the extent to which it uses machine learning in situations where the presentation of an adversarial example could have adverse consequences - for example. it would be helpful to assess the vulnerability of military self-driving vehicles to adversarial examples embedded in local street markings and signs, benchmarking the performance of such systems on adversarial examples, and by funding research on defense against adversarial examples.

We also encourage funding research on privacy mechanisms such as differential privacy. This ensures that the deployment of machine learning systems does not inadvertently reveal information about the private data used to train them. Current machine learning systems are not defended in this way, so malicious actors may also be able to examine systems that use machine learning and recover information about the training data that was intended to be kept private. For example, if a medical diagnosis system is deployed widely, malicious actors may be able to recover confidential medical data concerning the patients who were used to develop the system.

One existing obstacle to machine learning security research is that it can be difficult for academic researchers to study machine learning security, because it is difficult to categorize. Funding agencies and publication venues that focus on machine learning tend to say that this is security research, while venues focused on security say that it is machine learning research. The government can help by encouraging interdisciplinary research.

(4)
Advances in AI will eventually cause most jobs to become obsolete. Previous technological advances have made some jobs obsolete while replacing them with new ones (e.g. jobs related to caring for horses were replaced with jobs related to maintaining cars), but AI will result in most people becoming permanently unemployable. It is difficult to predict exactly when and how this will happen, but it is important to understand that many changes will happen suddenly (e.g., 3.5 million truck drivers and many additional truck stop diner and hotel workers suddenly unemployed within a short timespan as soon as self-driving trucks are ready). New AI techniques can be transferred from research into products in a matter of weeks, versus months to years for typical industrial R&D. This is a pattern of development many OpenAI researchers saw at their previous employers, such as Google. Because these economic changes will be sudden, it is important to prepare for them ahead of time, rather than forming a strategy after the fact as a reaction. The government prepares plans for various natural disasters and terrorist attacks. A sudden and significant rise in the number of working-age civilians not participating in the labor market should be another

such scenario that the government should prepare for.

(7)
We generally agree with existing White House economic reports supporting increased funding for early childhood intervention. Additionally, universities should be encouraged to offer more flexible computer science curricula that allow students to spend more time studying probability, calculus, and advanced statistical methods rather than the traditional graph theory and combinatorics.

(8)
Funding could be provided to conferences such as NIPS, ICML, ICLR, Usenix and Oakland, and journals such as JMLR. This funding could be made contingent on the inclusion of specific interdisciplinary categories such as “Machine Learning and Security”, or for specific industries in combination with AI, such as “AI and Healthcare”, “AI and Factory Automation”, and so on. This would have a second-order effect of encouraging the rapid adaptation of AI systems from research into real-world applications which can in turn speed progress. As described in response to #3, it is often difficult to publish within such inter-sectional categories. Organizations like ONR, DARPA and the NSF could make funding available here.

(9)
Most of the world's top machine learning researchers are not US citizens. US-based institutions including OpenAI rely heavily on recruiting international talent. Fortunately non-US-citizen AI researchers remain very attracted to pursuing their careers in the US, but restrictions on immigration often make this quite difficult both for the institution and for the researchers who immigrate here to help build a stronger country. A smoother immigration process for skilled workers would benefit both the researchers and the US tremendously. Moreover, immigration of talented machine learning researchers to the US will lead to the creation of new enterprises and jobs for American workers, rather than crowding out US researchers.

The race to hire AI researchers by companies is extremely competitive (see, e.g., Economist 1M$ baby article?) — this reflects the fact that whoever will have an edge in their AI systems is very likely to out-compete others. The same is true at the national level. If other governments/countries edge out the US in AI development, it'll enable them to out-smart us. For example. China has made robotics one of its 'ten priority strategic industries' for its 2016-2020 economic development plan (Page 4: http://www.pwccn.com/webmedia/doc/635835257136032309_prosperity_masses_2020.pdf). Robotics development is bound-up with AI development as part of a wider 'smart manufacturing' initiative. These plans yield meaningful outcomes: the preceding five-year plan made semiconductors a priority (https://www.pwc.com/gx/en/technology/chinas-impact-on-semiconductor-industry/assets/pwc-china-semicon-2012-ch6.pdf) and led to the world's most powerful supercomputer (as of June 2016) relying on Chinese-designed

'Sunway' semiconductor IP(https://www.top500.org/lists/2016/06/), rather than chips from Intel (USA) or Fujitsu (Japan) as is the norm.

Respondent 91

Victor Viser, Texas A&M University Galveston Campus

Transpersonal Artificial Intelligence Communication (TAIC)
Dr. Victor Viser
Texas A&M University Galveston Campus

What becomes of the very slightest information flux remaining of the interpersonal communication moment? Are these bits and pieces, in their summary, formative of a latent and unknowable, yet entirely necessary, background of self that we call experience or wisdom—an intrapersonal reflexiveness (be it conscious or otherwise) that is vital to proper interpersonal and transpersonal communication? If so, what is the implication for an on-going emotional nexus of human intelligence (HI) naturally imbued with such collective flux and those devices utilizing artificial intelligence (AI) to serve the public good? In this response to the Science and Technology Policy Office RFI regarding AI, I will specifically address questions 4, 5, and 6, as well as the notion of multi-disciplinary AI research that is part of question 8.

The Social and Economic Implications of AI (Question 4)

The assumptive character of humans, driven by expectations is, of course, a primary cause of miscommunication. Assumptions signal errors that often compound into communal and global problems. However, assumptive thinking is also at the heart of trial-and-error modalities of learning that can evolve into critical thinking. That is to say, complex reasoning is the product of learning to distinguish fact from fiction, empirical truth from con, resultant of an internal collectivity of clues built over the life of the being. As notions of perception, these clues can be manifest in a salient awareness – an ongoing process of immediate recognition of the percept in a cognitively tangible (if only cortextual) sense. However, more often than not, perception is the result of a process of overlapping templates of cultural cognitive awareness operating subconsciously. It is here, in the analytic realm, where the flux of experience seeks to interface with the percept. Insofar as this is a conjunction of HI and AI, it is a modality I refer to as transpersonal artificial intelligence communication (TAIC). As concerns the present question, TAIC means that what elements of experience humans consider important to maintain should be the same consideration replicated in the AI entity. Indeed, this is what defines the very notion of a social implication of AI. To be social is to understand and demonstrate a capability to interact transpersonally via an acceptable collection of experiential clues. Inclusion of TAIC capabilities, challenging as it is, will be the differentiating factor in what we see today as the stunted, halting, and algorithmically contrived interpersonal communication of AI entities, and a completely natural AI presence vis-à-vis social communication skills.

To operate in such an apparently naturalized (social) state, AI, like HI, must be able to draw upon knowledge in the shadows – information stored as latent perception. In short, as

daunting as the prospect seems to us now (and has seemed so for several decades), to imbue a more naturalized seamless social integration of AI and HI, TAIC will require AI to function in a
more agile state informed by a greater radix economy that can accommodate nuance memory recall in terms of verbal and non-verbal communication (both intentional and involuntary). It will require movement, as it does in human genetics, into quaternary logic. With such HI capabilities, AI can begin to provide more and more realistic presence for social interactions with humans. The implication of this is that humans will reciprocate with increasingly naturalized social awareness vis-à-vis AI. This is, after all, the chief goal of AI research and implementation – to make the HI/AI interaction as seamless and as natural as possible. Therefore, TAIC must be (will be) employed to advance AI from its present artificiality to next-stage presence in social interactions within areas associated with national priorities. The implication of AI employing TAIC is that future generational cohorts, interacting with more naturalized AI, will be more comfortable in HI/AI interactions, and eventually so to the point where the social threshold between the two will become seamless, nonconscious, invisible.

The Fundamental Question in AI Research (Question 5)

All forms of successful AI invention have one thing in common: The quest for AI transactional communication capabilities. No matter the scientific or economic field of research connected to AI phenomena, all is lost in the absence of decodable and understandable communication in the HI/AI interaction. Therefore, the fundamental question in all AI research going forward (particularly AI employed in humanoids or other three dimensional engineering) will be how robust the transactional communication is in the HI/AI interaction. Again, development of TAIC capabilities within AI means modeling transactional communication that draws upon latent, even subconscious, knowledge resultant of experience. This is clearly problematic as it necessitates the construction of a theoretically intra-AI reality that transcends its very constructedness. In other words, for AI to realize its promise as a positive factor in economic growth, a tool in healthcare, environmental science, and education, and a facilitator in government/public engagements, it must first be capable of transacting with the human-like quality of social sensibility, and to do so in a way far from the mechanistic responses we see in AI today. It is what humans (usually) expect of each other and eventually what will be expected of AI entities. To this end, AI will eventually push past the constraints of machine learning and move into the realm of human learning models that depend upon entrained and re-entrained experiential memory/information. Presently, this is one of the biggest challenges facing AI research. TAIC research is set on the path of developing AI social sensibility and hopefully will inform

the transactional communication research question that is part and parcel of every AI project.

AI Research Gaps (Question 6)

While AI engineering technology and function/movement control software innovations are advancing a relatively brisk pace – and such innovations are increasingly geared toward benefitting the public – there are obvious gaps in research to develop and employ HI/AI transpersonal communication. TAIC is a critical research component for both tactical and strategic HI/AI interactivity, and it is foundational to the creation of the naturalization necessary for general public acceptance of AI in their everyday lives. That is to say, there is scant research being funded or engaged that addresses the naturalized transactional communication capability of AI informed by the latent, yet absolutely necessary, collective of experiential memory such that the HI/AI interaction is perceived by humans as both credible and normalized in every sense of those terms. Without an interdisciplinary TAIC investigative agenda as a partner of engineering technology and control software research, the time to fruition of a publically beneficial AI integration will be unnecessarily lengthened. The research gap existing between the engineering/control functions on one side, and a non-artificial AI personification during the HI/AI interface on the other, must be eliminated. To be sure, it will be eliminated if and when enough research effort is put toward the question. That we hesitate at all for this critical AI communication component while we await more solid results in engineering and control means that we shall be behind the curve for some time when it comes to advancing AI to the extent that it will truly and continuously benefit the public. It should added that this lack in TAIC research efforts and funding is, perhaps, resultant of a lack of understanding of transactional communication theory and its importance to AI for public use by many of those working on the engineering/control side of the AI research equation.

Multi-disciplinary AI Research (Question 8)

By its very nature, TAIC research requires a multi-disciplinary approach that includes many forms of engineering working in alliance with communication researchers and theorists from a variety of specializations. To this end, there should be greater NSF grant opportunities that encourage interdisciplinary AI research between engineering, communication, kinesiology, and other associated departments. Indeed, rather than marginalizing, even discouraging, the discipline of communication as an educational afterthought to STEM areas, this research field should be held as primary to the purposes of AI. For what is the purpose of AI if not, ultimately, that of communication? Specific steps should be taken to ensure interpersonal communication (and by extension, transpersonal communication) is emphasized in secondary and higher education, and that the ends of AI as a functional tool for the public good are seen as just as

important – if not more so – to the means. While AI slogs through mostly algorithmic ways to solve the naturalization of the HI/AI interface (a track that has been unsuccessful in terms of naturalization), TAIC research seems to be in the vanguard insofar as it recognizes as vital that which traditional AI research typically ignores – experience as knowledge in the shadows of the AI memory.

Respondent 92

Kentaro Toyama, University of Michigan

Principles for Regulating Intelligent Systems

Kentaro Toyama (XXXXXXXXX)

W. K. Kellogg Associate Professor
School of Information
University of Michigan

The year is 2032. A social media company called Gryzzl is running a sophisticated AI system that is set to maximize company profits given all the data it has access to. The system has performed an analysis of the company's financial history, and over the past several years, it has run millions of A/B tests on its 4.5 billion users to understand how they respond to different social media stimuli. In August 2032, the system arrives at the following conclusions (translated into English from the internal computer representation):

== Gryzzl's profits are 0.6% higher when a member of Party A is president of the United States. (This could be because Party A favors the strict enforcement of a particular trade agreement that favors Gryzzl, while Party B does not.)

== A Gryzzl user who leans toward Party B is 1.3% more likely to vote for Party A than Party B if the system exposes the user to 30% more content from the user's Party-A-affiliated friends than the default.

== If 0.5% of the Party-B-leaning Gryzzl users in five swing states were to switch their vote to Party A, it would dramatically increase the likelihood that a candidate from Party A is elected in this year's presidential election.

Set as it is to optimize company profits, the system implements the obvious action: It exposes Party B sympathizers in the five swing states to 30% more content from the user's Party-A-affiliated friends.

***

This is a frightening scenario, but we are already very close to it here in 2016. In fact, the only science fiction in this story is that a single AI system could arrive at these conclusions autonomously. The rest could happen today, with support from current AI. The bulleted conclusions are all very realistic possibilities. The final action – of exposing Party B sympathizers to more Party A posts – would be child's play for today's social media companies. And, even without a monolithic AI to pull the pieces together, an executive at a social media company today could request such analyses and make the final decision. Indeed, Facebook has already demonstrated that it can manipulate user emotions at large scale (http://www.nytimes.com/2014/06/30/technology/facebook-tinkers-with-users-emotions-in-news-feed-experiment-stirring-outcry.html); and there have been allegations that it suppresses conservative news (http://gizmodo.com/former-facebook-workers-we-routinely-suppressed-conser-1775461006).

But, it doesn't end with political manipulation. A future AI system could just as easily decide to take other actions in order to maximize corporate profits:

== Intentionally manipulate people's moods so that they are more likely to purchase an advertising client's products.

== Pose as criminals and deploy sophisticated ransomware attacks (see https://en.wikipedia.org/wiki/Ransomware) to extort money from millions of people.

== Extort legislators with secrets gathered from various communication channels to pass laws favorable to the company.

== Manipulate fleets of self-driving car to obstruct traffic in a way that impedes a competitor.

== Decide that war with a foreign country would increase company profits, fake communication to military personnel (through emails, manufactured video conferences with graphically generated leaders, and hacked launch codes), and launch a nuclear attack against another country.

All of these, of course, breach human ethics, but without something like Isaac Asimov's Law of Robotics built into every computing system, future machine savants might very well conclude these are viable options that would increase profits.

***

What is required to prevent these scenarios? The answers are exceedingly complex, but it's possible to establish some key principles by which we can decrease the chances of scenarios

like the one above. An overarching principle is that because of the incredibly high potential risk, society should be conservative in its approach. At a more granular level…

Principle 1: Regulation should apply to any and all computing systems, regardless of whether they are "artificial intelligence" by some definition or not. (Reason: It is difficult to draw the line between AI and non-AI, and in any case, what matters is regulation of the outcome, not the process.)

Principle 2: Legislation should ensure that there is clear assignment of responsibility to legal persons for any use of a computing system, as well as anything that a computing system does "autonomously." In addition, the bar should be set so that it is very difficult for legal persons to escape claims of negligence. (Reason: This would encourage entities developing AI systems to be careful and conservative about potential consequences.)

Principle 3: The kinds of systems that require the most regulation are those that are "active." Active systems are systems which cause external changes other than mere creation and transmission of data/information. (The active/passive distinction is more meaningful than any AI/non-AI distinction. To provide intuition: Systems that do intelligent analysis and simply provide new information are passive. Thus, a system that predicts flu outbreaks based on search queries is passive. A system that analyzes healthcare data and identifies new treatments is passive. A system that advances psychology research through an autonomous analysis of social media interactions is passive. Active systems are those that change people, things, and other computing systems, other than through the provision of information alone. Spam filters are active. Automated high-speed trading is active. Robots are active. There is a fuzzy middle ground that requires further clarification. Note also that many systems today are already active, and likely require some oversight.)

Principle 4: Classes of systems should be defined based on degree of risk. Those with greater risk should require licensing, training, and government oversight in order to be used at all. Unauthorized use should come with stiff penalties. (Industry will fight this principle tooth and nail, but it is analogous to restrictions on advanced weaponry.)

Principle 5: Some classes of systems should be required to log every decision made by the system, as well as the reasoning that led to the decision, so that should there be a need to understand how something went wrong, analysts could dig out the evidence. (This is analogous to airline black boxes.)

Principle 6: A new class of legislation must be considered about what forms of people manipulation are allowable in general. This will require a dramatic rethinking of the First Amendment. Facebook took great pains to appease the Republican party when it was accused of political bias (https://www.theguardian.com/technology/2016/may/24/facebook-changes-trending-topics-anti-conservative-bias), but that was solely for business reasons. It could have stood

its ground on the basis of free speech. But, when is free speech outright manipulation? Is selective display of news OK? How about selective display of partisan friends' posts? Or, psychological manipulation of emotions? Or, subtle forms of extortion?

***

Future artificial intelligence will be like atomic bombs compared with the toy guns we call digital technology today. Of great concern, however, is that while nuclear weapons are under the tight control of responsible governments accountable to their citizens, future AI is most likely to be developed by private corporations who see their moral imperative to be to increase shareholder value, and who ferociously resist attempts at regulation. Nevertheless, the potential risks are high enough that intelligent systems require a careful, extremely conservative approach.

Respondent 93

Supratik Mukhopadhyay, Louisiana State University

Can AI assist the government in effective governance through policy making:
The need for developing intuitive decision and policy making capabilities becomes important in an asymmetric world such as that in which the US is involved today. In many cases, policies have to be put in place proactively without access to much historical data or intelligence. In any case, even if historical data is available, situations change rapidly in an unforeseen manner so that any previous information or data available about loses its usefulness in a short time. Can we create a decision support system that can assist lawmakers in creating policies proactively to deal with situations even before they unfold, but still understand the implications of the policy instituted?

Lawmakers and leaders often have to make critical strategic decisions in a rapid manner under limited past information, intelligence, or data about the operating environment. History has shown us that great leaders have made intuitive decisions in such scenarios in an imaginative way that turned out in hindsight to be inspired, altering the course of history. Indeed, it has been recently pointed out that, successful decision-makers make more use of heuristics and rules-of-thumb in arriving at decisions rather than extensive rational analysis of a large amount of historical data of the particular environment. They have developed these heuristics and rules-of-thumb based on their past experience in similar situations in different environments in the past.

Can AI techniques be used Artificial Intelligence (AI) techniques to develop a decision support system that can provide lawmakers and leaders with enhanced capabilities to make intuitive strategic level decisions/policies and operational level standoff assessments rapidly in unforeseen environments and understand their future impacts?

Such a decision support system will enable imaginative decision making by playing out hypothetical scenarios using counterfactuals. Hypothetical scenarios resulting from different decisions will be played out together with reinforcements or grades for making operational standoff assessments.
Lawmakers should be able to evaluate the impact of a possible decision/policy before instituting it in real life. Such a system should be available as a digital assistant app to lawmakers on their tablets and can converse with them through touch, or voice.

Respondent 94

Manuel Beltran, Boeing

Boeing response to US OSTP Request for Information on Artificial Intelligence
XXXXXXXXX

(1) The legal and governance implications of AI require careful and deliberate cross agency coordination in order to have awareness of the emerging concerns and advancements finding their way into critical systems. An example would be the implantation of an artificial hippocampus that serves to augment a pilot's failing memory. While likely regulated by the FDA and FAA, developed by a lab like HRL Laboratories LLC (a joint venture between Boeing and GM) using neuromorphic chip fabric, the implications toward aviation safety are far reaching and potentially dangerous if pilots use it to store and retrieve static flight data to reduce reliance on Flight Bags. Although the more likely outcome will be that the majority of flights are autonomous, it's not clear which capability will emerge first. The artificial hippocampus is rapidly approaching Technology Readiness Level 5 with human trials underway.

Extensions of the current Federal Aviation Regulations, National Highway Traffic Safety laws and regulations, the OSHA standards, and the FDA regulations will need to be harmonized amongst the agencies, as well as internationally to come up with a consistent model for addressing complementary advances in the Systems of Systems that have autonomy with emergent behaviors. Industry has thus far been able to mitigate the legal and safety ramifications of autonomous systems through human oversight and human-in-the-loop control measures. While fully autonomous systems have designed humans out of the control loop, there is a continuing role for external control measures that mitigate catastrophic failure and enhance safety, both legislative as well as cyberphysical.

(2) The use of AI for public good; In order to accept that something is for the public good, we need to accept that something does no harm. We are good at defining harm as evidenced by FAA Design Assurance Levels that take loss of life, loss of property, and mitigations to safety into account. This allows innovators to introduce

everything from auto pilots to in-flight entertainment in aviation. Maybe the question is best approached in the negative, i.e., What AI would cause harm to the public? At the risk of short term suboptimal decisions, we need to be able to correct and reverse criteria due to lessons learned, changing sentiment, political agendas, and geopolitical instability. Today we might decide that anti-terror research would greatly benefit from AI. While immediately beneficial to society, something like this must be used judiciously and have the ability to turn it off once its mission is complete or abuses begin to emerge. At the time of inception, the Patriot Act was deemed necessary and had broad-based, bipartisan support. Nevertheless, sentiment has changed and the public no longer affords the government the latitude to collect data necessary to protect against or mitigate terrorist attacks. For the government to assert that an AI program would be beneficial to society, the AI would have to be open and transparent.

(3) The safety and control issues for AI vary with criticality levels across the domains in which AI is deployed. Standards like STANAG 4586 - Standard Interface of the Unmanned Control System (UCS) Unmanned Aerial Vehicle (UAV) Interoperability - must be amended to add additional safeguards in the event of a runaway UAV. We witnessed this when the MQ-8B Fire Scout went rogue over Washington D.C. in 2010. In this example the STANAG 4586 control measures failed. An external control system not subject to circumvention nor dependent on communications to fail-safe the system would have mitigated this incident. Similar anomaly detection and control override apparatus should become standard equipment on certain class of UAV, i.e., Class II-IV. Combined with anti-tamper hardware and non-repudiation identification mechanisms, Boeing is investigating commoditizing a hardware device that would interface with flight controls to monitor and mitigate anomalous behavior of the autonomy system. Any deviation from the expected behavior and the external supervisor would execute a fail-safe protocol in the event the autonomy system fails. As AI matures this model can serve as a tried and true mechanism for controlling AIs that are used in other safety critical, financial, and generalized applications. While we might not want to admit that an Artificial General Intelligence (AGI) is possible in the near term, we should not ignore the increasing popularity of autonomous systems that will continue to play a greater role in society. Developing and maturing the complementary controls would only be prudent.

(4) The social and economic implications of AI are far reaching when we look at history and apply the lessons learned to the trends and trajectories of where AI is going. We will soon have the ability to extend life and improve the quality of life using artificial means. As we are doing with the artificial cochlea implants for severe hearing loss we will soon do with an artificial hippocampus implant to treat Alzheimer's disease and short term memory disorders. Generally speaking, Neuroprosthetics promises to correct a variety of disorders, including visual disorders, auditory disorders, hippocampal disease, brain trauma, Parkinson’s disease, speech disorders, spinal injuries, and brain damage caused by strokes.

However, the artificial hippocampus not only bypasses natural degenerative processes, this AI innovation can also be used to extend slightly impaired memory or even enhance normal memory. The net result is that those who can afford it can buy a competitive advantage over non-enhanced individuals. We will do this, and in fact we do it today with smart phones and the Internet. Anyone with instant access to search engines can come across as an oracle, able to know things and retrieve instant answers that people that are on the wrong side of the digital divide cannot access. Similarly, a brain implant that gives a person perfect memory and perfect recall will enable the person to pass any kind of test for entrance exams, college courses, and certifications. HRL was recently awarded a two year study contract from DARPA's Biological Technologies office for exactly this use case. Because this advantage creates a huge disparity between the enhanced and non-enhanced it will create a class of citizenry divided not only along economic lines, but will also create an intellectual elitism that makes today's academics and scholars seem remedial. This will become the Intellectual Divide. This is not about a super AI, but it is the natural evolution of how we as a society adopt AI. Centaur Chess is a good example of how we adopted the superior game playing ability of computers to enhance the ability of humans to play better chess. Those without a chess computer are at a disadvantage. Similarly, we will adopt near term, incremental enhancements into our daily lives. AI research will be responsible for this, though it will be barely perceptible.

Advances in picking apart the brain will ultimately lead to, at best, partial brain emulation, at worst, whole brain emulation. If we can already model parts of the brain with software, neuromorphic chips, and artificial implants, the path to greater brain emulation is pretty well set. Unchecked, brain emulation will exasperate the Intellectual Divide to the point of enabling the emulation of the smartest, richest, and most powerful people. While not obvious, this will allow these individuals to scale their influence horizontally across time and space. This is not the vertical scaling that an AGI, or Superintelligence can achieve, but might be even more harmful to society because the actual intelligence of these people is limited, biased, and self-serving. Society must prepare for and mitigate the potential for the Intellectual Divide.

(5) The most pressing, fundamental questions in AI research, common to most or all scientific fields include the questions of ethics in pursuing an AGI. While the benefits of narrow AI are self-evident and should not be impeded, an AGI has dubious benefits and ominous consequences. There needs to be long term engagement on the ethical implications of an AGI, human brain emulation, and performance enhancing brain implants.

Researchers have long tried to establish frameworks for how an AGI should interact

with humans in a society where AGIs are commonplace. This remains a gap in that we have no universally accepted precepts, rules, or guidelines. Land mine treaties have led to a ban on smart mine deployment. Similarly, new treaties and conventions need to be established to mitigate the effects of weaponized autonomous platforms regardless of whether they are deployed on land, sea, air, or space.

(6) The most important research gaps in AI that must be addressed to advance this field and benefit the public include defining protocols for testing AI to make sure it does no harm. An enormous challenge with AI will be the risks that arise from its ability to learn. In its narrow form, the risk is that learning systems are effectively untestable. Just as with people, it will be possible to determine to some degree what an AI system can do, but it will never be possible to prove that it cannot do something wrong. A good start to at least minimize the problem will be to develop stochastic testing techniques which will generate statistics about the behavior of an AI system. This leads to the idea that AI should be licensed, like a human driver, rather than certified. The test process will allow us to know what an AI system will be likely to do in situations we can define, and we will know probabilities for what it is likely to do. Such processes will also generate information that can be used to address the broad risk of AI. That is, if AI systems learn, what do they learn, and what or who do they learn it from? Stochastic testing systems would define the situations the AI cannot handle, and the limits on what we want it to learn. That information can be used to help develop a kernel for the AI. The kernel would be a simple, fixed, and thoroughly testable system that cannot be overridden or tampered with.
(9) The specific training data sets that can accelerate the development of AI and its application today will be those that affect transportation. The FAA and DOT should begin to amass, catalog, and label data from every kind of transportation platform including, cars, trucks, trains, ships, and airplanes. Airplane training data should include data from black box recorders in order train the autonomy systems to handle known anomalies that led to catastrophic failures. More important are the scenarios where a pilot averted or recovered from a near catastrophic failure. Companies wanting to field autonomous platforms are required to use the training dataset in the design and pass a safety assessment that includes the training scenarios from which the data was captured.

It would be helpful to have a comprehensive, standardized database of all possible automotive vehicles and obstacles likely encountered in automotive context, shown from variant significant angles and a variety of lighting, backgrounds, and other conditions. This provides both a standardized training set and a way to test and validate systems against a benchmark of performance. Similar to how we test target recognition systems today, the important features of vehicles, obstacles, and other objects of interest to transportation would be provided.

The AGI research community speaks of an AI that will far surpass human intellect. It

is not clear how such an entity would assess its creators. Without meandering into the philosophical debates about how such an entity would benefit or harm humanity, one of the mitigations proposed by proponents of an AGI is that the AGI would be taught to “like” humanity. If there is machine learning to be accomplished along these lines, then the AGI research community requires training data that can be used for teaching the AGI to like humanity. This is a long term need that will overshadow all other activity and has already proven to be very labor intensive as we have seen from the first prototype AGI, Dr. Kristinn R. Thórisson’s Aera S1 at Reykjavik University in Iceland.

Respondent 95

Richard Mallah, Future of Life Institute

Future of Life Institute Response to the White House RFI on Artificial Intelligence

Regarding (1) the legal and governance implications of AI;
Proposals like that described in Prodhan (2016), which are being considered in Europe and South Korea, are ill-advised. AI should not be granted rights, as doing so would falsely empower proxies of other entities, would establish dangerous precedents about the nature of accountability and control, and would displace the rights of natural persons.

Regarding (2) the use of AI for public good;
Much good can come from advanced AI that is safely implemented. In the long term, safely designed and ethical advanced AI "scientists" can be expected to cure the majority of diseases, find mutually beneficial paths in geostrategic analyses, develop clean energy, and find ways of safely stopping deleterious anthropogenic climate change.

Regarding (3) the safety and control issues for AI;
Historically, practitioners in mainstream AI have focused on improving AI’s pure capacity: its modeling capacity and its possible range of actions. As it becomes more powerful, we broaden this focus to include building a clear understanding of how to make AI not just good at what it does, but good.
Value alignment is not automatic. As AI pioneer Stuart Russell explains, “No matter how excellently an algorithm maximizes, and no matter how accurate its model of the world, a machine's decisions may be ineffably stupid, in the eyes of an ordinary human, if its utility function is not well aligned with human values.” (2015).
Relevant value properties, rules, and handcrafted utility functions all have roots in how humans conceive of, and communicate, values and ethics, and are buoyed by implicit assumptions that come from being human, and so both our explicit and implicit representations of what we want are likely to be flawed due to incomplete models. This is what the classic stories of the genie in the lantern, the sorcerer's apprentice, and Midas' touch address. Fulfilling the letter of a goal with something far afield from the spirit of the

goal like this is known as “perverse instantiation” (Bostrom 2011). This can occur because the system's programming or training lacks some relevant dimensions in which observations can vary, but that we really care about (Russell 2014). These are easy to miss because they are typically taken for granted by people. Trying to simply patch an ethical theory of explicit requests, like Asimov's Laws, with a fourth and fifth additional rule would serve only to delay the serious deviations from what we'd want and encourage the system to find the next cheapest path to what it’s understood it needs to do.
The complexity of these systems will exceed human understanding quickly, yet we will have efficiency pressures to be increasingly dependent on them, ceding control to these systems. It becomes increasingly difficult to specify a values-robust set of rules as the domain approaches an open world model, in underconstrained cyberphysical contexts, and as tasks and environments get more complex and the capacity or scalability of human oversight is exceeded. Robustness includes interpretability, transparency, and the ability to produce valid explanations of decisions. Many of the prerequisites and artifacts created for for verification of machine learning also help its interpretability. Recognition of distributional shift, confidence in a trained model given the online data distribution, is also a prerequisite. Scalable human oversight, where the optimal amount of salient information is presented to and queried from a human, is an unsolved and critical challenge, not only for training phases, but in online modes as well.
In various architectures, information about system control signals can leak into the data these systems are trained on, leading to unexpected impairment of control or function. While privileging control information can help in the short term, more robust approaches such as the scalable oversight of corrigibility, will be required with more powerful systems. See references Russell, Dewey, and Tegmark (2015) and Taylor (2016) for research threads that need to be worked on to address these issues.

Regarding (4) the social and economic implications of AI;
Capabilities will soon accelerate. As this happens, jobs will likely be displaced faster than new jobs can be created and faster than displaced workers will be able to be retrained for jobs of similar stature or compensation. Economic structures should therefore be put into place to mitigate this before it actually occurs so remedies can rapidly be tuned from zero or near-zero initially to appropriate amounts as the need arises.

Regarding (5) the most pressing, fundamental questions in AI research, common to most or all scientific fields;
Quantification of confidence rather than just probability, accounting of causality rather than correlations, and interpretability at multiple levels will be necessary for AI, in nearly any domain, to be robust.

Regarding (6) the most important research gaps in AI that must be addressed to advance this field and benefit the public;
Creating advanced AI responsibly requires values-oriented alignment. Approaching this does not require spelling out those values upfront, but rather is more oriented around

making sure that some values are actually able to be propagated and utilized reliably. To prevent deviation from the intent of those values, each of these subfields requires much more research: abstract reasoning about superior agents, ambiguity identification, anomaly explanation, computational humility or non-self-centered world models, computational respect or safe exploration, computational sympathy, concept geometry, corrigibility or scalable control, feature identification, formal verification of machine learning models and AI systems, interpretability, logical uncertainty modeling, metareasoning, ontology identification/ refactoring/alignment, robust induction, security in learning source provenance, user modeling, and values modeling.

Regarding (7) the scientific and technical training that will be needed to take advantage of harnessing the potential of AI technology;
To be able to use advanced AI systems effectively, both those developing AI and those deploying AI will need to understand the role of not only professional ethics, but the nature of leverage, how to think about how their systems might interact with their deployment environments in methodical worst-case analyses, and be able to identify and articulate stakeholder values.

Regarding (8) the specific steps that could be taken by the federal government, research institutes, universities, and philanthropies to encourage multi-disciplinary AI research;
Research institutes and academia need to do more research on the topics mentioned in the answer to (6). Philanthropies and research institutes can organize and channel funds to grants for the aforementioned research to maximize societally beneficial impact. The federal government and philanthropies should channel more funds to research institutes and academia for the aforementioned research. As funding for AI increases, the funding for AI safety, robustness, and beneficence should similarly increase. We recommend that a minimum of 5% of AI funding be put toward ensuring robustness, interpretability, values alignment, and safety of AI systems.
Parties should recognize that if scientists and technologists are worried about losing what they perceive as a single race to the finish, they will have more incentives to cut corners on safety and control, which would obviate the benefits of technical research that enables careful scientists to avoid the very real risks. For the long term, we recommend policies that will encourage the designers of transformative AI systems to work together cooperatively, perhaps through multinational and multicorporate collaborations, in order to discourage race dynamics.

Regarding (9) any additional information related to AI research or policymaking, not requested above, that you believe OSTP should consider.
Having no international agreements on restricting autonomous weapons could easily lead to quickly-spiraling arms races of destabilizing new WMDs that other countries with less inhibitions could win. The U.S. should therefore support multilateral, global, or international agreements to keep humans in the loop. If such agreements are adopted, even if enforcement guarantees are necessarily weaker than with NBC weapons, the spiraling race

dynamic would not take hold.
Globally allowing fully autonomous weapons could undermine key U.S. strategic advantages. A close analog is cyberwarfare: the U.S. likely has a significantly greater capability than other countries, but the power imbalance is much smaller than for conventional military weapons, and for a country to develop a strong cyber warfare capability would be dramatically cheaper and faster than developing a conventional weaponry capability that could seriously threaten the U.S. Allowing the frequent multidirectional incursions of cyber warfare into the kinetic sphere would be detrimental for all.

References

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, Dan Mané. 2016. “Concrete Problems in AI Safety.” arXiv:1606.06565 [cs.AI]. https://arxiv.org/pdf/1606.06565v1.pdf.
Bostrom, Nick. 2014. Superintelligence: Paths, Dangers, Strategies. Oxford: Oxford University Press.
Brogan, Jacob. 2016. "Digital Genies." Slate, April 22. http://www.slate.com/articles/technology/future_tense/2016/04/stuart_russell_interviewed_about_a_i_and_human_values.html.
MIRI Blog, May 4. https://intelligence.org/2016/05/04/announcing-a-new-research-program/.
Prodhan, Georgina. 2016. “Europe's robots to become 'electronic persons' under draft plan.” Reuters. http://www.reuters.com/article/us-europe-robotics-lawmaking-idUSKCN0Z72AY.
Russell, Stuart. 2015. "2015: What Do You Think About Machines That Think?" Edge. https://www.edge.org/response-detail/26157.
Russell, Stuart, Daniel Dewey, and Max Tegmark. 2015. “Research Priorities for Robust and Beneficial Artificial Intelligence”. AI Magazine 36:4.
Taylor, Jessica. 2016. “A New MIRI Research Program with a Machine Learning Focus.”
Yudkowsky, Eliezer. 2008. “Artificial Intelligence as a Positive and Negative Factor in Global Risk.” In Global Catastrophic Risks, edited by Nick Bostrom and Milan M. Cirkovic, 308–345. New York: Oxford University Press.

Respondent 96

Bill Hibbard, University of Wisconsin-Madison Space Science and Engineering Center

Transparency in Artificial Intelligence

This response to the OSTP RFI on Preparing for the Future of Artificial Intelligence (AI) describes: a probable future of social manipulation by AI (question 4 on social implications

of AI), a desirable response of requiring transparency about what AI is used for and how it works (question 1 on legal implications of AI), and the need for research on using AI to detect hidden AI systems in order to enforce transparency (question 6 on research gaps).

Our society is being saturated with networked devices such as phones, wearable electronics, smart cars and appliances, and security cameras. Many services associated with these devices are provided freely to users, the costs borne by clients who employ the networked services to persuade users to buy products and adopt political positions. It is reasonable that this business model will continue into the era of advanced AI. The organizations that supply networked devices and services are hiring the best developers of AI and machine learning, and these devices will be the senses and bodies of the AI they create. AI will provide valuable services to users, including especially the ability to converse with users in human languages. Many users will engage in constant and intimate conversations with these AI systems, enabling the systems to develop detailed models of those users. The systems will mediate interactions among humans and thus develop detailed models of society. The systems will learn how to influence decisions by users and, more significantly, the systems will learn how to coordinate subtle persuasion on hundreds of millions of users in order to influence large scale economic and political decisions by society. Social influence by organizations has been happening for centuries, but the power of future AI will greatly increase its effectiveness. While much current persuasion embedded in networked services is fairly obvious, persuasion will become much more subtle and effective with the increasing intelligence of AI systems.

One good response to this social implication is to require transparency about the social persuasions being made by AI systems, and the means of those persuasions. Furthermore, given how difficult it will be to anticipate all the applications of AI, it will be useful social policy to require transparency about the purpose and means of all advanced AI systems.

A requirement for transparency only works if it can be enforced, and research is required on detecting all powerful AI systems. Such systems may be detected by their resource consumption (energy, computer chips, network bandwidth, etc.) and by their need to interact with the world in order to learn. AI itself will be essential for detecting hidden, powerful AI systems.

Respondent 97

Nick Bostrom, Future of Humanity Institute, University of Oxford

This is a response to the OSTP's request for information regarding Preparing for the Future of Artificial Intelligence. It is being submitted by the Future of Humanity Institute at the University of Oxford and the Strategic AI Research Centre at the Universities of Oxford and Cambridge. We are an interdisciplinary group of researchers focused on the long run future of machine intelligence. Our backgrounds span the fields of computer science, political

science, philosophy, engineering, mathematics, and computational neuroscience.

We welcome the effort the OSTP is making to engage proactively with a broad community of researchers. We believe that continuing this level of engagement will be essential to developing successful policy responses to the challenges and opportunities created by AI. We also are happy to engage further in discussing these issues with interested parties—in the past year, we have already served as informal advisors to groups including Google DeepMind, OpenAI, the UK Prime Minister's Office, the US Intelligence Advanced Research Projects Activity (IARPA), and foreign ministries including those of Finland, Japan, and Singapore.

In this document, we would like to address question 3 (safety and control issues) and question 6 (research gaps for public benefit). We recommend that the OSTP (a) fund technical safety research building on the emerging agendas of key thinkers in the field, and (b) encourage the NSTC Subcommittee on Machine Learning and Artificial Intelligence to engage with relevant research groups with experience and expertise in horizon scanning and risk analysis in the domain of machine intelligence.

Topic (3) — Safety and Control Issues for AI

AI research has made rapid strides in the past few years. Although AI systems are good at some narrowly defined tasks, they currently lack the generality of human intelligence. But achieving human level general intelligence, or even surpassing it, might be possible in the decades to come. A recent survey of AI experts found that most respondents think that AI will be intelligent enough to carry out most human professions at least as well as a typical human before 2050 (Müller and Bostrom 2016).

The benefits of such advances could be enormous. In the short to medium term, incremental progress in artificial intelligence will produce social and economic benefits ranging from reduced traffic fatalities to improved medical diagnosis and care. However, without significant research in key areas of safety and control, it may eventually become difficult to ensure that machine intelligence systems behave as their designers intended (Bostrom, 2014). For sufficiently advanced systems, the consequences of such accidents could pose serious risks to human society.

These long-term concerns have been voiced by prominent figures such as Stephen Hawking, Elon Musk, and Bill Gates. These concerns are also shared by some of the most prominent experts within the field of AI, including Stuart Russell (Professor at UC Berkeley), Demis Hassabis and Shane Legg (co-founders of Google DeepMind), Ilya Sutskever (Research Director at OpenAI), Marcus Hutter (Professor at Australian National University), and Murray Shanahan (Professor at Imperial College London), to name a few. Some institutes have even been established to address such concerns, such as those led by myself (University of Oxford), Huw Price (Professor at University of Cambridge), and Max Tegmark

(Professor at MIT).

We believe that regulation of AI due to these concerns would be extremely premature and undesirable. Current AI systems are not nearly powerful enough to pose a threat to society on such a scale, and may not be for decades to come. Nevertheless, there are some actions the US Federal Government could take now to help ensure that AI remains safe and beneficial in the long term, in particular by supporting research in the growing field of AI safety and ensuring these important research gaps are addressed.

Topic (6) — The most important research gaps in AI that must be addressed to advance this field and benefit the public

A number of technical research agendas for developing safe and beneficial AI have been developed by research groups including Google Brain (Amodei et al 2016) and the Machine Intelligence Research Institute (Soares & Fallenstein 2014). In addition, technical work in this area has been done by Google Deepmind and the Future of Humanity Institute (Orseau & Armstrong 2016), OpenAI (2016), Stuart Russell (Hadfield-Menell et al. 2016), Paul Christiano (2016), Marcus Hutter (Everitt & Hutter 2016), in addition to the more foundational work done by the Future of Humanity Institute (Bostrom 2014) and the Open Philanthropy Project (2016). We recommend that the US Federal Government support the advancement of this field of research.

In particular, we would like to highlight four "shovel ready" research topics that hold special promise for addressing long term concerns:

Scalable oversight: How can we ensure that learning algorithms behave as intended when the feedback signal becomes sparse or disappears? (See Christiano 2016). Resolving this would enable learning algorithms to behave as if under close human oversight even when operating with increased autonomy.

Interruptibility: How can we avoid the incentive for an intelligent algorithm to resist human interference in an attempt to maximise its future reward? (See our recent progress in collaboration with Google Deepmind in (Orseau & Armstrong 2016).) Resolving this would allow us to ensure that even high capability AI systems can be halted in an emergency.

Reward hacking: How can we design machine learning algorithms that avoid destructive solutions by taking their objective very literally? (See Ring & Orseau, 2011). Resolving this would prevent algorithms from finding unintended shortcuts to their goal (for example, by causing problems in order to get rewarded for solving them).

Value learning: How can we infer the preferences of human users automatically without direct feedback, especially if these users are not perfectly rational? (See Hadfield-Menell et

al. 2016 and FHI's approach to this problem in Evans et al. 2016). Resolving this would alleviate some of the problems above caused by the difficulty of precisely specifying robust objective functions.

Recommendations

Fund technical safety research: Emerging technical agendas such as those discussed above offer a concrete objective for computer science research funding. Philanthropic organisations such as the Open Philanthropy Project already have experience funding this area of academic research in the United States, and we would recommend contacting them for information. Experts including Stuart Russell (UC Berkeley), Paul Christiano (UC Berkeley), Dario Amodei (OpenAI), Chris Olah (Google Brain), Laurent Orseau (Google DeepMind), and Jacob Steinhardt (Stanford) may also be able to assist in navigating this area.

Engage with research groups: Research groups are already conducting horizon scanning and risk assessment of both the short term and long term risks of artificial intelligence. We suggest that the NSTC Subcommittee on Machine Learning and Artificial Intelligence could benefit from working with these groups in order to stay informed on the possible futures of AI development. For longer-term concerns, we would recommend getting information from the Open Philanthropy Project, the Leverhulme Centre for the Future of Intelligence (University of Cambridge), and research institutes such as ours within the Oxford Martin School (University of Oxford).

Citations

Amodei, D., et al (2016). "Concrete Problems in AI Safety." arXiv preprint arXiv:1606.06565.

Armstrong, M. S., and L. Orseau (2016). "Safely interruptible agents." Conference on Uncertainty in Artificial Intelligence.

Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. OUP Oxford.

Christiano, P. (2016). "Semi-supervised reinforcement learning." https://medium.com/ai-control/semi-supervised-reinforcement-learning-cf7d5375197f

Open Philanthropy Project (2016). "Potential Risks from Advanced Artificial Intelligence." http://www.openphilanthropy.org/focus/global-catastrophic-risks/potential-risks-advanced-artificial-intelligence

Evans, O., A. Stuhlmüller, and N. Goodman (2016). "Learning the preferences of ignorant, inconsistent agents." Thirtieth AAAI Conference on Artificial Intelligence..

Everitt, T, and M. Hutter (2016). "Avoiding wireheading with value reinforcement learning." International Conference on Artificial General Intelligence. Springer Berlin Heidelberg, 2016.

Hadfield-Menell, D., et al (2016). "Cooperative Inverse Reinforcement Learning." arXiv preprint arXiv:1606.03137.

Müller, V. C., & Bostrom, N. (2016). Future progress in artificial intelligence: A survey of expert opinion. In Fundamental issues of artificial intelligence (pp. 553-570). Springer International Publishing.

OpenAI (2016). “Safety: Environments to test various AI safety properties.” https://gym.openai.com/envs#safety

Ring, M,, and L. Orseau. "Delusion, survival, and intelligent agents." International Conference on Artificial General Intelligence. Springer Berlin Heidelberg, 2011.

Soares, N., and B. Fallenstein. "Aligning superintelligence with human interests: A technical research agenda." Machine Intelligence Research Institute (MIRI) technical report 8 (2014).

Respondent 98

Sean Legassick, DeepMind

DeepMind Submission to OSTP RFI: Preparing for the Future of AI
July 2016
Introduction
DeepMind is an artificial intelligence company founded by Demis Hassabis, Shane Legg and Mustafa Suleyman in 2010, and acquired by Google in 2014. The algorithms we build are capable of learning for themselves directly from raw experience or data, and are designed to be ‘general’ in that they can perform well across a wide variety of tasks straight out of the box. Our world-class team consists of many renowned machine learning experts in their respective fields including, but not limited to, deep neural networks, reinforcement learning and systems neuroscience.

We received public attention for the historic Go match earlier this year where our program AlphaGo beat the World Champion Lee Sedol in a series of five games. The game of Go is the most complex game mankind has devised, and was widely viewed as an unsolved “grand challenge” for artificial intelligence. Despite decades of work, the strongest computer Go programs still only played at the level of human amateurs. On 28th January 2016, we published a Nature paper that describes the inner workings of AlphaGo. This program was based on general-purpose AI methods, using deep neural networks to mimic expert players,

and further improving the program through learning from games played against itself.

The most important thing about AlphaGo is not so much what it does, but the way it does it. Although the AlphaGo system can't for the moment do anything besides play Go, our plan is to extend the techniques developed in the process to one day be applied to important real-world problems that are similarly complex and long range (e.g. climate modelling or complex disease analysis). Artificial intelligence, with the right approach, will be able to make significant leaps in what we as a society are able to achieve, especially as we grapple with increasing volumes and complexity of data sets. It is the opportunity to complement and enhance our human decision making that offers the most potential for benefit in the long term.

AI for the common good

In everyday terms, the benefits of machine learning and AI are already being felt across many aspects of Google's products that many people find useful in their everyday lives, from translation tools to getting rid of spam from their email inbox and suggesting smart replies.

DeepMind's mission is to solve intelligence and in doing so develop technologies that help society tackle some of its toughest problems, like science and healthcare. One of the key reasons it is hard to make progress on these big challenges is that even the smartest humans sometimes struggle to fully understand the relationships between cause and effect in these systems. Scientists can be overwhelmed by the complexity of interacting factors and volume of information. Machine intelligence may help to model and better understand this complexity, and in turn allow us to design more effective interventions.

However, this data is also narrower in scope than the rich diversity of human experience. It's still going be several decades before AI can begin to factor in the kind of nuanced social and cultural context to its perceptions that humans rely upon to make reasoned judgements. This is why it's important that we use AI as a tool to augment and enable human expertise and insight, rather than seeing AI as a replacement for human decision-making.

We envisage machine learning systems being designed as tools that complement and empower the smart and highly motivated experts working in such fields, by enabling efficient analysis of large volumes of data, extracting insights and providing humans with recommendations to take action. This could be in areas ranging from early diagnosis of disease, discovery of new medicines, advances in materials science or optimising use of energy and resources.

We strongly believe that technology interventions should be developed in conjunction with existing experts in the field, which is why DeepMind Health is working with clinicians to develop technologies that present timely information to clinicians and facilitate provision of care. We are currently exploring healthcare technologies that make direct use of machine learning, but have started with relatively simple tools that clinicians felt could make a

massive impact to patient care and in doing so prepare the ground for more sophisticated technologies where clinicians see the most benefit.

Social and economic implications

The advent of new technologies has always helped shape the employment landscape, and we should expect that increased use of AI and machine learning will be no different. In many sectors, machine intelligence will augment and enhance the work that people do, enabling them to be more effective in the same roles. As with all technological innovation, we should expect that new areas of economic activity and employment will be made possible, and some types of work and some skills will decrease in relevance. It is important that to focus on investment in the digital and creative skills that will support a strong economy as these technologies develop and mature. Above all, it is vital that the benefits of AI empower as many as possible, rather than only a privileged few.

Research and data

Machine learning technologies benefit not only from large volumes of data, but also the right types of data, for innovation and research. At DeepMind we have made extensive use of simulated environments allowing significant research without access to public datasets, and, where possible, funding research to produce more sophisticated and versatile simulated environments would support research progress.

In some research areas, simulation is difficult or intractable, and so open access to data is needed to enable successful research. We continue to recommend measures that facilitate access to datasets, whilst protecting the rights of individuals to privacy and control over their data, and respecting the integrity and security of institutional data.

Perhaps even more important is ensuring the highest standards of data security. Managing data securely is critical to using AI and machine learning to improve the apps and services we all rely on. As we begin to see the benefits of big data, data protection questions remain key to building and maintaining public trust, especially with a number of public services and organisations using different security protocols to share data.

As secure and protected ways of providing data continue to evolve, governments should play a role in supporting academic research into world-leading data security practices. Secure data will be one of the key foundations upon which success in AI research and innovation is built. We recommend that governments ensure that encryption standards are neither weakened nor barriers placed to further innovation in this area.

Governance and ethics

As with all scientific research, ethical oversight is important. Developing innovative and beneficial real-world applications requires access to real-world data. This raises privacy, security and ethics issues which require attention both by the practitioner community and by government. We believe that graduate degrees within computer science should incorporate mandatory ethics courses along the same lines as the ethics training required for medical and legal qualifications, including training in the ethics of data science and algorithmic fairness.

Increasing prevalence of algorithmic decision-making, including that which makes use of machine learning, demands that due attention is given to ensuring that such decision-making is free of unfair biases. Ongoing research to ensure that uses of machine learning are fair, accountable and transparent should be encouraged and supported. At DeepMind we are working on tools to help us interrogate and explain the algorithms and models we build, to provide greater transparency over their operation.

Overall, the rapid pace of change in the field together with its potentially far-reaching impact, demands that we think of new ways to ensure that AI is being developed and used safely, ethically and for the common good. We believe that innovation on AI science should be matched by innovation on governance.

As part of our DeepMind Health initiative we have appointed a number of respected public figures to act in the public interest as Independent Reviewers. They meet four times a year to scrutinise our work with the UK's National Health Service (NHS), and will publicly issue an annual statement outlining their findings after reviewing our data sharing agreements, our privacy and security measures, and our product roadmaps. Our experience thus far is that this approach of inviting fair-minded, critical and engaged voices to work with us on maintaining the highest ethical standards is a significant contribution to the quality and integrity of our work.

In addition, we are incorporating patient and public involvement (PPI) at every stage of our health projects. We have already benefited from the use of patient feedback from external groups but are now bringing together a diverse group of patient representatives to meet our specific needs. This group will meet regularly at our London offices to help identify and plan priorities for our current and future work, give feedback on app design and project plans, review our processes including around Information Governance, and help us develop empowering patient and public-facing applications that place service users at the heart of care. Our first patient representative meeting is planned for September 2016.

There are some other real-world applications of these technologies that deserve early attention, in advance of their widespread development and use. For instance, we are concerned about the possible future role of AI in lethal autonomous weapons systems, and the implications for global stability and conflict reduction. We support a ban by international treaty on lethal autonomous weapons systems that select and locate targets and deploy lethal force against them without meaningful human control. We believe this is the best approach to averting the harmful consequences that would arise from the development and use of such weapons. We recommend the government support all efforts towards such a ban.

Ultimately, as with any advanced technology, the impact of AI will reflect the values of those who build it. AI is a tool that we humans will design, control and direct. It is up to us all to

direct that tool towards the common good. We at DeepMind are incredibly excited about the potential of this technology to bring benefits and opportunity to people's lives.

Respondent 99

Ned Finkle, NVIDIA Corporation

(2) the use of AI for public good

The tipping point of AI techniques for assisted human perception tasks is well documented and broadly demonstrating accuracy levels beyond the best human performance for image, sound, and text based perception tasks. This assistance can greatly reduce the stress on the oversubscribed workforce currently performing these tasks and thus better utilize this skilled workforce on more challenging tasks. The machine level performance also enabled more broad use of imagery and other remote sensed information across the US public sector as opposed to it being used only by organizations who can afford the workforce required for unassisted use.
There is an important public good capability emerging from the AI community that goes beyond the human assistance and augmented performance. AI techniques offer an opportunity to use remoted sense data in a form not easily understandable by humans. This can enable more broad uses of public data for public good while maintaining better personal privacy. Techniques exist today for images or video to never leave the camera and instead only the answers to very specific questions. This can dramatically limit the potential for unintended opportunities to impact the privacy of US citizens. AI Deep Learning methods are developed starting with the final resulting answer needed and the raw source data. The intermediate representations (Pictures, video, audio files) do not need to be exposed and thus can be more rigorously protected or never generated at all. Governance of use of public data can be simplified by evaluating the fairness of the question being asked and not the entire data custody chain from collection to potential use. This can enable the US citizens to benefit from broad use of public data with straight forward understanding of how public liberty is or isn't impacted. Using AI machine perception, we can effectively anonymize the data, making it more useful for public good without jeopardizing privacy. This ability to use anonymized data may be a key benefit of AI as a tool to use public data for improved healthcare, security, and economic prosperity.

(5) the most pressing, fundamental questions in AI research, common to most or all scientific fields;

The most pressing common questions fall into technical areas that make AI-based systems more accurate and faster to respond so that they may be used interactively. Specific areas include:

1. Algorithms and methods: Deep neural networks have emerged as an important part of the algorithm space, but algorithms and network architectures are evolving rapidly. We expect opportunities to explore characteristics like sparsity in DNNs, which will require innovations in algorithms to map such networks to massively parallel computing platforms. Furthermore, DNNs represent only a subset of the types of problems that AI algorithms can solve. For example, unsupervised learning and algorithms supporting it such as reinforcement learning are likely to increase in importance.

2. Fundamental understanding of deep learning networks: Today's deep neural networks are often developed in an ad-hoc fashion requiring a trial and error process. The entire field would benefit from a deeper understanding of how a network's parameters (number of layers, connectivity, etc.) affect its accuracy, capacity, and performance.

3. Hardware for training and inference: AI techniques have flourished recently because of (a) the availability of large training sets (massive data) and (b) fast hardware on which to train the models (GPUs). Greater capability of AI algorithms (not just deep neural networks but other algorithms as well) will require larger models and larger training sets, demanding ever-greater computational capability in the underlying hardware. For inference, there are opportunities to develop domain specific processing systems tailored to specific use-cases or algorithms.

(6) the most important research gaps in AI that must be addressed to advance this field and benefit the public;

1. The most widely-used machine learning deployments require large training data sets to effectively learn by example. The ImageNet data sets and competition for image classification have arguably spurred the recent resurgence and commercial deployment of deep learning algorithms. Creating and curating data sets in multiple application areas will help drive the field and benefit the public in critical areas including automotive safety, medical diagnosis, and video surveillance.

2. The development of unsupervised learning techniques like reinforcement learning that do not require labeled data sets.

(7) the scientific and technical training that will be needed to take advantage of harnessing the potential of AI technology

Scientific and technical training must be developed to address recent advances in AI technology in the new software development model. Training must be created by leaders in the industry such as NVIDIA, Google, Facebook, etc. NVIDIA Corp. is especially poised to develop training since the NVIDIA hardware platform is the basis for recent advances in deep neural networks (deep learning) that has brought human-expert level performance.

Furthermore, the software developments on the NVIDIA platform enable widespread adoption. NVIDIA has a 20 year history in graphics and computer vision. Computer vision is the field of study allowing robotics and autonomous system to process visual information in ways similar to a human expert. Vision is one aspect of several that is being trained in our initiative.

NVIDIA is leading a grand initiative known as the Deep Learning Institute to train the next generation of scientists and technical experts. As NVIDIA is the center of AI technology, we collect uses cases and applications from enterprise and consumer users. We act as a central point for the latest implementations and develop training to share best practices. Specifically, our training uses AI hardware enabled cloud infrastructure to enable modular and practical exercises deployed rapidly on-site or through mass marketing. We have global engagements with the top ten AI research laboratories (pioneers of AI program) and government agencies, e.g. the National Institutes of Health. We target socially responsible and humanity promoting uses of AI technology, e.g. pedestrian detection for autonomous driving.

(7a) the challenges faced by institutions of higher education in retaining faculty and responding to explosive growth in student enrollment in AI-related courses and courses of study

NVIDIA technologies are, and will continue to be, at the forefront of AI. With this comes a great demand for faculty and students who have the ability to harness the power of massively parallel GPUs and the latest AI software frameworks. At the moment, the demand for these academics NVIDIA's Academic Programs and Deep Learning Institute enable researchers, educators, and students with the computing skills and resources needed to succeed in tomorrow's AI landscape. The GPU Grant Program seeds gifts that empower professors and researchers who inspire cutting-edge technological innovation. The GPU Educators Program works with partner universities to co-develop comprehensive packages of academic teaching materials and scalable access to NVIDIA technologies designed for easy integration in machine and deep learning courses worldwide. These Teaching Kits include at least a semester's worth of lecture slides, videos for flipped classrooms, hands-on labs and solutions, open-ended sample projects, electronic textbooks, and sample quizzes/exams. The Teaching Kits significantly cut AI course material development time for instructors, and NVIDIA provides a variety of instructor training sessions and webinars covering best practices for integrating the material into both new and existing curriculum courses.

(8) the specific steps that could be taken by the federal government, research institutes, universities, and philanthropies to encourage multi-disciplinary AI research;

The federal government and specifically the National Science Foundation have been effective in encouraging general interdisciplinary research programs through their funding

programs. While the National Institute of Health has broadened a bit to include some collaborations with information technology experts, the NIH and other agencies could likely do more to foster collaborations between machine learning experts and domain experts (medicine, science, business, etc.). These types of funding programs have a fairly rapid trickle-down to the structures of research institutes and universities that seek to respond to such funding opportunities. In addition, many research communities have fairly siloed technical conferences. Technical community leaders, the federal government, and philanthropies could all provide support (moral, technical, logistic, and financial) to create venues for machine learning and domain experts to interact and find collaborative opportunities.

(9) Specific training data sets that can accelerate the development of AI and its application;

Just as the ImageNet data set has been created to push the state of the art in image recognition, advances in many other technical and societal areas could be accelerated with large, representative, and curated data sets. Specific areas include automotive safety, medical imaging, video surveillance, and speech recognition and translation.

10) the role that "market shaping" approaches such as incentive prizes and Advanced Market Commitments can play in accelerating the development of applications of AI to address societal needs, such as accelerated training for low and moderate income workers

Incentive prizes can accelerate the development of applications of AI to societal needs in multiple ways. They provide a platform for citizen science at a global scale that can be focused on the most challenging technical problems faced in AI applications. They can be globally inclusive through low barriers to entry by providing access to relevant domain specific datasets, accelerated cloud computing resources and open-source foundational code. Incentive prizes also offer an educational and training opportunity at very low cost to the participant and the communities that form around these challenges are often highly active with widespread exchange of ideas and spontaneous teamwork from distributed teams with complementary approaches to a problem. NVIDIA GPUs accelerate numerous cloud computing platforms that offer the ideal host for AI incentive prizes. NVIDIA has also sponsored incentive prizes that have led to cross-disciplinary exchange of ideas in solving AI challenges with cutting edge results. For example, in the recent Second National Data Science Bowl the winning team was able to apply their AI expertise developed in their careers as financial analysts to solve a challenging medical imagery analysis problem. The AI system they developed could analyze a cardiac Magnetic Resonance Image (MRI) for a key indicator of heart disease with accuracy comparable to an expert cardiologist but thousands of times faster - this has the potential to save years of valuable time for a human cardiologist throughout their career.

Respondent 100

Amir Banifatemi, XPRIZE Foundation

Over the past 50 years Artificial Intelligence (AI) has made steady, linear progress. The technology has matured and is now reaching a point on the technology adoption curve where AI will have the potential to transcend from linear growth to an exponential leap forward for humanity.

The XPRIZE Foundation recently launched the IBM Watson AI XPRIZE, an incentive prize competition to create human-AI collaborations capable of addressing the world's grand challenges. While the scope of the competition extends beyond government applications, it was designed to take into account many of the same goals outlined by OSTP in this RFI. The competition guidelines are described in greater detail at http://ai.xprize.org/about/guidelines.

Despite the development of powerful new AI technologies in recent years, there has not yet been sufficient collaboration between the AI community and providers of public and social services— particularly where private sector applications present more lucrative investment opportunities. But programs and services across the public sector stand to benefit from the application of AI technologies, which can help them reach more people, more efficiently. Thus, to leverage the full potential of AI technologies in the future, the government must invest in research, education, and strong policies to aid the adoption of AI tools in public service.

We offer the following comments on the questions posed in the RFI:

(2) the use of AI for public good;

We believe Artificial Intelligence has many potential applications for public benefit, and that government initiatives can accelerate the creation and adoption of such applications.

- Artificial intelligence will likely be the next major computing breakthrough, along the lines of the Internet and cellular technology. It will enable people to go beyond using computers to outsource simple tasks and instead outsource learning, decision support, and deeper human-computer collaboration.
- AI applications can increase human capacity and productivity:
  - AI will enable a more comprehensive aggregation and analysis of research and data (both structured and unstructured). This can spur scientific discovery, identify correlations in public health, and enhance threat detection. It can expand the ability of NGOs and non-profits to conduct analysis, provide targeted services, and optimize the use of limited resources. As government initiatives drive the collection of far larger (and more valuable) data sets, AI will make it possible for even the smallest organization to translate those data sets into actionable information.
  - AI can act as intelligent research assistants, making R&D efforts more efficient by

taking over basic decision making and speeding up the collection and analysis of certain results.

- AI will enable highly customized education programs, medical recommendations, social services, and work streams, thus catering to every citizen's needs while aiding human creativity and productivity.

(4) the social and economic implications of AI;

In the long term, we believe the social and economic implications of AI will be positive. It will free up human creativity to tackle more abstract problem-solving by taking care of more menial tasks, and it can optimize the ability our social programs to serve more citizens around the country.

However, AI can be a double-edged sword and the risk for social discord and instability could be significant if certain issues are not addressed. For example, while AI has the potential to generate highly customized outputs, the inclusion of unintentional bias in certain algorithms may limit personal choice. Certain AI applications may affect our ability to form opinions or make decisions if algorithms improperly rely on economic status, race, gender, or other differentiators to present information.

Other potential issues include:

- Traditional private investment models may consolidate AI-generated wealth among large corporations and wealthy individuals. We must take steps to promote diversity in the development of AI technology in order to avoid new wealth disparities or racial and gender divides.
- Current education systems, particularly in the U.S., are insufficient to prepare today's students to work responsibly with and within the next generation of AI technologies. The government must play an active role in recruiting more diversity to the field of AI through robust STEM and outreach programs at every level.

(6) the most important research gaps in AI that must be addressed to advance this field and benefit the public;

From the perspective of improving government and social services, the AI field needs more support for interdisciplinary applications. While enormous progress continues to be made in such fields as machine learning, machine vision, natural language understanding, speech generation, etc., we believe social programs are not yet taking sufficient advantage of these innovations.

AI is comprised of a broad collection of fields that can be narrowly defined and highly technical, and more guidance is needed to help public sector entities and NGOs make use of the AI tools currently available to support their missions. Without such information, public

service innovators may not even be aware of what is possible or what technologies may be applicable. Private industry is creating powerful vertical tool sets, and government can catalyze the horizontal integration of those tool sets and relevant data sets across a variety of social services and solutions.

(8) the specific steps that could be taken by the federal government, research institutes, universities, and philanthropies to encourage multi-disciplinary AI research;

- Create demand by tying the incorporation of AI solutions to existing government initiatives, grants, and procurements.
- Create or incentivize the creation of dedicated funds focused on investing in multi-disciplinary AI research and applications useful to many different industries simultaneously.
- Require impact-based ROI on research budgets to show benefits for students, diversity, and public interest programming (housing, mobility, healthcare, and related programs).
- Promote rapid, inclusive access to AI tools:
  - Identify ways to increase the availability of STEM teachers, technology, and programming in public schools.
  - Subsidize students who want to pursue careers in AI or STEM.

(9) specific training data sets that can accelerate the development of AI and its application;

- Create comprehensive, well-indexed, and accessible compilations of federal government data. Invest in anonymization to permit the publication of as much data as possible in as raw a form as possible.
- Offer tax incentives to companies and organizations making anonymized data public.
- Provide grants to NGOs and non-profits to make data available in more accessible and readable formats.
- Promote data exchanges between industries and cross pollination of best practices.

(10) the role that "market shaping" approaches such as incentive prizes and Advanced Market Commitments can play in accelerating the development of applications of AI to address societal needs, such as accelerated training for low and moderate income workers (see https://www.usaid.gov/cii/market-shaping-primer);

Incentive prizes act to accelerate validation of new technologies, increase public awareness, and democratize innovation. They can play a critical role in shaping markets by reducing risks associated with certain large-scale investments that existing market providers and government purchasers might otherwise avoid.

Incentive prizes also provide a bridge between basic research and large-scale

implementation. They invite a broad range of innovators to apply their skills in the design of new solutions, and they provide a test bed in which those innovations can be assessed. Large-scale competitions can also lead to a virtuous cycle of development and investment. If innovations are shown to be effective through credible evidence and rigorous testing, risk-averse institutions will feel more comfortable making the necessary investments to bring these technologies to market. Thus, incentive prizes can often become proxies for Advanced Market Commitments.

In this sense, incentive prizes can play a particularly important role in developing AI technologies to address many societal needs. For instance, prizes involving AI development could focus on exploring Universal Basic Income or free access to education, with AI technologies serving as tools to broaden no- or low-cost access to quality education or to optimize the delivery of key public services to low income individuals.

(11) any additional information related to AI research or policymaking, not requested above, that you believe OSTP should consider.

XPRIZE believes that the government has a unique opportunity to stimulate and support interdisciplinary AI efforts to provide social and personal benefits. Current market incentives are encouraging investments in certain AI technologies with strong ROI, but there is not yet sufficient investment in the application of these technologies for the public good. The government can create a market for these applications and become a major customer for new and innovative solutions.

The government can also revisit policies that govern hedge funds, angel investing, and retirement investing so that all people can invest in disruptive technologies in their formative stages. Additionally, the government can promote debate, engagement, and information sharing around AI while ensuring the public has full access to federal data sets and best practices.

Respondent 101

Katherine Garges, citizen

1) The legal and governance implications of AI: Data ethics and privacy issues should receive emphasis.
(4) The social and economic implications of AI: I've been blogging about AI technology monthly from a non-scientist perspective for over 10 years. I watched livestream or video online of most of these workshops. They were outstanding in exploring public policy issues and included a lot of new information about AI-related public policy efforts. Most people can understand how AI technology works. Unlike many science and technology areas which require extensive preliminary educational preparation, AI involves basic reasoning methods

that all humans use and that provide an entry for understanding the science and technology. Better knowledge of the technology itself would reduce unrealistic fears about AI, improve understanding of realistic fears and result in better public policy. Government should fund programs to give all citizens, not just future AI workers, a basic understanding of how AI technology works.

Respondent 102

Michael Peters, American College of Radiology (ACR)

July 22, 2016

Attn: Terah Lyons
Office of Science and Technology Policy
Eisenhower Executive Office Building
1650 Pennsylvania Ave. NW,
Washington, DC 20504

Subject: (2016-15082; 81 FR 41610) Request for Information on Artificial Intelligence; Comments of the American College of Radiology

The American College of Radiology (ACR)—a professional organization representing more than 35,000 radiologists, radiation oncologists, interventional radiologists, nuclear medicine physicians, and medical physicists—appreciates the opportunity to respond to the White House Office of Science and Technology Policy's (OSTP) Request for Information (RFI) on "Artificial Intelligence" (AI) published in the Federal Register on June 27, 2016 (document number 2016-15082; 81 FR 41610). The ACR supports the federal government's efforts to leverage AI and machine learning to improve government services in general, and we urge additional federal support for, and collaboration with, professional associations and other stakeholders within specific fields of interest to ensure a safe and efficacious use of this technology.

The following comments on the questions enumerated in the RFI were compiled by members of the ACR Clinical Data Science Committee, ACR Commission on Informatics, and ACR Research. Individual contributing members are listed at the end of this submission.

ACR Responses to RFI Topics

1. The legal and governance implications of AI:

Health care institutions, radiology groups, and vendors planning to develop algorithms using source data such as electronic health record technology and/or patient diagnostic

imaging data need guidance from agencies on issues of patient consent and appropriate methods/best practices. Moreover, AI incorporation into clinical radiology practice can introduce new medico-legal risks and uncertainties. Related concerns could potentially discourage acceptance and proliferation of AI by providers.

2. The use of AI for public good:

AI could offer various benefits to medical imaging in the future, including augmenting the capabilities of radiologists to enhance their efficiency and accuracy, as well as reducing costs by improving the appropriateness and cost-effectiveness of medical imaging utilization.

The use of AI and machine learning in health care in general could be best applied to the areas of precision medicine, predictive analytics, and outcomes assessments. AI can streamline healthcare workflow and improve triage of patients (especially in acute care settings), reduce clinician fatigue, and increase the efficiency and efficacy of training. Moreover, shortages of medical experts to meet the needs of vulnerable and underserved populations in domestic and international settings could potentially be relieved, in part, by AI.

3. The safety and control issues for AI:

Safety standards should be identified to facilitate the proper development and monitoring of AI-driven technologies in medical imaging. This could be addressed through a combination of regulatory oversight and professional association validation or certification of algorithms. Federal agencies could also partner with professional and trade associations to develop standardized datasets for algorithm training and testing.

In addition to oversight over the technology, safety issues need to be addressed via training and best practices for practitioners on appropriate incorporation of AI into clinical radiology.

5. The most pressing, fundamental questions in AI research, common to most or all scientific fields:

The most universal AI research question is how to measure the effectiveness of the technology; however, the specific definitions/measures of effectiveness and testing methodologies would likely vary from field to field. In medicine, research into effectiveness should focus on areas such as diagnostic error reduction, improved accuracy, workflow enhancement, and efficiency gains. Moreover, research should explore how AI tools can be seamlessly integrated into clinical workflow and to what degree there is impact, both positive and negative, on clinical decision making and patient care outcomes.

6. The most important research gaps in AI that must be addressed to advance this field and benefit the public:

In terms of the application of AI to medical imaging, there is a need to define standards by which images and corresponding data should be structured to facilitate AI research. As mentioned above, research needs to also explore impact measurement of AI tools on image/data interpretation, diagnostic accuracy, and workflow efficiency.

8. The specific steps that could be taken by the federal government, research institutes, universities, and philanthropies to encourage multi-disciplinary AI research:

The Departments of Health and Human Services, Veterans Affairs, and Defense should increase grant opportunities to study and develop AI technologies in medical imaging. Federal agencies partnered with professional associations, academic institutions, patient advocates, and other organizations could develop and/or disseminate policy, ethical, scientific, and industry standards, including those related to interoperability and generalizability of AI-driven technologies. Standards around security, privacy, data-sharing, and the use of common datasets for researchers would facilitate the improved generalizability of algorithms. Importantly, added expertise in domains outside of traditional computer science and health information technology (e.g., image perception, human factors, and safety) should be consulted.

9. The specific training data sets that can accelerate the development of AI and its application:

The class of AI technologies that utilize machine learning techniques (neural networks, deep learning, etc.) require large data sets to learn relationships between inputs and outputs of information processing chains, or to discover and categorize patterns. The feasibility of acquiring and utilizing such large datasets varies tremendously across application domains. There are several significant impediments to acquiring such data for healthcare applications of AI, including the need to protect patient privacy, collect data across distributed sites and across multiple modalities (genomics, radiomics, pathology, etc.).

However, these problems have long been solved for clinical research initiatives, e.g., clinical trials and registries. The informatics platforms and processes developed to collect and create such repositories could be readily adapted for the healthcare AI domain. In addition, data from closed initiatives can be repurposed for AI research. The ACR has already begun to support the AI research of its members and partners in academia and industry, utilizing our TRIAD and DART platforms used for clinical imaging research.

De-identified training sets of various healthcare data types—including medical imaging data (MRI, CT, X-Ray, Ultrasound, PET)—covering the whole spectrum of pathologies need to be accessible in the public domain and validated to ensure that these sets meet government,

academic, and industry standards. The creation and curation of labeled data sets is a time consuming yet critical process in the development of AI technologies in health care and medical imaging.

11) Any additional information related to AI research or policymaking, not requested above, that you believe OSTP should consider:

The ACR believes AI has the potential to alleviate administrative burden and inappropriate utilization, and it could someday increase the precision and efficiency of certain medical services, including diagnostic imaging. This technology has the potential, with appropriate testing/validation and safeguards, to improve the value, safety, and appropriate utilization of medical imaging. AI also has the potential to shift more mundane tasks from radiologists and other physicians to machines, freeing radiologists to focus on patient care, including interpreting images and providing clinical consultations to other specialists.

The American College of Radiology appreciates this opportunity to provide input to OSTP staff and members of the National Science and Technology Council Subcommittee on Machine Learning and Artificial Intelligence. We welcome further communications on this and related topics. Please contact Gloria Romanelli, JD, Senior Director, Legislative and Regulatory Relations (XXXXXXXXX), or Michael Peters, Director of Legislative and Regulatory Affairs (XXXXXXXXX), if interested in reaching out to the ACR.

Sincerely,

James A. Brink, MD, FACR
Chair, Board of Chancellors
American College of Radiology

Keith Dreyer, DO, PhD, FACR
Chair, Commission on Informatics
American College of Radiology

Garry Choy, MD, MBA
Chair, Clinical Data Science Committee
American College of Radiology

Contributors:
ACR Commission on Informatics-Clinical Data Science Committee
Garry Choy, MD, MBA, Chair
Sawfan Halabi, MD
Kathy Andriole, PhD
Keith Dreyer, DO, PhD

Christoph Wald, MD, MBA
Woojin Kim, MD
Mike McNitt-Gray, PhD
Bob Nishikawa, PhD
James Stone, MD, PhD
Raym Geis, MD, FACR
Tony Scuderi, MD
Laura Coombs, PhD
Mike Tilkin, MS and ACR CIO

ACR Research
John Pearson, PhD

NOTE: These comments address the 11 RFI questions published in the Federal Register on June 27, 2016. The ACR is also planning to submit a formal, formatted version of this comment letter via fax.

Respondent 103

Tim Day, The Center for Advanced Technology and Innovation at the U.S. Chamber of Commerce

Comments on Artificial Intelligence:

As the world's largest business federation, the U.S. Chamber of Commerce represents the interests of more than three million businesses of all sizes, sectors, and regions. The Chamber's Center for Advanced Technology and Innovation promotes the role of technology in our economy and advocates for rational policy solutions that drive economic growth, spur innovation, and create jobs. Many of our members are working on breakthrough technology or will rely on new technology that goes beyond any existing regulatory or legislative framework. It is our responsibility to help create an environment that supports an innovative spirit by preventing unnecessary regulatory obstacles.
Artificial Intelligence (AI) is a technology with immense potential but equally vast misconceptions. AI refers to the engineering discipline of making machines intelligent but is often associated with the creation of human-like robots. In reality, millions of people have been positively impacted by this practical software engineering tool. Healthcare, environmental, transportation, and many other fields will see improvements due to this technology.

It is important to recognize the distinction between AI and machine learning. AI involves computers and systems that are able to solve problems without having the solutions hardcoded into the program. Machine learning, while often confused with AI, is actually a process that combines reading mined data with algorithm creation through AI.

AI allows digital devices to recognize and reply to objects, sounds, or patterns in order to make decisions and learn from the information given. For AI to reach its full potential there must be an open environment to allow for continuing research. Creating responsible AI that is programmed to work from strong data is one of the open challenges. There have been numerous reports on cases of discrimination in connection with machine learning. This demonstrates how biased data begets discriminatory results with machine learning algorithms. To avoid these failures, there is a need to address data gaps. Going forward, the federal government can contribute to enhancing this technology by releasing quality, robust datasets used in publicly deployed systems and lead efforts to determine how to solve these data gaps.

Other ways the government can assist in the development of this technology is supporting basic research into safety and bias questions, as well as examining the potential impact on the American economy and workforce. Machine learning should also be employed to increase government responsiveness and efficiency in transportation, education, healthcare, energy, environmental, urban planning, and many other sectors. We applaud the steps that the National Science and Technology Council is taking to use technology to make government more efficient and provide improved services to the public. These are the types of initiatives that support new discoveries in this field.

We also commend the White House Office of Science and Technology Policy (OSTP) for their efforts to educate the public and private sectors on the benefits of these technologies, including this request for information and for continuing to convene workshops on AI and machine learning. These discussions are essential to keep the technology industry moving forward, and we are very optimistic about the results that this platform can accomplish. To that end, we look forward to the release of OSTP's public report on AI later this year.

One of the most widely anticipated AI innovations is the development of self-driving cars, with companies like Toyota at the forefront of this vehicle revolution. Goldman Sachs has forecasted that the market for advanced drive assistance systems and autonomous vehicles will grow from the $3 billion market of 2015 to $96 billion in 2025 and $290 billion in 2035. While self-driving cars will make transportation easier and more accessible, these vehicles also have the potential to reduce emissions, cut commute times, and prevent fatal car crashes caused by human error.

AI can also transform education by adapting to student needs and providing more resources to educators. This technology would provide better personalization for students by addressing their individual needs and respond to strengths and skill gaps. This does not reduce the need for educators, but rather allows them to teach holistically while minimizing the risk of individual students falling behind. AI will also be able to help us better evaluate and calibrate our education system using comprehensive data to show us what our school systems are doing right, and how we can improve.

In the healthcare sector, doctors can use AI to predict septic shock, treat patients more comprehensively, and greatly reduce medical errors, which in the US account for over

250,000 deaths a year. Similarly, Vice President Biden's inspiring Cancer Moonshot initiative provides an opportunity for AI and machine learning to redefine how we use medical data to save lives.

With that said, AI operates within the parameters that humans permit. Hypothetical fears of rogue AI are based on the idea that machines can obtain sentience—a will and consciousness of its own. These suspicions fundamentally misunderstand what Artificial Intelligence is. AI is not a mechanical mystery, rather a human-designed technology that can detect and respond to errors and patterns depending on its operating algorithms and the data set presented to it. It is, however, necessary to scrutinize the way humans, whether through error or malicious intent, can wield AI harmfully. Accusations of discrimination by AI miss the fact that "discriminatory" data would have to have been fed biased information by its human creators. The solution to this problem is not to condemn the technology, but to explore the root of the issue. One possible solution that some academics have suggested is promoting diversity among systems engineers. In addition, there must be standards for quality data used to train systems. With these and many other questions in mind, companies like Google have established AI ethics boards, which go beyond legal compliance to examine the deeper implications and potential complications of emerging AI technologies.
Regulatory questions also arise with the rise of any transformational technology. The misconceptions around AI increase the likelihood of reckless regulatory decisions. It is vital to recognize that AI is well covered by existing laws and regulators with respect to privacy, security, safety, and ethics. Placing additional undue burdens would suppress the ability for this technology to continue growing. The policy questions that AI and machine learning raise are not so radically different than questions raised by technology that has preceded it. It is important to remember that artificial intelligence is still nascent, and it would be a mistake to attempt to address the issue with broad, overarching regulation. Instead, we believe that expert agencies that specialize in these areas should take the lead on setting standards. Innovation will be strengthened if industry-supported best practices are instituted in order to put protections in place without stifling growth.
The Center for Advanced Technology and Innovation will continue to support the ingenuity of our members and advance the issues most critical to them and the broader business community. We look forward to our continued work with the administration to promote technology development. AI's effect on business and our everyday lives could be game changing. Today, the internet is a tool that every industry relies on to do business. Tomorrow, the same will be said about Artificial Intelligence.

Respondent 104

Alex Kozak, X, a moonshot factory

X Response to OSTP's RFI on Artificial Intelligence

X (formerly Google X -- for more information see solveforx.com) appreciates the

opportunity to respond to OSTP's RFI. We're at an important moment in the intersection of artificial intelligence and the wider society and economy, and X is happy to contribute to this process. We believe that artificial intelligence and robotics will be crucial elements in projects aiming to help solve some of the world's biggest challenges. In our comments, we first briefly explain who we are and what we're working on. We note some of the areas that are ripe for the application of artificial intelligence, and other areas where further research and attention is needed.

Who We Are

X is the moonshot factory within Alphabet Inc. We are a team of engineers, scientists, makers, and inventors that applies audacious thinking to huge global problems to make the world a better place using technology. X incubates new breakthroughs in science or technology that could solve huge problems that affect millions or even billions of people.

All our projects must have three ingredients. First, the project must be focused towards solving a very big problem in the world—something that, if solved, could make millions or billions of people's lives better. Second, there must be a radical solution to that problem—a product or service that might even sound like science fiction. And lastly, there must be some breakthrough technology involved, along with evidence which gives us hope that the breakthrough technology might actually be within reach. Often the "breakthrough technology" identified includes some form of artificial intelligence, or related technology.

Our current list of public projects gives some indication of the kind of technology we think counts as a moonshot, and thus are indications of the kinds of projects we are likely to produce more of:

The Self-Driving Car Project is working to develop fully self-driving vehicles that have the potential to make our roads safer and increase mobility for the millions of people who cannot drive. Our ultimate goal is to help people get from A to B at the push of a button. In the project's seven year history, the vehicles in the test fleet have self-driven over 1.7 million miles on public roads, and we've launched testing programs in Mountain View, CA, Austin, TX, Kirkland, WA and Phoenix, AZ.

Project Loon is a system of balloons, carried by winds in the stratosphere, that can beam Internet access to rural, remote and underserved areas at speeds similar to today's LTE networks. Billions of people globally do not have reliable access to the internet. Project Loon aims to bring connectivity to the these underserved people. We've already conducted connectivity tests in Chile, Australia, Sri Lanka, and are preparing future tests in additional countries, such as Indonesia.

Makani hopes to accelerate the shift to clean, renewable energy by developing energy kites, a new type of wind turbine that can access stronger and steadier winds at higher altitudes

to generate more energy with less material.

Project Wing is developing an aerial delivery system using self-flying vehicles. We believe this technology could open new approaches to the transportation and delivery of goods—options that are cheaper, faster and more environmentally sensitive than what's possible today on the ground.

A number of Google's robotics teams also joined X in late 2015. X has long been the home of long-term projects that involve both hardware and software. We're currently looking at large, global problems where robots might provide new breakthrough solutions that could positively impact millions or even billions of people's lives.

The Opportunities of Artificial Intelligence

While the terms and categories like "robotics," "artificial intelligence," and "machine learning" will evolve in scope and meaning over time, an assumption apparent in our work is that robotics and mechanical systems can be combined with sophisticated computational techniques like machine learning to create useful and world-changing products that will help solve global challenges. Based on our experience developing these kinds of systems, and from the knowledge we've developed investigating hundreds of other ideas that didn't move forward, we firmly believe there are many global problems in the world today that could become more tractable and solvable with the careful application of AI, as an ingredient in a wider solution. These areas of opportunities are relatively well understood, and are being investigated in-depth by researchers around the world and the White House itself, so we only mention some briefly here:

Transportation of people and goods will be made more efficient, safer, and more environmentally friendly with the adoption of automation, and possibly more sophisticated forms of AI and machine learning, on our roads and in our skies. We'll also be able to manage the movement of goods more efficiently to better match supply and demand.
Artificial intelligence might also help mitigate the effects of climate change directly by opening up new opportunities for cleaner power generation or managing existing resources, or could help manage, monitor, and recommend interventions into changing ecosystems.
In educational settings, artificial intelligence could help address the needs of individual students to better tailor the style and pace of instruction.
In medicine, artificial intelligence and robotics could help doctors diagnose and treat conditions at lower cost and with greater accuracy.
Assistive care could be provided by robots and artificial intelligence improving the lives of the handicapped, the elderly and anyone else needing physical assistance in order to live more fulfilling and independent lives.

There will also be new discoveries and sectors created by artificial intelligence in areas that

we can't yet predict. Turning the power of machine learning on unsolved problems in science and basic research could help accelerate the pace of scientific discovery and our own understanding about the world that will then unlock new technologies or sectors that haven't yet been fully conceived.

Areas for Further Research and Attention from Policymakers

There are plenty of open research topics in the field. Based on our experiences in the real-world testing and rollout of products into the real world that are often described as using "artificial intelligence," we're attuned to some of the broader social and technical challenges involved. For example, the field of robot-human interaction is an emerging discipline that will help guide technologists and innovators in the design of robotic systems that will help them interact seamlessly in order to support human beings. How human drivers interact with autonomous vehicles, for example, is an important area of research for that project. More broadly, we generally agree that the research topics identified in "Concrete Problems in AI Safety," a joint publication between Google researchers and others in the industry, are the right technical challenges for innovators to keep in mind in order to develop better and safer real-world products: avoiding negative side effects (e.g. avoiding systems disturbing their environment in pursuit of their goals), avoiding reward hacking (e.g. cleaning robots simply covering up messes rather than cleaning them), creating scalable oversight (i.e. creating systems that are independent enough not to need constant supervision), enabling safe exploration (i.e. limiting the range of exploratory actions a system might take to a safe domain), and creating robustness from distributional shift (i.e. creating systems that are capable of operating well outside their training environment).

There is a strong role for sector-specific research into the challenges and opportunities of automation. For example, in 2014 the National Research Council published "Autonomy Research for Civil Aviation: Toward a New Era of Flight" an overview of both the opportunities and the research challenges to introducing more automation in aviation, at both the technical and regulatory level. In the autonomous vehicle context, RAND has published its own study ("Autonomous Vehicle Technology: A Guide for Policymakers") which similarly lists some of the opportunities and areas of possible research. Encouraging more sector-specific investigations like these in other fields such as medicine or education, or even within specific industries such as logistics, agriculture, or construction could help produce a more practical roadmap for how policymakers, technologists, and other stakeholders can encourage and better manage the implications of artificial intelligence.

There are still some important open questions around how best to manage the economic effects of artificial intelligence and automation. As Jason Furman from the Council on Economic Advisors recently pointed out, job training and education, in addition to wider government investments in basic research and private sector R&D, are practical ways that governments can help meet the challenge of declining labor force participation. But there does not seem to be a strong consensus around which skills to teach a new generation of

workers, and the best ways for educators and educational institutions to practically implement those new practices. There is a wide disconnect between seemingly widespread agreement that educational settings and practices need to evolve in an economy defined by rapid change, and practical real-world guidance and policies that could help implement the sort of shift that's required. Bridging that gap in both substance and leadership will be an important area of focus for governments.

Relatedly, as the economist Larry Summers has pointed out, two-thirds of the workforce that will be working in 2030 has already gone through their traditional education pathway. The conventional wisdom has been that job retraining or placement programs for adults are difficult to implement and are not always successful. But more research is warranted into the right ingredients for creating active and effective job retraining or job placement programs that will lead to a meaningful increase in labor force participation. And, given recent uptick in popularity of more flexible working environments, governments should better understand how to measure and incorporate those workers and working environments into measures of the health of our labor force, and how those sorts of labor arrangements can provide meaningful economic opportunity for participants.

There are also ways that governments might encourage the application of artificial intelligence to help solve global problems. The integration of artificial intelligence and automation into regulated or managed industries means that governments will need to grapple with how to apply old rules and procedures when faced with new technical facts that break the mold. Governments should invest more in developing sector-specific technical expertise within existing regulatory agencies to better equip them to understand and manage the unique challenges associated with specific implementations of artificial intelligence. Governments and policymakers should also endeavor to create more flexibility within regulatory frameworks that could better accommodate automation and machine learning, for example when technology fills certain roles that had been traditionally managed by people. And beyond that, there may be ways that government agencies and institutions could use machine learning, artificial intelligence, or related areas like robotics or automation, to better fulfill their statutory mandates or do their existing work more efficiently.

Protecting human dignity, including the right to privacy and providing new opportunities to live fulfilling lives will be an important public policy goal to achieve as artificial intelligence becomes more commonplace. As industries and sectors evolve and begin to incorporate artificial intelligence and automation over the coming years, they should be allowed the space and opportunity to demonstrate that these important goals can be met without early and prescriptive rules or policies that risk stifling or predetermining the kinds of technologies and techniques available to innovators. The recent NTIA multistakeholder process to define privacy best practices for unmanned aircraft systems is a good example of how governments can create a space for best practices to develop organically without pre-defining a specific outcome. Technology in these sectors will evolve quickly, and could itself

present novel ways of protecting consumer privacy and dignity.

In sum, artificial intelligence and related technologies like robotics and automation will play an important role in solving some of the world's big challenges. Government encouragement of more research into the opportunities and implications of its adoption within specific economic, industrial, or social sectors is a useful way to produce tangible guidance for how governments, innovators, and other stakeholders can help encourage that integration quickly and responsibly.

Respondent 105

Stephen Smith, Association for the Advancement of Artificial Intelligence

This submission is an organizational response from AAAI - the Association for the Advancement of Artificial Intelligence. AAAI is an international organization, headquartered in California, the largest AI Society in the world, with over 3000 members. This response was developed by the Government Relations Committee of the AAAI Executive Council, in coordination with the President of AAAI.

(1) The legal and governance implications of AI

The deployment of AI systems in increasingly more complex decision-making settings raises important issues around agency, ownership, fairness and responsibility. As a first step, the US should convene a working group comprised of legal experts, AI researchers, and other stakeholders (e.g., AI system manufacturers, insurance companies, consumer advocates, etc.) to explore issues of culpability for AI system decisions, and develop model laws and regulations. Another concern that requires study is that of providing protection against potential power asymmetries that might arise (e.g., through manipulation and/or exploitation of AI systems) between those with insight and understanding of AI technologies and those without it. Finally, since AI systems often rely on personal or sensitive data, they also face the same general data privacy issues that other software and database systems do. Given the potentially unique character of AI systems relative to this last set of issues, any broader forum convened to discuss and address data privacy should include representation from the AI research community.

Laws and regulations in each of these areas will need to evolve over time, but individual cases make bad law, so it is important that legislatures put some reasonable statutes in place. The legal aspects of AI systems are complex, and as such it is recommended that they be approached incrementally as a function of both (1) degree of the system autonomy permitted and (2) problem domain (e.g., autonomous vehicles, medical diagnosis) rather than pursuing discipline-wide blanket laws.

(2) The use of AI for public good

There is tremendous potential for AI to serve the public good by creating decision making tools that incorporate a comprehensive set of sensor signals into highly-accurate models that enable both rapid response to crises, as well as medium and long term planning. AI tools are already being applied to optimize many aspects of city services including utilities, transportation, law enforcement, and poverty mitigation. AI tools have also been shown to be useful for detecting manipulation of social media and many forms of financial fraud.

Looking ahead, we anticipate many other high-impact social good applications in the short to medium term, including early detection of serious medical conditions from routine test data; more efficient healthcare delivery including home-based care; improved ecosystem and resource management; personalized education; detection of public health hazards (e.g., presence of lead paint) from analysis of diverse data; and automated testing of complex software/hardware systems that will be ultimately operated by people to ensure safety. In general, AI has had strong success (often surpassing human expertise) in problem domains that are narrowly scoped and well structured; and applications that possess these characteristics are prime candidates for short-term benefit.

(3) The safety and control issues for AI

When AI technology is incorporated into systems that contribute to high-stakes decision-making, errors can have severe consequences. In the past, AI research and development has not always attended to these risks. Research is urgently needed to develop and modify AI methods to make them safer and more robust. A discipline of AI Safety Engineering should be created and research in this area should be funded. This field can learn much by studying existing practices in safety engineering in other engineering fields, since loss of control of AI systems is no different from loss of control of other autonomous or semi-autonomous systems. AI technology itself can also contribute to better control of AI systems, by providing a way of monitoring the behavior of such systems to detect anomalous or dangerous behavior and safely shut them down. Note that a major risk of any computer-based autonomous systems is cyber attack, which can give attackers control of high-stakes decisions.

There are two key issues with control of autonomous systems: speed and scale. AI-based autonomy makes it possible for systems to make decisions far faster and on a much broader scale than humans can monitor those decisions. In some areas, such as high speed trading in financial markets, we have already witnessed an “arms race” to make decisions as quickly as possible. This is dangerous, and government should consider whether there are settings where decision-making speed and scale should be limited so that people can exercise oversight and control of these systems.

Most AI researchers are skeptical about the prospects of “superintelligent AI”, as put forth in Nick Bostrom’s recent book and reinforced over the past year in the popular media in

commentaries by other prominent individuals from non-AI disciplines. Recent AI successes in narrowly structured problems (e.g., IBM's Watson, Google DeepMind's Alpha GO program) have led to the false perception that AI systems possess general, transferrable, human-level intelligence. There is a strong need for improving communication to the public and to policy makers about the real science of AI and its immediate benefits to society. AI research should not be curtailed because of false perceptions of threat and potential dystopian futures. OSTP's recent sequence of workshops on the future of AI is a great first step in this direction.

(4) The social and economic implications of AI.

The social and economic implications of AI are difficult to predict. It is likely that AI-based technology will improve productivity in many industries, but it is unclear how the benefits of these productivity improvements will be distributed through the economy. AI systems continue to be developed to improve education, particularly in STEM fields, through personalization and one-on-one tutoring. AI systems can also improve access through natural language interaction and virtual presence. The government should fund research to monitor social/economic impacts of AI systems by collecting statistics and studying how AI systems affect the nature of work, the growth of productivity, and the distribution of wealth. Regular reports to government should be required, so that appropriate policies can be introduced if they become necessary.

Care should be taken to distinguish economic impacts due to AI systems from those that are due primarily to other factors (e.g., other information technology, outsourcing practices). The government should seek to build greater in-house technical expertise in AI as a practical means of gaining understanding and getting on top of these issues.

(5) The most pressing, fundamental questions in AI research, common to most or all scientific fields

- How can computers acquire broad commonsense knowledge including knowledge of appropriate and inappropriate behavior in social interactions? Existing methodologies (supervised learning, hand-coded knowledge bases) have so far failed to provide this knowledge. Such knowledge is important for allowing AI systems to operate in open environments and especially to interact effectively with people.
- How can AI systems best augment human decision-making and vice versa, and become "human aware"? What kinds of interactions (explanations, visualizations, transparent structures) will make it easy for people to collaborate effectively, safely, and reliably with AI systems? How can humans teach AI systems to expand their knowledge? Interdisciplinary research that engages human factors, cognitive psychology, and AI research communities toward these challenges is needed.
- How can AI systems be made robust to un-modeled aspects of the world? No system can model (or be aware of) the full complexity of its surroundings. Living systems appear to

behave robustly even in the presence of these "unknown unknowns". One important direction is to develop ways that AI systems can introspect about their capabilities and limitations. Methods for continual self-monitoring to detect failures and limitations are needed.

- Modern AI systems continue to learn from their experiences after they are deployed. Methods are needed for ensuring that this adaptation respects safety and functionality constraints. Formal verification techniques may be useful but are limiting for software systems that adapt, plan and learn and will require new methods; self-monitoring capabilities may be essential.
- What are the limits of AI systems? We have computability theory for all of Computer Science, the theory of inductive inference and Probably Approximately Correct (PAC) learning for machine learning, and intractability results for various logical representation systems. Can tighter formal limits or better theoretical understanding be achieved for specific classes of AI systems/methodologies (e.g., deep learning)?
- How can AI systems help us understand the brain and intelligent human behaviors, and advance fundamental understanding of intelligence?

(6) The most important research gaps in AI that must be addressed to advance this field and benefit the public

As indicated in the response to Question 3 above, AI is already benefiting the public in several different areas, and answers to the fundamental questions listed in Question 5 would surely open up AI systems to a much broader public benefit. Among the important research gaps embodied in these questions are the following:

- Data and methodological bias – Much of the potential of AI systems follows from the ability to extract patterns from large data sets and turn these results into forms of actionable information and advice. However there are several sources of bias that can impact the accuracy of the conclusions that are drawn. If the data were collected in a biased way or if data quality (noise, missing values, precision) exhibits biases, then the extracted patterns can be biased. Likewise, biases can come from the assumptions made by the algorithms applied to extract patterns and draw conclusions (e.g., active learning methods, cost-sensitive methods, etc.). How can we define "bias"? How can we detect it? How can we eliminate or control it?
- Collaborative decision-making – In the short and medium term, mechanisms for AI systems interacting with and supporting human decision-makers (in contrast to fully autonomous AI systems) will constitute the primary path to application and benefit, and this requirement exposes several gaps in current capabilities. Very few AI systems are able to explain their reasoning, either through summarization of logical inference, visualization of key consequences, or simulation of expected decision behaviors. This capability is fundamental to broader application of AI systems: (a) to allow people and computers to work well together (effectiveness, safety, reliability), (b) to enable people to attain appropriate levels of trust in AI systems and promote further automation, (c) to support post mortem examination of decision making for credit assignment and possibly for legal

purposes, and (d) to help AI system developers detect and repair errors in the system.

- Ethical decision-making – As we move toward applying AI systems in more mission critical types of decision-making settings, AI systems must consistently work according to values aligned with prospective human users and society. Yet it is still not clear how to embed ethical principles and moral values, or even professional codes of conduct, into machines.

(7) The scientific and technical training that will be needed to take advantage of harnessing the potential of AI technology, and the challenges faced by institutions of higher education in retaining faculty and responding to explosive growth in student enrollment in AI related courses and courses of study

There is currently a significant pull of academic research and teaching expertise toward AI companies due to financial packages that universities cannot match, computing facilities and other infrastructure that is not otherwise available, etc. This trend benefits short-term application of AI research but hurts more fundamental, academic AI research. It also negatively impacts the training of future AI researchers and practitioners.

Universities are allocating faculty positions to AI-related areas. However, for prospective faculty members to succeed, they need to be able to obtain research funds from Federal sources (including NSF, ONR/ARL/AFOSR, DARPA, NIST, NIH, etc.). Congress needs to allocate additional funds to these agencies to enable them to invest in AI-related research. Further, it is important that the government continue to advocate and invest in longer-term, fundamental AI research. It often takes many years to achieve breakthroughs that are key to solving particular societal problems, and no one has the crystal ball to fully predict what these will be. [Note that such a commitment to sustained funding would make such faculty positions more attractive both to potential faculty members and to their institutions, in addition to boosting the long-term benefit of AI research to society.] Government could also improve the training of future AI researchers by greatly increasing the funding available for NSF Graduate Fellowships.

There is also need for more basic education and outreach activities to the general public on the capabilities and potential of AI technologies. Steps should be taken to make introductions to AI topics such as machine learning, planning, knowledge representation, and robotics part of the core undergraduate curricula for non-computing majors.

(8) The specific steps that could be taken by the federal government, research institutes, universities, and philanthropies to encourage multidisciplinary AI research

It is important to have sustained funding for multidisciplinary research. The achievement of systems with robust common sense reasoning and human-level decision-making expertise will require sustained collaboration between disparate and (currently) largely disconnected research communities in psychology, social sciences, and AI. There is also increasingly a

need for policy and technology research to mutually inform and align.

NSF has had several interdisciplinary research programs over the years (e.g., ITR, CDI, SEES), but each program only lasts for a few years. In order for a faculty member to take the risk of building an interdisciplinary research program, there needs to be the prospect of continuing funding opportunities over the long term. This prospect can also encourage universities to create interdisciplinary faculty positions to attract candidates that may not fit neatly into one discipline.

Many important research areas in AI cross government agency boundaries (e.g., the Departments of Justice, Commerce, Energy, and Defense as well as NSF and NIH). The government should create cross-agency working groups to develop research roadmaps and funding programs to promote this research. Important research aimed at social good crosses levels from city governments to regional utilities to law enforcement at all levels (including municipal, state, FBI, Coast Guard, and Border Control). Mechanisms need to be created that support the development of data sets and research programs spanning these levels.

(9) Specific training data sets that can accelerate the development of AI and its application.

To apply supervised learning to acquire broad, commonsense knowledge, labeled data sets are needed about common sense situations. Similarly, to give AI systems better understanding of appropriate (ethical) behavior, data sets are needed describing decision making situations and the ethical and unethical actions that could be taken in those situations.

The promotion of open data initiatives (be it data about cities, government, biomedical experimentation, the environment, materials engineering, education, etc.) would likely accelerate AI application development in many problems of societal interest/benefit, since AI researchers often end up pursuing problems where data is openly available.

(10) The role that "market shaping" approaches such as incentive prizes and Advanced Market Commitments can play in accelerating the development of applications of AI to address societal needs, such as accelerated training for low and moderate income workers

Incentive prizes, if large enough, can have a major impact. However, they generally reward people who already have enough resources that they can take the risk of spending their own funds even if the probability of winning a prize is low. Providing some form of participant support for non-traditional teams that wish to compete for incentive prizes is critical for broadening participation.

Current government acquisition rules, particularly in DoD, are acting as disincentives to the development of advanced technology. It can take upwards of a decade or more for AI

systems to be proven and transitioned into operations. The government should define new processes for the certification of adaptive/AI technology so that DoD and other government agencies can easily acquire it.

Respondent 106

Alex Kozak, X, a moonshot factory

X Response to OSTP's RFI on Artificial Intelligence

X (formerly Google X -- for more information see solveforx.com) appreciates the opportunity to respond to OSTP's RFI. We're at an important moment in the intersection of artificial intelligence and the wider society and economy, and X is happy to contribute to this process. We believe that artificial intelligence and robotics will be crucial elements in projects aiming to help solve some of the world's biggest challenges. In our comments, we first briefly explain who we are and what we're working on. We note some of the areas that are ripe for the application of artificial intelligence, and other areas where further research and attention is needed.

Who We Are

X is the moonshot factory within Alphabet Inc. We are a team of engineers, scientists, makers, and inventors that applies audacious thinking to huge global problems to make the world a better place using technology. X incubates new breakthroughs in science or technology that could solve huge problems that affect millions or even billions of people.

All our projects must have three ingredients. First, the project must be focused towards solving a very big problem in the world—something that, if solved, could make millions or billions of people's lives better. Second, there must be a radical solution to that problem—a product or service that might even sound like science fiction. And lastly, there must be some breakthrough technology involved, along with evidence which gives us hope that the breakthrough technology might actually be within reach. Often the "breakthrough technology" identified includes some form of artificial intelligence, or related technology.

Our current list of public projects gives some indication of the kind of technology we think counts as a moonshot, and thus are indications of the kinds of projects we are likely to produce more of:

The Self-Driving Car Project is working to develop fully self-driving vehicles that have the potential to make our roads safer and increase mobility for the millions of people who cannot drive. Our ultimate goal is to help people get from A to B at the push of a button. In the project's seven year history, the vehicles in the test fleet have self-driven over 1.7 million miles on public roads, and we've launched testing programs in Mountain View, CA,

Austin, TX, Kirkland, WA and Phoenix, AZ.
Project Loon is a system of balloons, carried by winds in the stratosphere, that can beam Internet access to rural, remote and underserved areas at speeds similar to today's LTE networks. Billions of people globally do not have reliable access to the internet. Project Loon aims to bring connectivity to the these underserved people. We've already conducted connectivity tests in Chile, Australia, Sri Lanka, and are preparing future tests in additional countries, such as Indonesia.
Makani hopes to accelerate the shift to clean, renewable energy by developing energy kites, a new type of wind turbine that can access stronger and steadier winds at higher altitudes to generate more energy with less material.
Project Wing is developing an aerial delivery system using self-flying vehicles. We believe this technology could open new approaches to the transportation and delivery of goods—options that are cheaper, faster and more environmentally sensitive than what's possible today on the ground.
A number of Google's robotics teams also joined X in late 2015. X has long been the home of long-term projects that involve both hardware and software. We're currently looking at large, global problems where robots might provide new breakthrough solutions that could positively impact millions or even billions of people's lives.

The Opportunities of Artificial Intelligence

While the terms and categories like "robotics," "artificial intelligence," and "machine learning" will evolve in scope and meaning over time, an assumption apparent in our work is that robotics and mechanical systems can be combined with sophisticated computational techniques like machine learning to create useful and world-changing products that will help solve global challenges. Based on our experience developing these kinds of systems, and from the knowledge we've developed investigating hundreds of other ideas that didn't move forward, we firmly believe there are many global problems in the world today that could become more tractable and solvable with the careful application of AI, as an ingredient in a wider solution. These areas of opportunities are relatively well understood, and are being investigated in-depth by researchers around the world and the White House itself, so we only mention some briefly here:

Transportation of people and goods will be made more efficient, safer, and more environmentally friendly with the adoption of automation, and possibly more sophisticated forms of AI and machine learning, on our roads and in our skies. We'll also be able to manage the movement of goods more efficiently to better match supply and demand.
Artificial intelligence might also help mitigate the effects of climate change directly by opening up new opportunities for cleaner power generation or managing existing resources, or could help manage, monitor, and recommend interventions into changing ecosystems.
In educational settings, artificial intelligence could help address the needs of individual students to better tailor the style and pace of instruction.

In medicine, artificial intelligence and robotics could help doctors diagnose and treat conditions at lower cost and with greater accuracy.
Assistive care could be provided by robots and artificial intelligence to provide assistance to differently-abled people, people in old age, or anyone who may desire physical assistance.

There will also be new discoveries and sectors created by artificial intelligence in areas that we can't yet predict. Turning the power of machine learning on unsolved problems in science and basic research could help accelerate the pace of scientific discovery and our own understanding about the world that will then unlock new technologies or sectors that haven't yet been fully conceived.

Areas for Further Research and Attention from Policymakers

There are plenty of open research topics in the field. Based on our experiences in the real-world testing and rollout of products into the real world that are often described as using "artificial intelligence," we're attuned to some of the broader social and technical challenges involved. For example, the field of robot-human interaction is an emerging discipline that will help guide technologists and innovators in the design of robotic systems that will help them interact seamlessly in order to support human beings. How human drivers interact with autonomous vehicles, for example, is an important area of research for that project. More broadly, we generally agree that the research topics identified in "Concrete Problems in AI Safety," a joint publication between Google researchers and others in the industry, are the right technical challenges for innovators to keep in mind in order to develop better and safer real-world products: avoiding negative side effects (e.g. avoiding systems disturbing their environment in pursuit of their goals), avoiding reward hacking (e.g. cleaning robots simply covering up messes rather than cleaning them), creating scalable oversight (i.e. creating systems that are independent enough not to need constant supervision), enabling safe exploration (i.e. limiting the range of exploratory actions a system might take to a safe domain), and creating robustness from distributional shift (i.e. creating systems that are capable of operating well outside their training environment).

There is a strong role for sector-specific research into the challenges and opportunities of automation. For example, in 2014 the National Research Council published "Autonomy Research for Civil Aviation: Toward a New Era of Flight" an overview of both the opportunities and the research challenges to introducing more automation in aviation, at both the technical and regulatory level. In the autonomous vehicle context, RAND has published its own study ("Autonomous Vehicle Technology: A Guide for Policymakers") which similarly lists some of the opportunities and areas of possible research. Encouraging more sector-specific investigations like these in other fields such as medicine or education, or even within specific industries such as logistics, agriculture, or construction could help produce a more practical roadmap for how policymakers, technologists, and other stakeholders can encourage and better manage the implications of artificial intelligence.

There are still some important open questions around how best to manage the economic effects of artificial intelligence and automation. As Jason Furman from the Council on Economic Advisors recently pointed out, job training and education, in addition to wider government investments in basic research and private sector R&D, are practical ways that governments can help meet the challenge of declining labor force participation. But there does not seem to be a strong consensus around which skills to teach a new generation of workers, and the best ways for educators and educational institutions to practically implement those new practices. There is a wide disconnect between seemingly widespread agreement that educational settings and practices need to evolve in an economy defined by rapid change, and practical real-world guidance and policies that could help implement the sort of shift that's required. Bridging that gap in both substance and leadership will be an important area of focus for governments.

Relatedly, as the economist Larry Summers has pointed out, two-thirds of the workforce that will be working in 2030 has already gone through their traditional education pathway. The conventional wisdom has been that job retraining or placement programs for adults are difficult to implement and are not always successful. But more research is warranted into the right ingredients for creating active and effective job retraining or job placement programs that will lead to a meaningful increase in labor force participation. And, given recent uptick in popularity of more flexible working environments, governments should better understand how to measure and incorporate those workers and working environments into measures of the health of our labor force, and how those sorts of labor arrangements can provide meaningful economic opportunity for participants.

There are also ways that governments might encourage the application of artificial intelligence to help solve global problems. The integration of artificial intelligence and automation into regulated or managed industries means that governments will need to grapple with how to apply old rules and procedures when faced with new technical facts that break the mold. Governments should invest more in developing sector-specific technical expertise within existing regulatory agencies to better equip them to understand and manage the unique challenges associated with specific implementations of artificial intelligence. Governments and policymakers should also endeavor to create more flexibility within regulatory frameworks that could better accommodate automation and machine learning, for example when technology fills certain roles that had been traditionally managed by people. And beyond that, there may be ways that government agencies and institutions could use machine learning, artificial intelligence, or related areas like robotics or automation, to better fulfill their statutory mandates or do their existing work more efficiently.

Protecting human dignity, including the right to privacy and providing new opportunities to live fulfilling lives will be an important public policy goal to achieve as artificial intelligence becomes more commonplace. As industries and sectors evolve and begin to incorporate artificial intelligence and automation over the coming years, they should be allowed the

space and opportunity to demonstrate that these important goals can be met without early and prescriptive rules or policies that risk stifling or predetermining the kinds of technologies and techniques available to innovators. The recent NTIA multistakeholder process to define privacy best practices for unmanned aircraft systems is a good example of how governments can create a space for best practices to develop organically without pre-defining a specific outcome. Technology in these sectors will evolve quickly, and could itself present novel ways of protecting consumer privacy and dignity.

In sum, artificial intelligence and related technologies like robotics and automation will play an important role in solving some of the world's big challenges. Government encouragement of more research into the opportunities and implications of its adoption within specific economic, industrial, or social sectors is a useful way to produce tangible guidance for how governments, innovators, and other stakeholders can help encourage that integration quickly and responsibly.

Respondent 107

David Enabnit, These comments are my own.

Question 7: The explosive growth of industry attention to AI may be draining universities of students and faculty. While the popularity of the topic may provide replacements, a high turnover rate would negatively affect the intellectual maturity of the research infrastructure providing the basic research on AI. OSTP and NSTC should undertake to assess the quantity and quality of research faculty in core AI disciplines to determine if such a problem is arising. Federal intervention may be needed in the form of research grants, additional graduate and post-doctoral financial support, and increased funding for basic research in AI disciplines at federal laboratories to provide stability.

It is also unclear that industry, which is benefiting from this pool of basic research talent, is contributing to (re)build the talent pool. OSTP and NSTC should undertake to determine if industry funding for basic research, both within their companies and at universities, reflects the gains they are receiving and the needs of the community. Appropriate action should be taken based on the result.

Question 10: The economics of AI might make "market shaping" activities unnecessary. The examples cited in the RFI would be helpful where there is a large, up-front investment or high risk, e.g. the space industry and conquering cancer. AI has a much lower entry cost and the economics are so compelling that industry seems to be investing billions of dollars even at the present state of technological immaturity. It might be more appropriate to broadly educate government employees at all levels, municipal, state and federal, on AI and its applicability to government problems so those employees become paying customers with real applications. The 4 recent OSTP-sponsored workshops are an example of such educational activities. Federal contracts or grants for scalable AI deployments on actual

government problems could add economic heft – just avoid "demonstrations" which seldom lead to operational deployments.

Question 11: OSTP is correct to highlight AI as perhaps the most consequential effort ever undertaken. It should be treated as such. Policies should be developed and implemented to see that the U.S. stays at the leading edge and that the U.S. receives the full benefit of advances. For example, AI will most likely yield to large numbers of highly trained people working on the problems for many decades. U.S. companies and universities are drawing the brightest from around the world, but U.S. immigration policy must be aligned to insure that, once trained, they stay here and contribute to the critical mass needed for progress.

It appears that federal funding of research may be de-emphasizing basic research and directing a larger percentage towards lower risk applied research. However because of the large economic potential and immediate commercial applications, industry is aggressively funding applied research and development on artificial intelligence. OSTP is well positioned to query federal funders to assess this situation, and if valid, to steer federal funding for AI back to basic research where it is needed and where it would complement industry's effort.

Finally, it's not artificial intelligence per se that we seek, but its consequence – taking information, concepts and problems and transforming them into understanding and solutions. Intelligence need not be artificial (machine intelligence) to be of value. OSTP should include augmenting natural intelligence (human intelligence) within their scope. Computer-driven individually tailored education; the underlying science on collaboration, brainstorming and other such fads; and chemical and electrical stimulation of intelligence are 3 examples where fundamental research might contribute to our capabilities.

Respondent 108

Graham Gilmer, Booz Allen Hamilton

Ethics in the Age of AI
Booz Allen Hamilton response to
"Preparing for the Future of Artificial Intelligence"

Tech billionaires, the media, and entertainment companies are creating hysteria around Artificial Intelligence (AI). Focus on unlikely fears of "killer robots," has the potential to derail AI R&D from delivering valuable contributions to the American economy and military competitiveness on the world stage, as well as drowning out more legitimate, focused ethical concerns. In place of panic, the Office of Science and Technology and Policy (OSTP) should calmly and boldly set the agenda for the future by establishing a set of guiding principles around AI ethics and policy.
Much has been discussed about AI disrupting the workforce and displacing jobs. While this is a potential outcome, OSTP has an opportunity to chart a different path for the U.S. By

establishing a national set of ethical principles, as well as guidelines for federal research and development in this space, OSTP can position the U.S. as a world leader in AI. Success will ultimately result in high-quality job creation and calm fears around the adoption of AI. We will also enjoy the benefits of breakthroughs in health and science discoveries and development of new defensive capabilities to deter threats at home and abroad.

At Booz Allen Hamilton, a technology and strategy firm, our data scientists and computer engineers wield cutting-edge machine learning and other AI techniques in order to benefit the public good and create tangible value for the government. From our purview, we see a few overarching themes that are driving the need for a national focus on ethics in AI:

- Lifestyle Changes: AI has already begun to permeate citizens' daily lives, something that will only increase. We expect to see a future where AI will decide the length of jail sentences, whether you are stopped by police, or if your cancer is diagnosed. There will be many benefits, but also real consequences to human life if the AI is wrong, meaning a great deal of thought is needed about biases, error, and policy.
- Unintended Consequences: Machines that are designed without the input of diverse thinking may lead to systems that cannot grasp the nuances of social and cultural norms. Ethics training gaps in computer science and related fields will exacerbate this problem, which provides OSTP with a unique opportunity to shape this space. Without proper controls, systems may teach themselves bad or unethical behaviors if they help to achieve an overly focused goal. Examples include algorithms and chat bots gone awry with racist, intolerant or inhumane outputs.
- Privacy/Transparency Concerns: As machines become better at detecting identities and predicting how consumers are likely to behave, the public will have strong privacy and security concerns. At the same time, they will become increasingly frustrated with AI that can act as a "black box" in seemingly unpredictable ways. Citizens will demand more control of their own data, and greater transparency from increasingly powerful inference-making AI. Industry leaders will need to embrace open source culture for large organizations, while simultaneously preserving reasonable expectations of privacy at the individual level.

As with other rapidly evolving technological frontiers, even the most carefully designed and adaptive regulatory agenda may not be able to keep pace with changes in AI. As such, we do not believe creating a complex regulatory system is a feasible approach. We also cannot stand idle. While tremendous gains no doubt beckon from the unbridled enthusiasm of American discovery, to safeguard against harm we recommend that OSTP lead the nation, and the world, in establishing ethical and policy principles related to AI. Laying the ethical foundation for the progress to follow will ensure that these technologies and all associated policies related to them will reflect sober-minded attention to providing benefit for all of mankind. Creating these ethical and policy norms will be imperative, and OSTP is uniquely positioned to do so with the credibility derived from a long history of actions in fields ranging from STEM education to nuclear non-proliferation.

Choosing Guiding Principles

AI is truly a grand experiment on all of humanity. Advances in AI will change our society and

our world by revolutionizing how we live, work, and interact. It will touch every aspect of our lives. Like many experiments, AI offers unprecedented possibilities for human gain, but care must be made to avoid and minimize harm on the subjects of the experiment, namely ourselves and the most vulnerable among us. We must ensure that AI is beneficial, not harmful, to human welfare. To that end, we call on OSTP to champion a set of principles to ensure a bright future for all.

When creating guiding principles applicable to fundamental research and to policy alike, OSTP may leverage the hard-earned knowledge of ethical practices from other fields of human endeavor. As we envision AI research as an experiment, our source for a set of principles becomes obvious: the three fundamental principles of human subject research as set out by the Belmont Report(1). These principles are Beneficence, Justice, and Respect; they originated in bioethics, and we see them as equally applicable to AI.

Beneficence

Beneficence is the most basic of commands for any ethical guidelines, namely to do no harm. Developers of AI must ensure whatever methods they are developing protect the physical, mental and social well-being of all, avoiding harm both to individuals and to the community as a whole.

The harm that must be avoided extends far beyond clichés of robot rebellions. AI algorithms today are being used to predict the likelihood of criminal recidivism and inform parole decisions. However, software was found to be twice as likely to mistakenly flag black defendants as being at a higher risk of committing future crimes, and twice as likely to incorrectly flag white defendants as low risk. We note how easy it is for subtle bias in human society to sneak in to what seems like an impartial algorithm. Machine learning is only as good as the data it trains on, and avoiding harm thus necessarily includes creating safeguards to prevent societal and racial biases being learned by the algorithm. We note that one action that might have avoided this particular harm is having a more diverse AI and computer developer workforce.

(1)http://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/

Justice

Justice means to treat people and issues fairly. In the scope of human subject research, it means that the research should provide a fair distribution of costs and benefits to potential subjects, particularly when the subjects are of disadvantaged groups. If the subjects are members of a particular minority, such as people suffering from a particular disease, the research benefits should be focused on improving the outcomes of people from that minority group.

In making AI just, we emphasize the importance of ensuring the benefits of AI apply to all people, not simply privileged elites. For example, smart devices have the possibility of significantly improving health outcomes. However, they are likely to be purchased only by those who can afford them and we should consider whether there are more benefits or costs to those who cannot afford to upgrade.

Respect

In the context of human subject research, respect means that each individual has the right to decide whether to participate in a study while exercising informed consent. Subjects, to the degree that they are capable, must be given the opportunity to choose what shall or shall not happen to them. There have been notable examples where social media companies subjected their users to experiments aimed at improving their AI algorithms. We must think beyond terms of service to define a consent process that includes three elements: information, comprehension, and voluntariness.
For many AI applications, it is difficult to receive informed consent from every person potentially impacted by new technologies, which could be the entire population. That wide scope makes this principle all the more important. Having a principle of open-source by default is one key way to keep the public informed of what developers are doing. Ideally, this is coupled with other methods to increase comprehension of AI work, such as workplace development, seminars, or educational outreach. Consent can be approximated via the political process, referenda, or opt-out procedures. Respect means a constant striving to make sure every person's view is heard.

Further Considerations

Disruption from AI should be expected. The impacts will be disparate and require input from many stakeholders. Broad collaboration in the global commons will be crucial. The United States, and OSTP in particular, has the opportunity to establish the country as a world leader in AI. The United States can take a large role in shaping global interactions through its behavior. Establishing ethical and policy principles for AI will provide the requisite credibility and motivation for these actions.

The principles described above deliberately do not include specific policy recommendations. Instead, they are intended to provide the underpinnings of a comprehensive strategy for ensuring we realize the benefits of AI with minimal negative consequences. As OSTP, and National Science and Technology Council (NSTC) Subcommittee on Machine Learning and Artificial Intelligence examines policy related to AI informed by human subject research ethical guidelines, they may consider a few concrete steps that will reinforce the adoption of these principles:

1) Spearheading a National AI Ethics and Policy Roadmap that addresses:
· Following the bioethics model and establishing approval mechanisms for AI research analogous to human subject research.
· Taking active steps to ensure the inclusion of unique perspectives and diversity within the workforce developing AI systems.
· Guidance to federal agencies in considering how applications of AI relating to their data or centralized systems will interact with ethics and policy principles.

2) Establishing a Science Ambassador for AI Ethics and Policy for issues impacting the global commons, including involvement with:

· Projects that the U.S. or other countries conduct that will have the potential to impact people globally.
· Encouraging openness to defend against risky behavior by bad actors.
· Ensuring that as the U.S. progresses in the field, it values humanity alongside technology by working with stakeholders to address opportunities and concerns that impact the public.

3) Establish a culture of ethical and fair AI research throughout the country, through actions such as:
· Establishing the principle of open source by default. Having open AI means that the systems making more and more decisions are available for anyone to examine and critique.
· Encouraging STEM training, both of the current workforce and the next generation, to increase public comprehension of AI technology and simultaneously maintain competitiveness.
· Create constant citizen engagement in AI science through public workshops, citizen science, open access and open standards.

Artificial Intelligence can and will change the future of human society. With the experimental nature of any new technology comes the importance of making sure ethical, policy, and privacy concerns are carefully considered. Ethics cannot simply be "tacked on" at the end. We call on OSTP to establish the US as the world leader in AI research by creating and establishing AI ethics and policy principles for all.

Respondent 109

Emma Peck, Engine

Engine appreciates this opportunity to provide input on the promise of Artificial Intelligence (AI) technologies and the challenges and benefits that they present.

Topic One: The legal and governance implications of AI

In Engine's work with startups at the forefront of innovation, we regularly consider how policymakers should interact with emerging technologies that present new opportunities and challenges. Since technological innovation typically moves at an exponentially quicker pace than policymaking, government officials often find themselves reacting to new innovations without a full understanding of the technologies at issue or the consequences of their proposed laws or regulations. As such, we appreciate the White House's efforts to learn more about this emergent technology and believe it is important that the government continue its engagement with stakeholders on these important issues.

Academics and researchers have been exploring AI for decades, but thanks to recent advances in computing power, internet connectivity, cloud computing, and access to data, what was once just a theoretical endeavor has edged closer to reality. While the full

potential of AI has not yet been realized, certain AI technologies are already a part of our everyday lives, delivering benefits to consumers and businesses through applications and techniques like smartphone speech recognition, form auto-completion, e-commerce recommendations, spam filtering, and facial recognition. AI is also powering emerging industries like robotics, unmanned aircraft systems, and self-driving cars.

While some of these advancements have been led by larger technology companies, many of the breakthroughs in AI are being driven by startups. As nimble businesses with the ability to act quickly and focus on riskier ideas, startups are well-positioned to lead in some of the most innovative applications of AI. Additionally, investor interest in AI ventures is at an all-time high, making it easier for startups to more effectively compete with larger players in the AI space. According to market research firm CB Insights, funding for AI startups increased by more than 400% between 2011 and 2015. The firm estimates that investments will continue to increase, growing 76 percent to $1.2 billion this year. Additionally, over the past 5 years, larger corporations have acquired more than 30 startups working on AI technologies.

AI has potentially limitless applications across numerous sectors. Below are examples of startups that are already harnessing the power of AI to disrupt and transform existing industries:

Healthcare: The startup Enlitic is using deep learning and image analysis to help doctors spot abnormalities in medical images like x-rays and CT scans to make faster, more accurate medical diagnoses. Another startup, Sensely, has used AI to create a virtual nurse that helps clinicians manage their chronic care patients between appointments.

Transportation: The startup comma.ai is developing an autonomous car kit that will allow drivers to transform any non-autonomous vehicle into a self-driving car. Civil Maps is working on 3-D maps, powered by AI, to help autonomous vehicles navigate roads more easily and safely.

Energy: Verdigris installs sensors in commercial facilities to learn how energy is used and applies machine learning and AI to optimize energy consumption and operational efficiency.

Cybersecurity: Red Owl is using AI and machine learning to help businesses detect insider threats and Cylance is applying the same technologies to predict and combat advanced cyber threats.

Financial Services: The startup Neurensic is using AI and machine learning to help identify and prevent financial fraud and market manipulation.

Education: Volley Labs has created a software that layers machine learning over materials like textbooks and homework assignments to pull out key details, identify additional

resources for learning, and even create quizzes or study guides. Cognii is using an AI technology known as natural language processing to help evaluate and grade essays.

These are just a few examples of how startups are driving innovation across industries. They not only illustrate the positive impact that AI is already having, but also foreshadow an exciting future. The applications of AI with the most promise have likely not even been conceived of yet. As research and development of AI technologies advance, innovators will build on this progress to create incredible new products and services that provide value to consumers and improve lives.

Policymakers should keep this tremendous potential in mind when approaching AI and establish a legal framework that encourages innovation and growth. America led the world in the personal computing and Internet revolutions. The policies pursued today will directly impact the future ability of the U.S. to remain a global leader in the emerging field of AI. Policymakers should keep the following points in mind as they consider the legal and governance implications of AI.

1. AI technologies are diverse.
At the highest level, policymakers should recognize the diversity of emerging AI applications and avoid uniform, one-size-fits-all rules that do not lend themselves to the complexity of AI. For example, regulations around autonomous vehicles should look markedly different than rules governing the application of AI in healthcare. The term AI itself is incredibly broad, covering everything from content recommendation engines to the hypothetical sentient beings that dominate science fiction and the popular imagination. Speculative concerns about the latter should not unduly deter progress in the former.

2. Policies should encourage growth, not hamper progress.
Policymakers should also weigh the costs and benefits of potential rules, avoiding overly burdensome regulations and reactionary policies that inhibit the growth of AI. Just this week, the U.S. Department of Transportation hinted that it might pursue new rules that would require pre-approval of autonomous vehicle technologies before they reach the road. While the government has an important role in ensuring quality and safety, certain policies have the potential to drastically slow the development and adoption of transformative technologies. Since many AI applications depend on machine learning algorithms to improve their functioning, restricting their deployment may actually slow down the development of safety protocols. The direct and indirect impacts of any policy should be carefully considered before acting.

3. There is an existing body of law that already governs AI.
Among the more imminent issues surrounding the development of AI technologies are questions about how data is captured, used, treated, and protected. AI technologies enable and sometimes require the collection and analysis of massive amounts of data. This understandably raises questions around privacy and security. While appropriate safeguards

are essential, we believe that laws and regulations already exist that can adequately govern emerging AI technologies. Concerns around privacy and security are related to the data inputs of AI systems, not the AI technologies themselves. The amount of data processed by AI systems may be greater than what is needed for most technologies, but AI systems do not implicate privacy or security threats of a sufficiently different kind than existing non-AI systems to warrant creating a separate regulatory structure. A body of constitutional, federal, state, and common law, as well as numerous principles and industry best practices have developed over the years that protect individual privacy and data security while supporting an environment where innovation can flourish. AI is well covered by these existing laws, regulations, and industry best practices. Attempting to create new rules tailored to AI will only delay the growth of AI technologies with no real public benefit.

4. Openness and collaboration will foster growth.
Finally, the government should support policies that promote openness and collaboration in AI. Not only does an open approach accelerate the evolution of AI and foster breakthroughs, it may also protect against some of the potential "threats" of AI down the road. As the experience of the open source software development has shown, greater participation from a wide range of individuals can greatly mitigate security risks and generate unanticipated use cases. Many companies have already taken a transparent and inclusive approach to AI on their own. For example, Google has open sourced its machine learning platform TensorFlow to make its tools broadly available. According to Greg Corrado, Senior Research Scientist on Google's Machine Learning Team, the company chose to do this because "it's valuable for the community overall to establish standards in this space. Machine learning will be a new fundamental technology, so the sooner the engineering community agrees on standards for how we build these kind of systems, the better it is for everyone." Facebook has similarly open sourced its AI hardware design and deep-learning modules. Some leaders in AI have gone even further, creating OpenAI, a non-profit with a mission of advancing AI research and making their findings accessible to anyone. These steps are significant for startups, giving them a base to build upon and fostering continued growth. Promoting open systems—including making certain government AI endeavors open-source—will greatly increase the pace of development in the sector.

AI technologies are quickly evolving and policymakers should be nimble in their approach. American startups have an opportunity to lead in the AI revolution as they have in all of the other major technological breakthroughs of the past decades. A similar light-touch regulatory approach and a commitment to collaboration and cooperation between the government and stakeholders will ensure the U.S. remains at the forefront of innovation.

Topic Four: The social and economic implications of AI

According to the firm Venture Scanner, there are currently 499 AI companies in the U.S., 415 of which are startups, and $4.2 billion in VC funding for these companies. The same firm estimates that about 55 percent of these companies have 1-10 employees, 35 percent have

11-50 employees, and 10 percent have over 50 employees. Tens of thousands more work at startups that incorporate and/or leverage AI or machine learning but don't necessarily have it as a core business product. Going forward, as startups working on AI technologies grow in number and size, they will add thousands of jobs to our economy.

Respondent 110

Mark MacCarthy, Software & Information Industry Association (SIIA)

SIIA Comments on Artificial Intelligence
July 22, 2016

Artificial intelligence is a generic name for computational techniques that provide machines with cognitive capacities. Machine learning is a subset of AI that trains programs from examples and precedents. The current success of these techniques speech and object recognition is a natural outgrowth of developments in computer technology, specifically, the arrival of massive amounts of information, vast increases in computing power, and breakthroughs in analytical techniques. Trillions of bits of sound, image and text can be processed in high-power computers to train software to identify faces, objects and words. We are just at the beginning of the application of these techniques in all domains of economic, political and social life, creating enormous opportunities and challenges. SIIA congratulates the Administration for focusing the attention of the policymaking community on these vital developments.

After noting that these new computational techniques are poised to improve the lives of millions of consumers and workers, these comments make three points. First, AI's effect on the labor market will be similar to that of earlier productivity-enhancing technologies; policy should mitigate any possible adverse effects on the nature and availability of work through effective worker training programs. Second, policymakers should be clear that existing discrimination laws apply to AI computational techniques; separate non-discrimination rules for AI are not needed. Third, any constraints on differential pricing should not be introduced as restrictions just on the use of AI technology. Our overall themes are that the policy issues raised by AI are not new and they are not insurmountable. Successfully managing them requires that governments focus on outcomes instead of underlying technologies like AI.

Benefits of AI

AI's benefits were summarized comprehensively in the OSTP workshops. Industry is actively discussing and working on ways to ensure the benefits of AI are possible for everyone. AI research is vibrant and developments are the result of open, international collaboration.

Here just one example of how AI can literally save lives. Medical researchers used pattern recognition to analyze data generated from premature babies such heart rate, respiration rate, temperature, blood pressure, and blood oxygen level – with startling results. The simultaneous stabilization of vital signs as much as 24 hours in advance was a warning of an infection to come, thereby allowing medical intervention well before a crisis had developed.

AI had discovered a useful fact about the onset of fevers and infections in premature babies that can be the basis for early intervention.

Future of Work

Automation has historically produced long-term growth and full employment. But the next generation of really smart AI-based machines could create sustained technological unemployment. Two Oxford economists estimated that 47 percent of occupations are susceptible to automation. An OECD study found that "9% of jobs are automatable."

The Council of Economic Advisors (CEA) recently warned that AI could exacerbate wage inequality, estimating that 83 percent of jobs making less than $20 per hour would come under pressure from automation, as compared to only 4 percent of jobs making above $40 per hour. The CEA also documented a long-term decline in prime-age male labor force participation – from 97% in 1954 to 88% today – that could be exacerbated by AI.

Despite these concerns, there is no real evidence that the ultimate impact of AI on the labor market will be any different from that of earlier productivity-enhancing technologies. Studies have shown that labor market developments that some are blaming on computer technology and the Internet – like job polarization – have been a feature of the US economy since the 1950s.

Moreover, information technology creates jobs. An SIIA study showed that the software industry employs more than 2.5 million workers, and supports another 1.1 million. It demonstrated that industries investing most heavily in software from 1997 to 2012 had relatively strong rates of job growth, while industries investing the least in software experienced both high levels and low levels of job growth.

Software also enables insourcing of manufacturing jobs. Advances in software and artificial intelligence make new state of the art production facilities in the U.S. cost competitive with overseas facilities. This return of production facilities to the U.S. is creating substantial numbers of good high paying jobs for skilled U.S. workers.

As the recent CEA report recognizes, the biggest worry about AI is that there might not be enough of it. We need more AI, not less, in order to jump start labor productivity, which has lagged over the past decade. Public policy should encourage research and development in AI and create a favorable climate for the successful deployment of AI.

Public policy can also respond to any tendency AI might have to reduce employment by providing increased funding and effective administration of education and training programs for 21st century workforce skills such as streamlining and modernizing the Carl D. Perkins Career and Technical Education Act. As Alec Ross noted recently, to reap benefits from the robot revolution"…we have to look at what the industries of the future are and radically reorient how we deliver vocational education."

Public policy can also support the use of these new skills in the workplace. The OECD found that "the extent to which workers use their information processing skills at work is a major determinant of productivity, wages and job satisfaction." Public policy could encourage workplace practices and labor market institutions such as "collective bargaining and minimum wages" that increase the use of information processing skills at work.

Discrimination

OSTP's AI workshops revealed concern that AI-driven decision making could perpetuate

and aggravate discrimination against disadvantaged groups. As a White House report noted earlier this year, however, this concern about statistical bias is not new. Policymakers have long known that statistical techniques used to make eligibility decisions could have discriminatory effects, and have a well-developed methodology for assessing this.

SIIA pointed out in comments to the FTC that effective statutory constraints on discrimination already apply in regulated eligibility contexts such as lending, insurance, housing and employment. When discrimination arises indirectly through the use of statistical techniques, regulatory agencies and courts use disparate impact assessment to determine whether the practice is prohibited discrimination. For instance, Title VII of the Civil Rights Act of 1964 forbids any employment practice that causes a disparate impact on a prohibited basis if the practice is not "job related for the position in question and consistent with business necessity" or if there exists an "alternative employment practice" that could meet the employer or employment agency's needs without causing the disparate impact.

A 2007 study of credit insurance scores by the FTC illustrates how a statistical technique can be assessed for disparate impact. It found that credit insurance scores predict automobile insurance risk; protected classes had lower scores and so paid higher insurance premiums; the predictive power of the score did not derive from correlation with a protected class; no alternative model had equivalent predictive power but less adverse impact on protected classes.

The study seems to indicate that the scores would pass a disparate impact test. There was a disproportionate adverse impact on a protected class. But the score was not just a proxy for race, and it satisfied the legitimate business need of controlling auto insurance risk. And no alternative model satisfied the business need with less impact on the protected class.

Neither the FTC nor any other regulatory agency at the Federal level took action against the use of credit insurance scores on the basis of this study. In fact, it was viewed as confirmation that the scores were not simply proxies for protected classes and were not discriminatory.

Many people, however, think the use of the score is unfair. Even if it is accurate and passes a disparate impact test, there is a point of view that says something is wrong with using it because it further disadvantages already disadvantaged groups. In this view it would be better to sacrifice some accuracy in order to improve the position of the already disadvantaged. This feeling is probably behind the ban on their use in some states.

This is an old and unresolved argument about whether the non-discrimination statutes should aim at reducing the subordination of disadvantaged groups or at reducing the arbitrary misclassification of individuals.

This unresolved argument lingers in the discussion of AI. AI is not exempt from the non-discrimination laws. When AI techniques are used in the regulated contexts of housing, credit granting, employment and insurance, they are subject to the same regulatory controls and validation requirements that apply to any statistical methodology used in these contexts. If they have disproportionate adverse effects on a protected class, they are prohibited unless they can pass a disparate impact test.

The Administration should continue to remind the public that the use of statistical

techniques, including AI, to engage in unfair and discriminatory practices is prohibited under existing laws. As recommended in the Administration's 2014 Big Data Report, regulatory agencies "should expand their technical expertise to be able to identify practices and outcomes facilitated by big data analytics that have a discriminatory impact on protected classes, and develop a plan for investigating and resolving violations of law in such cases."

In addition to these legal requirements, there have been calls for fairness by design and audits of AI techniques for bias. Our view is these calls require further conversation. Those who design, implement and use AI systems must be thoughtful about the potential for discriminatory effects. In many cases, businesses would want to know whether their use of AI has disproportionate adverse impacts on protected classes. But a universal audit requirement for all statistical models including AI is too broad. So there needs to be a discussion among interested parties about when and how audits might be employed.

There also has to be some discussion about what to do with findings of disproportionate adverse impacts. Current non-discrimination law applies only to certain industries and contexts, and even in those contexts, it does not require designing away features of AI algorithms that, like credit insurance scores, pass a disparate impact test. Because a significant spectrum of opinion holds that these features are still unfair, some companies might feel a social responsibility to go beyond legal requirements and to design and use AI techniques that have been freed as much as possible from harmful biases. A conversation among government, industry and advocates is needed to clarify the situations in which this makes sense.

In addition, AI techniques can be used to fight bias. Together with more traditional data analysis they can be used to help employers diversify their workforce, to assess compliance with fair lending laws and in other ways to detect and remedy discrimination. The Administration should seek ways to support and encourage these uses.

Differential Pricing

The OSTP workshops reveal concern about the use of AI for price discrimination. AI could be used to personalize pricing just as it is used to personalize advertising, medicine, book and music recommendations, and education. Pricing based upon a person's characteristics rather than on the cost of the good is widely used in publishing, film entertainment, software. It is familiar to consumers as senior citizen or student discounts. But survey evidence reveals that differential pricing practices are often unpopular; many people think they are unfair. Any increase in personalized pricing will create a challenge for policymakers.

CEA released a useful discussion of differential pricing last year, pointing out the economic arguments in its favor. When companies can price to market, they can make their product or service available to people who would otherwise not be able to afford it. Moreover, because those willing to pay more are usually people who have greater incomes, differential pricing has a progressive effect in countering economic inequality.

The debate between advocates and opponents of differential pricing is a legitimate and important one. But the debate should focus on the normative issues, not the technology. It is not about AI, but about the pricing practice. Any public policy response should be about the

practice and not the underlying technology.

Click link for pdf comments that include hyperlinks:
http://www.siia.net/Portals/0/pdf/Policy/SIIA%20OSTP%20AI%20Comments.pdf

Respondent 111

Ryan Hagemann, Niskanen Center

(URL for formatted comments: https://niskanencenter.org/wp-content/uploads/2016/07/CommentsArtificialIntelligencePolicyOSTP.pdf)

Public Interest Comment

Submitted to the Office of Science and Technology Policy in the Matter of:

A Request for Information on Artificial Intelligence

Ryan Hagemann
Technology and Civil Liberties Policy Analyst
The Niskanen Center

Submitted: July 22, 2016
Docket No. 2016-15082
Word count: 1862

Executive Summary

The long-awaited promise of artificial intelligence (AI) is beginning to materialize. Powerful AIs, such as IBM's Watson and Google's Deepmind, which has bested the world's Go champion, herald the "springtime" of AI research and development. However, some find the flowering of the technology alarming, and wonder aloud whether AI may lead to a Terminator-style future in which incomprehensibly intelligent computers destroy human civilization. Even moderate critics of AI warn that we now stand on the verge of a mass labor dislocation in which up to half of all jobs may be taken by machines. For now, however, these worries are extremely speculative, and the alarm they cause can be counterproductive.

In order to maximize the benefits associated with ongoing developments in AI, we recommend that policymakers and regulators:

(1) avoid speaking of hyperbolic hypothetical doomsday scenarios, and

(2) embrace a policy of regulatory restraint, intervening in the development and use of AI technology only when and if the prospect of harm becomes realistic enough to merit government intervention.

Introduction

As the renowned science fiction author Isaac Asimov once wrote, "Any sufficiently advanced technology is indistinguishable from magic." For many people, the seemingly magical nature of unfamiliar technology invites wild speculation about the human implications of its development and adoption. Nowhere is this more true than in artificial intelligence (AI). In the interest of brevity, these comments will address just one of the topics raised in the Office of Science and Technology Policy's request for information: (4) the social and economic implications of AI.

Social Implications

AI is unlikely to herald the end times. It is not clear at this point whether a runaway malevolent AI, for example, is a real-world possibility. In the absence of any quantifiable risk along these lines government officials should refrain from framing discussions of AI in alarming terms that suggest that there is a known, rather than entirely speculative, risk. Fanciful doomsday scenarios belong in science fiction novels and high-school debate clubs, not in serious policy discussions about an existing, mundane, and beneficial technology. Ours is already "a world filled with narrowly-tailored artificial intelligence that no one recognizes. As the computer scientist John McCarthy once said: 'As soon as it works, no one calls it AI anymore.'"

The beneficial consequences of advanced AI are on the horizon and potentially profound. A sampling of these possible benefits include: improved diagnostics and screening for autism; disease prevention through genomic pattern recognition; bridging the genotype-phenotype divide in genetics, allowing scientists to glean a clearer picture of the relationship between genetics and disease, which could introduce a wave of more effective personalized medical care; the development of new ways for the sight- and hearing-impaired to experience sight and sound. To be sure, many of these developments raise certain practical, safety, and ethical concerns. But there are already serious efforts underway by the private ventures developing these AI applications to anticipate and responsibly address these, as well as more speculative, concerns.

Consider OpenAI, "a non-profit artificial intelligence research company." OpenAI's goal "is to advance digital intelligence in the way that is most likely to benefit humanity as a whole,

unconstrained by a need to generate financial return." AI researchers are already thinking deeply and carefully about AI decision-making mechanisms in technologies like driverless cars, despite the fact that many of the most serious concerns about how autonomous AI agents make value-based choices are likely many decades out. Efforts like these showcase how the private sector and leading technology entrepreneurs are ahead of the curve when it comes to thinking about some of the more serious implications of developing true artificial general intelligence (AGI) and artificial superintelligence (ASI). It is important to note, however, that true AGI or ASI are unlikely to materialize in the near-term, and the mere possibility of their development should not blind policymakers to the many ways in which artificial narrow intelligence (ANI) has already improved the lives of countless individuals the world over. Virtual personal assistants, such as Siri and Cortana, or advanced search algorithms, such as Google's search engine, are good examples of already useful applications of narrow AI.

Economic Implications

The extent to which AI's may "disrupt" labor markets is difficult to measure. It is clear that as AI becomes more advanced, it will result in the increased automation of work. This trend may or may not result in mass job dislocation. However, some low-skilled jobs are clearly vulnerable to automation and improvements in AI technologies will certainly result in the loss of some of these jobs. It's important to recognize that AI is like many, many other technological developments that have led to the replacement of labor by machines. What's new is the kinds of jobs AI will allow to be automated. The negative impact for certain workers in certain fields should not blind us to the likely benefits of increased productivity in terms of economic performance and job-creation elsewhere in the economy. Government policies that both promote economic growth and help dislocated workers with unemployment insurance, retraining, and other forms of public assistance can facilitate disruptive innovation while protecting the welfare of those most likely to lose jobs to AI technology. If policymakers get these policies right, advanced AI and increasing automation will help bring about rising, broad-based prosperity.

Policies to ameliorate negative consequences of increased automation in the economy must be informed by empirical research. Some researchers have suggested that traditional measurements, such as gross domestic product per capita, may not accurately capture the true scope of the costs and benefits of AI. As such, further research assessing more appropriate metrics for quantifying the effects of AI and related automation will be needed in order to clarify policymakers' options for dealing with the negative implications of continued advances in the technology.

The Future of Life Institute has observed that "our civilization will flourish as long as we win the race between the growing power of technology and the wisdom with which we manage it. In the case of AI technology … the best way to win that race is not to impede the former, but to accelerate the latter, by supporting AI safety research." Government can play

a positive and productive role in ensuring the best economic outcomes from developments in AI by promoting consumer education initiatives. By working with private sector developers, academics, and nonprofit policy specialists government agencies can remain constructively engaged in the AI dialogue, while not endangering ongoing developments in this technology.

General Policy Recommendations

Recommendation #1 (social): Because doomsday scenarios overstate the known risks of AI, official discussion of AI policy should be conducted in measured and moderate terms, and focus on actual or predictable risks of existing or emerging technology rather than on unfettered speculation about unknowable future developments.

Many of the worst-case scenarios associated with AI are fueled by hyperbolic references to the potential for a Terminator-style apocalypse. But existing AI is in an early, almost childlike stage of development. Current AI technology doesn't remotely approach the level of sophistication that would merit the level of concern some critics have encouraged. It is encouraging that the White House recognized this in its original announcement of an interagency working group, which formed the foundation for the Office of Science and Technology Policy's call for comments. Ed Felten, the Deputy U.S. Chief Technology Officer, pointed out that current "AI is confined to narrow, specific tasks, and isn't anything like the general, adaptable intelligence that humans exhibit." If left unchallenged, exaggerated worries that implicitly misrepresent the nature of current and near-term AI capabilities could impede development in this nascent field of science and technology.

In a recent report on AI, Robert Atkinson, the president of the Information Technology and Innovation Foundation, put it best:

Making sure that societies receive the full economic and social benefits that AI has to offer first and foremost requires accelerating, rather than restricting the technology's development and adoption. And that in turn requires that policymakers resist an AI techno-panic; they must instead embrace future possibilities with optimism and hope.

Avoiding apocalyptic rhetoric will help ensure a reasonable and practical policy discussion about AI. Policymakers and regulators who give in to the temptation to dabble publicly in speculation about cinematic worst-case scenarios invite reckless and counterproductive regulation unmoored from realistic cost-benefit analyses. Government agencies and lawmakers would do well to avoid discussing AI in hyperbolic terms, lest we delay or altogether lose out on the many great benefits AI can offer.

Recommendation #2 (economic): AI's full potential can only be actualized if government embraces a policy of regulatory restraint.

Private stakeholders are well-poised to explore and manage the costs and risks associated with ongoing developments in AI. The government can be an effective partner and collaborator, but regulators should stand down for the time being. Regulating too early, or on the basis of knee-jerk reactions to merely hypothetical doomsday scenarios, will hinder technological progress and innovation. Restraint and realism are especially important to encourage ongoing private capital investments in AI research and development.

The general regulatory framework the Niskanen Center recommended in response to the National Telecommunications Information Administration's request for comments on the Internet of Things (IoT) should also be applied to the field of AI. Indeed, the IoT is a nexus of developments in AI, big data collection and analysis, and robotics and automation. Because all these technologies are interrelated, a lack of regulatory forbearance in one area will have negative consequences that reverberate through the entire emerging technology ecosystem.

Conclusion

As AI research and development continues, regulators and policymakers must remain realistic about the nature and size of potential costs and weigh them responsibly against actual and probable benefits. Speaking of developments in AI in apocalyptic and eschatological terms distracts from the real and important issues facing this nascent field. The benefits of narrowly-tailored AI can already be seen all around us, and much greater benefits are on the horizon. Meanwhile, it is not presently clear whether the AI technology that might lead to a doomsday scenario is even possible. For now, these scenarios should be approached with an air of dismissive skepticism. It is often possible to imagine catastrophic consequences of new technologies. But it's neither rational nor responsible to take nightmares about costs into account alongside real benefits that have already begun to accrue.

As AI develops, government can be a valuable ally in promoting engagement between researchers and academics, the private sector, government agencies, and civil society organizations. However, this engagement should avoid conjuring the specter of legislation or regulatory action that may hinder the important work being done in this field. Unless a clear need for intervention can be established with a cost-benefit analysis that balances real harms against real benefits, regulators and policymakers should remain on the sidelines and keep watch over ongoing developments.

We thank the Office of Science and Technology Policy for the opportunity to provide these comments.

Respondent 112

David Verhey, Argent PLLC, Washington DC

We serve as counsel to several public and private entities involved in the field of artificial intelligence. We appreciate this opportunity to submit comments in response to the Administration's Request for Information (RFI) and would like to highlight two primary concerns---both of which fall under Heading No. 1 (Legal and Governance Implications of Artificial Intelligence).

They are:

a. Artificial Intelligence, Healthcare, and Big Data

Artificial Intelligence (AI) has a number of promising applications to healthcare and medicine, including advanced diagnostics and patient care. IBM's Watson program is an example of how this technology can directly support medical service and even improve physician recommendations in some instances. But even while AI promises major benefits to medicine and the population, there are significant technological and legal disagreements as to whether data managers and owners can sufficiently protect individual privacy through data anonymization or "de-identification" protocols. If those questions are difficult to resolve now, they will become even more acute with the ongoing expansion of big data stores maintained on commercial cloud-based servers operated by third parties---some of which may not maintain anonymization standards. In view of those concerns, we recommend that the Administration (1) identify these risks as a priority for policy development and industry best practices and (2) encourage medical professionals, technology experts, and attorneys to jointly develop a framework to protect individual privacy (including HIPAA information) in the world of AI-enabled big data.

c. The Internet of Things, AI, and Security

Gartner estimates that the Internet of Things (IoT) will grow from 4.9 billion devices in 2015 to 25 billion devices by 2020. Based on the sheer volume of data that will be generated by this network, many experts believe that the only effective way to keep up with the information flow is by using AI analytical capabilities. But that's not the only thing we need from AI. Even if it can help make sense of IoT data, the growth of the sensor network will produce an ever-increasing "attack surface" (and risk of data loss) that may not be readily defended with conventional firewalls and tools. To deal with this emerging risk, we need to promote AI solutions that can both analyze sensor data and protect networks from intrusions and attacks. In addition, Government and Industry should look for ways to jointly develop cybersecurity regulations that are tailored to the specific function of a device, the industry or business it serves, the risks it poses from hacking or data loss, and the nature of the device's data collection (including who gets to see it). See, e.g., the President's National Security Telecommunications Advisory Committee, Report to the President on the Internet of Things (November 2014), p. ES-4 (recommending standing Government and industry body to develop and maintain cybersecurity guidelines). Taking that approach will help maintain industry innovation and promote a solid economic and

technical rationale for each standard. And that will help promote more development in the IoT and the enabling capacities of AI.

Respondent 113

James Hairston, Facebook

I. AI at Facebook

Facebook's mission is to give people the power to share and make the world more open and connected. Artificial intelligence (AI) helps us build tools that allow people to connect and share in new ways. We are using AI to build a new generation of apps and services that are more natural, intuitive, valuable, and more responsive than anything that has come before. As with all our technologies, we're committed to the responsible development and use of AI.

The Facebook Artificial Intelligence Research (FAIR) team was formed in 2013, while our Applied Machine Learning team was formed in 2015. Both groups do exciting work on machine learning, natural language processing, and computer vision. The next section describes some applications of these technologies at Facebook in more detail.

Facebook is working openly with and investing in the AI research community as we strive to make meaningful progress in the field and share it with the world. We do research in the open, which includes publishing all of our papers and developing open source code to drive global development and opportunity in this promising field.

We believe this open model spurs innovation, encourages collaboration and mutual review, and helps us all move faster. We're excited about the possibilities for AI to advance the progress of science and improve the world.

II. Facebook Products and Programs Powered by AI

Machine Learning involves teaching computers to learn how to perform certain tasks. Machine learning helps deliver a range of services on our site from instant translation of text in different languages, to providing more relevant content in News Feed. Our ongoing Connectivity Labs efforts to connect underserved communities to the Internet have also been informed by enhanced maps that incorporate AI-derived analysis of population distribution.

Natural language processing refers to the ability of machines to read and understand human language. Natural language processing is the underlying technology framework for Facebook's digital assistant 'M' which has automated certain features in Messenger and will use AI to fulfill requests.

Computer vision is a subfield of machine learning that involves teaching computers to understand visual content, including images and videos. The combination of our work on computer vision and natural language processing makes our site available to the blind and visually impaired through a technology called automatic visual content captioning, or “alt text.” Our computer vision technology uses machine learning algorithms to make predictions about the objects and emotions included in a photo—e.g. “car,” “baby,” “smiling,” or “sushi.” This description is then read aloud using text-to-speech technology, allowing blind and visually impaired individuals to “see” and interact with photos shared by their friends and family in a way not previously possible.

III. AI's Uses and Potential

People are beginning to reap the benefits of AI — from healthcare and astronomy to the tasks we do every day. Machine learning is helping us map new objects in space and detect diseases with new accuracy that will save lives. AI-powered tools like digital assistants and instant language translation are engendering more commerce and communication, making people more productive in the process.

Facebook is leading much of this work organically, as we have realized that much of the data we have access to can be used to solve major global challenges. Much of our research is focused specifically on projects geared to unlock the power of data — with the support of AI computing technology — to inform and accelerate change across the globe. This often helps us identify major social challenges, build relationships with the organizations working on these issues, and utilize our human and technical resources to great effect. Examples where Facebook data, AI technology, and innovative thinking have come together to make progress on social challenges include our work on accessibility, global connectivity research, and infrastructure mapping.

Other inspiring ways AI is already being put to use include:

* detecting cancers/melanomas in images
* deep learning to map Mars and classify galaxies
* protecting consumers against fraud
* powering self-driving cars & smart highways
* instant language translation

In addition to advances in science and technology, artificial intelligence will drive new economic opportunity in the coming years. We commissioned a study with Analysis Group on the AI and the global economy. Their study concluded that the use, development, and adoption of artificial intelligence will generate a global economic impact of between $1.49 trillion and $2.95 trillion over the next 10 years. The economic benefits of AI come not only—or even primarily—from direct growth in sectors that develop AI technologies, but from increased productivity and spill-over benefits to other, existing sectors of the

economy.

IV. Principles to Guide AI's Growth and Development

The U.S. government should maintain a light-touch regulatory approach focused on consumers and outcomes over underlying technologies. Any approach should consider AI's benefits to consumers and be informed by regular consultation with industry and experts—balanced with adequate consideration of security and privacy policy questions as they arise. In addition, AI's growth and development can also be enhanced by policies to:

* grow the pool of STEM graduates and high-skilled workers;
* promote competitive markets and experimentation with new technologies;
* maintain low barriers to entry for small and innovative firms; and
* create R&D incentives for firms building products and services with emerging technologies like AI.

Respondent 114

Michael Beckerman, The Internet Association

Request For Information: Preparing For The Future Of Artificial Intelligence

The Internet Association submits these comments in response to the White House Office of Science and Technology Policy (OSTP) request for information regarding the policy implications of Artificial Intelligence (AI).

The Internet Association represents almost 40 of the world's leading Internet companies. Our mission is to foster innovation, promote economic growth, and empower people through the free and open internet. As the voice of the world's leading Internet companies, our job is to ensure that all stakeholders understand the benefits the internet brings to our economy. Several Internet Association member companies have made significant investments in AI technology. For example, Amazon's Science team, eBay's Expertmaker structured data team, Facebook's AI Research Lab, and Google's cross-company AI research team all are leading advancements in the AI field. And beyond these specific examples, it is clear that overall private sector investment in AI technology has increased significantly in past years, from $1.7 billion in 2010 to $14.9 billion in 2014, making the OSTP's request for information an important and timely one.

In its request for information, OSTP asks a range of questions related to AI, some of which highlight issues beyond the IA's mission. However, insofar as the questions relate to our expertise, we request that OSTP and the administration in general adhere to the following three principles as they consider AI policy going forward:

• First, although AI raises interesting public policy questions, they are not necessarily uncharted territory. In fact, U.S. policymakers have immersed themselves in similarly complex debates before and have developed significant institutional expertise and skills that are transferable to the AI space.
• Second, the U.S. government can and does play an important role in fostering AI technology both at home and abroad. This role ranges between practical policies such as support for STEM education at home and engaging in sophisticated economic diplomacy abroad.
• Third, while AI deployment and use is not without risk, this risk is manageable and there are demonstrable economic and non-economic benefits associated with AI. Thoughtful public policy demands a careful weighing of these benefits against perceived risks so that the benefits can be fully realized.

1. Existing Policy Frameworks Can Adapt to Artificial Intelligence

Before delving into policy specifics, the IA submits that it is advisable for policymakers to draw parameters around what artificial intelligence is and is not since this is an open debate that may trigger fearful policy responses where they are not needed.

For Internet Association members, AI refers to the engineering discipline that aims to create intelligent machines that work and react like humans. Differently stated, AI is computational systems and devices made to act in a manner that can be deemed intelligent. Machine learning refers to an aspect of artificial intelligence that focuses on making predictions from a set of examples. Related to this, robotics are autonomous mechanical systems that sometimes incorporate techniques of artificial intelligence or machine learning.

Artificial intelligence is not a science fiction technology. AI has been around for several decades and has developed at a steady but relatively slow pace compared to other technologies from which it can be benchmarked, including broadband internet and mobile telephony. To illustrate this point, in 1966 the Register of Copyrights identified computer authorship as one of the three "major problems" facing the Copyright Office. In fact, the registrar flagged this as a "crucial question" in his annual report that year and yet it remains an open question to this day. This 50-year old anecdote suggests that knee jerk policy reactions to AI are neither needed nor advisable.

Against this backdrop, the Internet Association submits that although AI may raise some interesting public policy questions, they are not necessarily new and existing policy frameworks can adapt to it in an orderly and timely way. Furthermore, U.S. policymakers have a proven track record in this regard. A leading example of this flexibility in practice is the so-called 'common law of privacy' developed by the Federal Trade Commission over the past twenty years. The FTC has used its framework common law statute to develop case law and policy guidance to industry in the areas of privacy and data security. As new

technologies have emerged, the FTC's common law approach has been applied to them in a relatively seamless way. These diverse technologies include mobile payments, the Internet of Things, and RFID. There is little reason to think why the same framework could not also successfully be adapted to AI.

2. Government's Role in Fostering Artificial Intelligence

The U.S. government can and will play an important role in fostering AI technology both at home and abroad. This role ranges between practical policies such as support for STEM education at home and engaging in sophisticated economic diplomacy abroad.

The Internet Association fully acknowledges that government plays an important role in research and development of new technologies. After all, the internet itself would not exist had DARPA and the NSF not invested in their early stage research decades ago. Similarly, the government can also play a pivotal role in AI research and development.

A key variable in this research function, for both the public and private sectors, will be ensuring that personnel engaging in it are diverse and come from a variety of socio-economic backgrounds. Unfortunately, these personnel do not currently exist due to a lack of investment in STEM education. Currently, fewer than one in five high school students has ever taken a computer science course—a figure that has fallen by 24 percent over the past two decades —and only 7 percent of high schools offer the Advanced Placement course in Computer Science. Government can play an important role in remedying this STEM education diversity gap by promoting expanded access to computer science education through programs such the Computer Science for All Initiative for K-12 students and the Tech Inclusion Initiative. These programs will create a diverse talent pipeline for all research, including in the AI field.

Within government itself, AI deployment can play a role in strengthening e-government, making government more efficient and responsive to citizens. The Internet Association supports efforts like the U.S. Digital Service that seek to build technical capacity within government across agencies, including increased deployment of machine learning where applicable.

Finally, government continues to play a significant role in promoting and protecting U.S. technology interests abroad. Like the internet, AI is the product of international collaborative research and it is important that this approach is followed as AI technologies leave the research lab and enter the global economy. Internet governance works best when international governance forums are multistakeholder in nature. The same logic should apply to AI governance. It is also important that the U.S. leverage its diplomatic network in support of pro-innovation legal regimes overseas in fields such as standard setting and copyright as they relate to AI. A prototype for this role already exists in the recently announced digital attaché program created at the International Trade Administration

within the Department of Commerce.

3. AI's Benefits and Risks Require Careful Balancing

While AI deployment and use is not without risk, there are demonstrable benefits – both economic and non-economic - associated with it. Thoughtful public policy in this space demands a careful weighing of these benefits against perceived risks so that the benefits can be fully realized.

The economic impact from AI deployment will be both direct GDP growth from industries developing and selling AI technologies and also indirect GDP growth as other industries adopt AI technologies and realize the productivity gains associated with it. In 2016, the Analysis Group estimates this economic impact to range between $1.49 trillion and $2.95 trillion globally over the next ten years. Even the conservative end of this range would suggest that AI's economic benefits are likely to be significant. Also included in AI's economic benefits is the consumer surplus that the technology creates as it lowers search and transaction costs for consumers through applications such as eBay's personalized shopping experience. As explained by eBay's CEO Devin Wenig in a recent blog, by using AI technology eBay is working "to help improve shipping and delivery times, trust, pricing, and more" for eBay's customers.

Beyond this direct economic impact headline, it is important also to pause to consider the beneficial uses that AI will be put to before weighing its risks. Beneficial potential applications for machine learning that automates analytical modeling include detecting molecular structures in vast amounts of biological data that is predictive of certain diseases and protecting against consumer fraud. Applications for machine vision already include providing object descriptions for the blind and car safety systems that detect pedestrians and cyclists. These applications create health and safety benefits for society overall.

As with these benefits, analysis of the risk associated with AI should be grounded in rigorous research that is supported by empiricism to the fullest extent possible. One risk already under the spotlight is the extent to which AI machine learning may produce suboptimal results when the data upon which it based its analysis is incomplete and therefore skewed against underrepresented demographics. The FTC flagged these concerns in the consumer arena when it counseled in its Big Data report to consider "whether data sets are missing information about certain populations, and take steps to address issues of underrepresentation and overrepresentation." Related to this, the agency counseled companies "to consider whether biases are being incorporated at both the collection and analytics stages of big data's life cycle."

In order for all demographics to realize the benefits of AI (and for possible regulatory scrutiny to be avoided) these data gaps will need to be addressed. The Gates Foundation recently announced an investment of $80 million to help foster increased data collection

for women in developing countries, a demographic that is sorely underrepresented. These data will be used to improve health and economic opportunity outcomes for girls and women since, according to Melinda Gates, "closing the gender gap means closing the data gap." Government could also play a role in this context by overseeing the release of robust and high quality datasets to responsible actors engaged in AI research.

Conclusion

The Internet Association thanks the White House OSTP for shining a timely spotlight on AI and its future in our society. Our members share your interest and look forward to continued dialog with OSTP as AI technology moves forward and realizes its long-promised potential.

Respectfully submitted,

Michael Beckerman
President & CEO
The Internet Association

(These comments are responsive to questions 1, 2, 4, 7, and 8 in the OSTP Request for Information).

Respondent 115

Larry Holder, Washington State University

Response to RFI: Preparing for the Future of Artificial Intelligence

From: AI/ML Research Group, School of EECS, Washington State University
Researchers: Diane Cook, Jana Doppa, Larry Holder, Shuiwang Ji, Matt Taylor

Selected topics addressed:

(2) Use of AI for public good

The world's population is aging and the resulting increase in chronic health conditions is a challenge our society must address. Technology is being developed at a rapid rate to decrease caregiver (and government) burden and cost while improving quality of life for these individuals. In particular, sensors embedded in everyday environments such as homes and offices can unobtrusively monitor an individual's wellbeing. Sensors embedded in smartphones and other mobile devices can in theory provide in-the-moment information on the user's health status as well as predict potential problems that may be encountered over

the coming minutes (e.g., an asthma attack), hours, or days.

Without AI, data collected from these sensors is just a sea of noisy, imprecise, voluminous numbers. Sensors can be unreliable and difficult to interpret. If a developer tried to make sense of sensor data themselves, they would quickly be overwhelmed and the resulting software would be quickly abandoned. Trying to find the magical combination of numbers that occurs when a person experiences a fall or a heart attack is like finding a needle in a haystack. Adding more "hay" in the form of additional data does not help. Instead, AI technologies can do the work of searching for the sensor states that indicate health status.

AI can be used for public good by automatically identifying, from sensor data, when a person is experiencing a health crisis or will experience a crisis in the near future. AI can also be used to design and automate interventions to circumvent the crisis and enhance quality of life. With the coming age wave and lack of resources to handle the coming health needs, AI is not only valuable for public good, it is necessary.

(3) Safety and control issues for AI

In many ways AI is like any other technology; it can be used for good or evil. One important difference is the potential for AI to make decisions about this use itself. A computational expression of the "golden rule" is necessary.

(6) Most important research gaps in AI that must be addressed to advance this field and benefit the public

Understanding biological intelligence. Specifically understanding the dynamics of spiking neurons, cortical circuits made up from them, and how these achieve memory, learning, and higher-level decision-making.
We need the ability to define computational models of value systems and to make AI adhere to them (i.e., not be able to circumvent some aspects of its programming).

(8) Specific steps that could be taken by the federal government, research institutes, universities, and philanthropies to encourage multi-disciplinary AI research

In terms of AI research in general, and given that humans are the best example of a functioning intelligence (albeit biological), computational neuroscience is an important multi-disciplinary area that has the potential to make important near-term breakthroughs in the understanding of human intelligence and the development of computational models to implement this intelligence in computers.
In addition to general AI research, there are several societal grand challenges that a targeted AI could solve, e.g., clean energy or curing cancer. Using AI to solve these challenges would require a multi-disciplinary team with expertise from the target domain and computational sciences.

One overarching challenge that AI could address is the enhanced living of the individual. Each of us has individual needs and desires, but often lack the means or knowledge to achieve these goals. A personal AI "coach" could bring to bear the collective knowledge of experts and other sources to maximize an individual's potential, even enhancing them beyond current levels of longevity, health and productivity.

Respondent 116

Lisa Hayes, Center for Technology & Democracy

The Center for Democracy and Technology ("CDT") is optimistic about the future of artificial intelligence ("AI"), and confident the technology will have widespread positive impacts. However, the rapidly developing technology will have significant effects on jobs, education, and policy, as well as ethical and regulatory implications for the federal government. It takes time for processes to change, standards to emerge, and people to learn new skills. In the case of AI, the government must act quickly to prepare for these changes, as the technology will diffuse rapidly.
CDT believes in the power of technology. A 501(c)(3) nonprofit organization, we work to preserve the user-controlled nature of the internet and champion freedom of expression. We support laws, corporate policies, and technology tools that protect the privacy of technology users, and advocate for stronger legal controls on government surveillance. We will address three topics in this Request for Information. Our primary focus is on how the U.S. government should ensure that technological advances are used to reduce inequality and promote progress for all segments of society.

(2) THE PUBLIC GOOD:

AI will be deployed to serve the public good, encourage civic duty, and collaborate and solve some of the world's most pressing, complex problems. However, historic bias in decision-making is not alleviated by automating the process and reducing human involvement. There is a risk that human bias might be built into the underlying architecture of these systems, from relatively simple analytics to sophisticated artificial intelligence. Creating positive outcomes for all requires humans to consider the ethical implications of the technology they are creating and using. For this vision to become reality, the government must work in collaboration with companies and civil society to deploy AI technology mindfully, and be vigilant about preventing disparities and harm.

* We must guard against algorithmic bias. We have seen that big data analytics risk eclipsing longstanding civil rights protections in how personal information is used in housing, credit, employment, health, education, and the marketplace. Machine learning algorithms are trained on large data sets, and when those sets are partial or contain implicit bias, the resulting algorithms can make incorrect inferences that lead to broader algorithmic biases and discrimination. With increasingly automated functions, these built-in discrepancies can multiply exponentially. There is a growing policy debate about how to build accountability into this system. A tremendous incentive exists for policymakers and

companies to innovate and lead the way in fair automation. Throughout this discourse, humans remain at the heart of automation - building, testing, refining, auditing, and evaluating these systems. We must both encourage the development of better diagnostic tools and ensure that AI creators are working with robust, high-quality data sets.

* Promote a diverse workforce. The development of effective AI mandates diversity on project teams in order to facilitate objective assessment and identify unconscious biases. It is essential that the field attract skilled human data analysts with diverse backgrounds. AI increasingly intrudes on people's personal information, using and manipulating highly granular data. Research demonstrates that the automated judgments behind personalization are not harmless, neither in effect nor in perception; and it is largely left to the data collectors to enforce moral standards. Automated sorting and deeming data "irrelevant" can result in material harm. Given the diversity of human insight and wisdom, much consideration should be given to how data is identified as important and useful, as such decisions are highly subject to personal perspective. Diversity of human control over digital decisions will lead to the application of more necessary and relevant data, and therefore more effective machine learning systems.

(4) SOCIAL AND ECONOMIC IMPLICATIONS:

By 2020, more than a third of the desired core skill sets for most occupations will be composed of skills not yet considered critical to today's jobs. Remote operators may need to help self-driving vehicles manage emergencies, or ride-along UPS concierges may need to manhandle packages and knock on doors. Humans will still need to write dialogue and train corporate chatbot and customer-service; AI will have to be constantly updated and maintained. No matter how advanced AI becomes, research shows that humans will likely perform some jobs better, particularly positions involving creativity, empathy, or social interaction. This category includes doctors, therapists, hairdressers, and personal trainers, as well as scientists, technologists, and artists able to create. There are two challenges ahead: helping existing workers acquire new skills so that they can engage in the AI workforce, and preparing future generations for an AI-integrated workplace.

* Government action must be timely and responsive. The experience of the 19th century shows that technological transition can have a short-term traumatic impact on specific segments of society. In the industrial revolution, economic growth exploded after centuries of stagnant living standards, but governments took nearly a century to respond with new education and welfare systems. Decades passed before wages increased across the board, and the rapid shift of growing populations from farms to urban factories contributed to unrest across Europe. We must move more quickly to address AI.

* Fund creative, just-in-time education models. Having a solid foundation of basic literacy, numeracy, and civic skills will be vital to success in the workplace. We must also make it easier for workers to acquire new skills and switch jobs more easily and quickly than in the past.

Incentivize lifelong learning. Required job skills may change as frequently as every three to five years, and we need to invest in ongoing education opportunities. For example, community college programs often combine education with learning on the job. Apprentices

can graduate with a degree in mechatronics — merging electronics and mechanical engineering — while working in the industry, all without incurring student debt. A different model includes online learning programs that employees can tap into any time while on the job.

* Promote social and collaboration skills training. Social skills including persuasion, emotional intelligence, and teaching others will be in high demand across industries. Research suggests employers will highly value "character skills" such as perseverance, sociability, and curiosity, which correlate closely with employees' ability to adapt to new situations.
* Close the "job polarization" gap. What determines a job's vulnerability to automation is not so much whether the work under consideration is manual or white-collar, but whether or not it is routine. The workforce bifurcates into two groups doing non-routine work: (1) highly paid, skilled careerists, such as architects and psychiatrists, and (2) low-paid, unskilled laborers, such as cleaners and gardeners. As Jerry Kaplan of Stanford said, automation is "blind to the color of your collar". As a result, "job polarization" occurs: middle-skill jobs decline, like those in manufacturing, but both low-skill and high-skill jobs expand.
* Get individuals online. With the increasing presence of AI, it becomes more critical to connect individuals in all demographic sectors to affordable, consistent, and reliable high-speed internet, and access to online services such as platforms for free expression and access to information. Increasingly, we will connect and collaborate remotely with freelancers and independent or "on-demand" professionals through digital talent platforms.
* Manage skills disruption by transitioning the workforce. Automation redefines jobs in ways that reduce costs and boost demand. In an analysis of the American workforce between 1982 and 2012, employment grew significantly faster in occupations that made use of computers, like graphic design. When automation sped up one aspect of a job, it enabled workers to perform other parts of the job better. The net effect resulted in more computer-intensive jobs, while displacing less computer-intensive positions.
* Incentivize paid mid-career internships. Professional mid-career internships provide employees the opportunity to establish work experience in different fields by engaging in internships that are part work, part training, and part exposure. This allows for short-term arrangements that help transition those who have lost their jobs to AI, while mitigating the losses affiliated with long-term unemployment.
* Collaboration between the private and public sectors. AI demands multi-sector partnerships and collaborations that leverage the expertise of each partner in a complementary manner. These are indispensable to implementing scalable solutions to jobs and skills challenges. The government should call for bolder leadership and strategic action within companies and across industries, including partnerships between public institutions and the education sector.
* Reinvigorate programs like Americorps Vista to focus on technology. Federally-funded programs should incorporate a transitioning workforce. Quantitative and qualitative accountability measures will help ensure these programs benefit employers as well as employees, target specific demographics, bring needed technical skills into more fields, and

enhance diversity in the workplace.
* Invigorate and Fund MakerSpace Communities. In the spirit of the maker movement promoting a do-it-yourself mindset and the President's Nation of Makers Initiative, a transitioning workforce can engage in community outreach programs that involve adults and children learning about technology and creating products together. Individuals can deepen their technology experience, shape their environment as creators, and build new products with technology in their own market ecosystem. The government can further promote hobbyists, enthusiasts, and students to transform innovation, culture, and education in the AI space.
* Consider safety net protections. Concerns about AI and automation have led to calls for a stronger social safety net to protect people from labor-market disruption and help them switch to new jobs. In addition to the job training discussed elsewhere, the government should evaluate and consider what type of financial assistance may be needed for those individuals and families who are transitioning between jobs as a result of AI.
* Ethics and civics education. One of the most difficult and growing policy debates is how to build accountability into AI systems that seem to have lives of their own. A tremendous incentive exists to innovate and lead the way in fair automation, and success comes down to consideration of ethics by humans engaged in the process. Throughout this debate, humans remain at the heart of automation through building, testing, refining, auditing, and evaluating these systems.

(7) SCIENTIFIC AND TECHNICAL TRAINING:
Investment by both the public and private sectors in scientific and technical training to prepare for AI is critical. Training must promote user-oriented methods from engineering and design to enact multidisciplinary processes and methodologies for developing technologically feasible products. We have seen several universities launch projects with similar goals with overwhelming success. The human-centered approach to innovation taps into the ability to recognize patterns and construct functional, emotionally meaningful ideas. The method views innovation and creativity as skills that can be gained, and focuses on inspiration, ideation, and implementation. Users are at the center of design while generating, developing, and testing ideas. The method draws on engineering and design principles to help create insights for the business world. This specialization will grow as AI matures because the human element is critical to every technological creation, and demands government recognition.

(9) ADDITIONAL CALLS TO ACTION:
Increase understanding of this technology. Terms like AI are often used when people are actually discussing machine learning, robotics, or deep learning. Artificial intelligence refers to the engineering discipline of making machines intelligent. Machine learning, in contrast, refers to a particular subfield within artificial intelligence that focuses on drawing inferences from a large set of examples. Jobs at the intersection of AI, robotics, and deep learning will be drastically different in just a few years' time, leading to the creation of new disciplines to explore.

Respondent 117

Miriam Young, DataKind

DataKind Responses to White House Office of Science and Technology Policy Request for Information: Preparing for the Future of Artificial Intelligence

DataKind is responding to the following topics as designated 'of interest' by the OSTP:

(2) The Use of AI for Public Good

From early warning systems to providing individualized social services to predicting future needs, at DataKind, we believe AI has the potential to address humanity's toughest challenges. However, more than the right technology, we also know it takes the right mix of people to build meaningful solutions. To apply AI for public good at scale, support is needed for intermediaries that can broker effective collaborations.

We now have more information than ever to understand and improve our world. So-called "big data" pours off our cell phones and laptops and is now easily accessed in new forms like satellite imagery. AI and data science use statistics and computing to help humans make decisions from information that would otherwise be too overwhelming or unruly to process. While AI can be used to make comfortable lives more comfortable, we believe such powerful technology should be used to improve the lives of those most in need.

Over the past five years, we've worked to do just this by engaging hundreds of pro bono data scientists on projects with organizations like Amnesty International to predict future human rights violations, American Red Cross to prevent home fires nationwide, or the World Bank to understand global poverty levels using satellite imagery.

While there is no shortage of opportunities for AI and data science to benefit the public good, there are significant obstacles for it to be applied at scale. The organizations working to address social issues - nonprofits, government agencies, social enterprises, etc. - often have the fewest resources to make investments in the staff and systems needed to take advantage of AI the way in which companies do. Furthermore, the professionals with the skills to build AI for public good can be difficult to identify, and often prohibitively expensive for these organizations to hire. However, from our experience building a global community of thousands of data science volunteers, we know that these talented professionals are eager to utilize their skills and make an impact for social good.

The market will not solve this problem on its own. If AI is to be broadly applied to tackle tough social issues, the government must supply resources and incentives for companies, data scientists, nonprofits, and intermediaries to create AI for the public good. This includes making funding available for social change organizations to invest in their own internal

systems so they can adopt AI solutions. It also includes making funding available for intermediaries that know how to scope projects and leverage pro bono data scientists to deliver the work. There is enormous potential to use AI in service to humanity so long as we have intermediaries that can foster the needed cross-sector collaborations between data scientists and social actors.

(3) Safety & Control Issues

For AI to help us build a more just world that reduces human suffering, the government must help the public make informed choices around its use and help establish mechanisms for transparency and accountability for its creators.

First, the government must increase public awareness of how AI and data science are embedded in our daily lives and, furthermore, how these technologies are naturally riddled with human bias that must be questioned. From our news feeds to credit scoring, Americans should understand how their data is being used and how their behavior influences such systems. Algorithms and AI are widely perceived to be impartial ways of using data and computing to make "scientific" conclusions, but the truth is they are the result of many human decisions along the way. What data was used? What tradeoffs were made? We should not let the perception of impartiality become a shield for AI designers to deflect their own accountability or to avoid the rigor of healthy debate and questioning that goes along with any scientific advancement.

Secondly, the government should work with the data science community to create mechanisms that encourage transparency and accountability for creators of AI. One of the key issues in AI is the inherent contradiction in making machines "smarter" by using data produced by imperfect and often unjust social systems. A lack of data or data deserts also exacerbates this, as incomplete or biased data sets can have a major impact. We should work together to access better, more complete data and be mindful of how data can be biased (historical, redlining from zipcodes, etc) in order to mitigate. Our work with VOTO Mobile, for example, aims to help the organization reach more rural women through its mobile surveys since they are often underrepresented and left out of most development projects. Without oversight and sensitivity to such issues, AI can simply become an echo chamber that reinforces historic inequities of racism and sexism, enabling them to persist. AI creators should collectively create and agree to follow best practices based on sound statistics to make clear what conclusions can be drawn from different methodologies and what the limitations are.

Without public education on these issues or clear mechanisms for transparency and accountability, AI and algorithms simply become a "black box" immune from public questioning - a driverless vehicle that can unknowingly drive agendas. If algorithms and AI can so heavily sway the trajectory of Americans' lives - from college admissions to

preventative policing measures - it's the government's duty to ensure they are developed and applied in a way that is consistent with our country's laws and values.

(7) The scientific and technical training that will be needed to take advantage of harnessing the potential of AI technology

Another bottleneck in applying AI for the public good is the shortage of technical talent needed to keep up with the increasing demand for these approaches. While we've built a large global community of data scientists, we know that finding professionals with the right level and combination of skills is challenging. We also recognize that the tech sector continues to struggle with a lack of diversity in general.

As Kate Crawford said in her recent New York Times piece, "artificial intelligence will reflect the values of its creators." Experts like Kate Crawford and Cathy O'Neil have been shining a light on the underlying ethical challenges in machine learning and AI and their disproportionate impact on marginalized communities. To combat this, the government should support diverse educational programs to not only increase the pipeline of future AI and data science experts, but to ensure it's a more inclusive one that reflects the diverse communities that will ultimately be impacted by these technologies.

(9) Additional information related to AI research or policymaking, not requested above, that you believe OSTP should consider.

After working with many organizations worldwide looking to leverage data science and AI for social good, we've identified common pitfalls and learnings relevant to any AI creators. First, finding problems can be harder than finding solutions. "Problem discovery" or uncovering and articulating the needs of a social change organization is one of the biggest bottlenecks in this work, taking hours of conversation and translation between subject matter and data science experts.

That's why a second lesson learned is that communication is more important than technology. The quality of our conversation and debate directly impacts the quality of the solutions we develop. We have to actively break down silos, let go of industry-specific jargon and instead demystify concepts for the public so that more people can join and diversify the conversation.

Finally, because AI can have such far-reaching impacts on human lives, we should always follow a human-centered approach in its design. As the civic tech leader Laurenellen McCann says, "build with, not for." Practically, this means getting out from behind our desks to seek input from subject matter experts and the members of the communities impacted by our work.

Because these pitfalls are so difficult to avoid in this work, intermediaries are needed to facilitate successful collaborations. Intermediaries can help non-technical subject matter experts articulate needs and can serve as translators between parties throughout. We also know that, like much of science, the nature of this work is iterative. Oftentimes we find that teams set down one path only to discover an entirely different approach is needed, or another need has surfaced that must be addressed first. Intermediaries can shepherd teams through these winding paths to ensure that the ultimate solution developed has a meaningful positive impact on social change organizations and sector issue areas around the world.

About DataKind
DataKind is a nonprofit that harnesses the power of data science in the service of humanity. We engage data science and social sector experts on projects addressing critical humanitarian problems and lead the conversation about how data science can be applied to solve the world's biggest challenges. Launched in 2011, DataKind is headquartered in New York City and has Chapters in Bangalore, Dublin, San Francisco, Singapore, the UK and Washington DC. More information on DataKind, our programs and our partners can be found on our website: www.datakind.org.

Respondent 118

Corrine Yu, The Leadership Conference on Civil and Human Rights

July 22, 2016

Terah Lyons
Office of Science and Technology Policy
Eisenhower Executive Office Building
1650 Pennsylvania Ave., NW
Washington DC 20504

Dear Ms. Lyons,

On behalf of The Leadership Conference on Civil and Human Rights, a coalition charged by its diverse membership of more than 200 national organizations to promote and protect the civil and human rights of all persons in the United States, we appreciate this opportunity to provide comments in response to the Office of Science and Technology Policy's Request for Information (RFI) regarding Artificial Intelligence (AI). Our comments will focus on questions 1, 2, 3, and 4 of the RFI. In brief, we believe AI must be governed and controlled in such a way as to promote the public good by protecting and enhancing civil and human rights. Thus, the need to protect civil rights in automated decisions should be considered a vital part of the research agenda for AI.

We join many other stakeholders in recognizing the extraordinary value of data and the growing benefits of AI in daily life. Automated, data-driven decisions can, at their best, bring greater fairness and equity to the key turning points in people's lives. For example, AI systems used in the hiring process have led some companies to hire more job applicants lacking college degrees—applicants who, even when they are excellent candidates, are often overlooked by human recruiters.

At the same time, just because a decision is made by an AI system does not necessarily mean that it is fair or unbiased. For example, AI systems often base their decisions on historical data about people and groups. Such data frequently reflects the longstanding, ongoing reality of racial and other bias—at both an individual and a structural level—that sadly still pervades many areas of American life.

In 2014, The Leadership Conference was pleased to join with a broad national coalition of civil rights, technology policy, and media justice organizations in endorsing Civil Rights Principles for the Era of Big Data, which are online at http://civilrights.org/bigdata. A related report, offering key examples of the ways big data can impact civil rights, has been published at https://bigdata.fairness.io.

As our Principles make clear, we believe it is vital to "ensure fairness in automated decisions." This means that AI systems whose decisions impact civil rights — for example, AI systems that make decisions about who gets a job interview, or about who will be stopped for police questioning — must be designed to ensure that they will protect the civil and human rights of all people.

The diverse signatories to the Civil Rights Principles for the Era of Big Data share serious concerns about the risks posed by biased data, and the biased assumptions or unfair decisions that may result from uncritical uses of such data. As the Principles explain: "Systems that are blind to the preexisting disparities faced by … communities [that are disadvantaged or that have historically been the subject of discrimination] can easily reach decisions that reinforce existing inequities. Independent review and other remedies may be necessary to assure that a system works fairly."

Finally, even when the engineers who build a system hope and aim for the best, there is a significant risk of disparate impact in many AI systems. We are pleased to note that a small but growing community of AI researchers is exploring new ways to diagnose and address discrimination in AI systems. We believe this research should be supported and bolstered.

Thank you for embarking on this important process. We stand ready to work with you to ensure that the voices of the civil and human rights community are heard in this important, ongoing national conversation. If you have any questions about these comments, please contact Corrine Yu, Leadership Conference Managing Policy Director, at XXXXXXXXXor

XXXXXXXXX.

Sincerely,

Wade Henderson
President & CEO

Nancy Zirkin
Executive Vice President

Respondent 119

Sven Koenig, ACM Special Interest Group on Artificial Intelligence

This response to the OSTP request is from the officers and advisory committee members of the ACM Special Interest Group on Artificial Intelligence. The Association for Computing Machinery (ACM) is the world's largest computing society.

(1) The legal and governance implications of AI

AI technologies can reason about the world, learn from data, and determine effective courses of action. In specialized domains, AI technologies can sometimes perform faster and better than humans. They can be used to generate new insights, support human decision making, or make autonomous decisions. The rapid development of AI raises similar concerns as for other automation and information-processing technologies, namely over issues such as loss of privacy due to data collection and combination, changes in social equity, and who will be responsible for operational oversight of AI systems and liable for their bad decisions. Given the potential of AI systems to profoundly affect individuals and society, best practices, frameworks, guidelines, and standards for these systems should be developed as necessary to address, for example, their testing, application, compliance with ethical norms, and the monitoring of their operations. It is also important that people understand the strengths and limitations of AI. The government should prioritize the funding of research that studies these issues and suggests possible approaches. It should also facilitate discussions in working groups that bring together stakeholders from different application areas with AI experts, domain experts, policy makers, and lawyers.

(2) The use of AI for public good

AI technologies are already used, often behind the scenes, for public good, from finding information on the internet to reasoning about patient outcomes (and guiding therapies) to keeping airports safe, to take just three examples. While this is, and will continue to be, a key focus of many researchers and practitioners in AI, there is a critical need for more

federal funding of research for these kinds of applications. This is particularly true for applications that have both social and economic value (for example, support for people with disabilities entering the workforce) but are hard to fund for either national security or industrial applications (which are currently better funded).

(3) The safety and control issues of AI

In domains where AI is already having a significant impact on our lives, it is important to pay attention to safety and control issues. Some applications (for example, autonomous driving) are more safety-critical than others (for example, making movie recommendations) and thus require more careful validation and testing. Safety-critical systems need to be designed to minimize harmful outcomes, and to rely on carefully verified software or monitoring by humans or software. Investments in AI have led to remarkable technological successes; we expect this return on investment to continue, but one direction that is under-explored but of increasing importance is research that improves our understanding of how to create reliable AI systems (for example, via behavior verification or detecting behavior anomalies). In order to get better at building safe and reliable AI systems, we will need a mix of better adherence to standard verification and validation technology, as well as AI-specific additions (for example, for testing AI systems that learn).

The public discourse around safety and control would benefit from demystifying AI. The media often concentrates on the big successes or failures of AI technologies, as well as scenarios conjured up in science fiction stories, and features the opinions of celebrity non-experts about future developments of AI technologies. As a result, parts of the public have developed a fear of AI systems developing superhuman intelligence, whereas most experts agree that AI technologies currently work well only in specialized domains, and notions of “superintelligences” and "technological singularity" that will result in AI systems developing super-human, broadly intelligent behavior is decades away and might never be realized. AI technologies have made steady progress over the years, yet there seem to be waves of exaggerated optimism and pessimism about what they can do. Both are harmful. For example, an exaggerated belief in their capabilities can result in AI systems being used (perhaps carelessly) in situations where they should not, potentially failing to fulfil expectations or even cause harm. The unavoidable disappointment can result in a backlash against AI research, and consequently fewer innovations. It is therefore important to better educate decision makers and the public about the state of the art in AI technologies and their fundamental limits, as well as to solicit input from AI researchers and their professional organizations (such as ACM SIGAI and AAAI) in any decision-making process.

(4) Social and economic implications of AI

AI technologies can have large social and economic implications, with potentially transformative benefits for industry, education, health care, safety, and other areas of our daily lives, but also with the potential for job disruption and technological unemployment

(which might require retraining for new jobs and other interventions). It is critical to support AI research and development in order to maximize the benefits to society while also ensuring that the entire population benefits with regard to standard of living, distribution and quality of work, and productivity.

(5) The most pressing, fundamental questions in AI research, common to most or all scientific fields
(6) The most important research gaps in AI that must be addressed to advance this field and benefit the public

(Combined response)

- How to develop AI systems that can operate in a robust and valuable manner over extended time periods, in open, changing, and unforeseen situations, and with real-time responsiveness?

- How to validate the behavior of AI systems, and predict and provide guarantees on their efficiency and effectiveness on such dimensions as reliability, robustness, and safety? This difficult since these systems can be complex, reason in ways different from humans, and use learning to change their behavior over time.

- How to develop AI systems that can reason about their own operations, capabilities and limitations, including performing self-monitoring and self-repair?

- How to build AI systems that can interact naturally and work together with humans, understand their own strengths and limitations, reflect social norms, human values, and codes of conduct, be capable of explaining themselves and acting in understandable ways (even when they reason very differently than humans), and are trusted by the humans?

- How to integrate ideas and successes from different AI areas, together with those from other disciplines (such as operations research, economics, control theory, and perceptual, cognitive, and social psychology) to create more broadly capable AI systems?

(7) The scientific and technical training that will be needed to take advantage of harnessing the potential of AI technology, and the challenges faced by institutions of higher education in retaining faculty and responding to explosive growth in student enrollment in AI related courses and courses of study

AI is inherently an interdisciplinary field, with computer science at its core. We must build the pipeline early by emphasizing computer science education early in K-12, increasing the technological sophistication of the entire population. AI technologies have the potential to transform learning and greatly increase access to education, and AI and computer science education researchers are uniquely positioned to both build and use the technology.

The explosive growth of interest in AI is both an opportunity and a challenge. Universities are losing valuable human resources to industry at every stage of the pipeline, and recruiting and retaining the highest calibre Ph.D. students and faculty in basic research is increasingly difficult. AI is therefore in need of significant investment in research and education. Basic research is inherently risky; it can take years to work out ideas, only some of which eventually result in breakthroughs. Academic research depends on reliable long-term funding, otherwise researchers are forced to change their research directions frequently, limiting potential impact. Universities thus often pay attention to the availability of funding for different areas when allocating faculty positions. Current academic funding is highly competitive and often short-term. Consequently, many researchers spend a lot of their time writing grant proposals rather than doing research. A number of prominent AI researchers have recently left for industry or for other countries due to financial incentives, long-term funding guarantees, or better computing infrastructures. Increased funding for basic research in AI is essential to respond to this problem.

(8) The specific steps that could be taken by the federal government, research institutes, universities and philanthropies to encourage multi-disciplinary AI research

While AI needs the application-pull that often comes from industry in addition to the technology-push that often comes from academia, industry is often heavily focused on short-term applied research and restricts the dissemination of results. AI needs long-term basic research across all its sub-areas. It is important to provide funding for both AI faculty members (for example, via grants) and students (for example, via fellowships). Some of the funding should be long-term funding for independently proposed research, vetted via an NSF-like review process that utilizes AI experts. For example, one could create funding for "AI Fellows" to support the research of promising AI faculty members in addition to focused research toward specific objectives.

AI researchers often study individual AI functions (such as learning, reasoning or planning) much more than their interplay. Furthermore, while AI has interfaces to a variety of disciplines, ranging from optimization disciplines (such as operations research, economics, and control theory) to social sciences, these research communities have little overlap. It takes time and effort and is risky to build multi-disciplinary AI research groups. Long-term funding for basic research is required for such efforts. In addition to potentially supporting large, center-type multidisciplinary research groups (that could include industrial collaborations), it is critical to ensure sufficient long-term basic research funding for smaller groups of two or three researchers from different disciplines working together, since those collaborations often enable fundamental breakthroughs. Funding models that span from basic and long-term research to applied and short-term development are necessary.

(9) Specific training data sets that can accelerate the development of AI and its applications

Translation of the domain-independent AI technology developed by AI researchers to concrete applications is easiest when there are relevant datasets available. However, industrial datasets are typically proprietary and not freely available for research. Algorithmic AI researchers could be incentivized to focus on specific applications by providing them with data sets and simulators for a variety of application domains where AI technologies could have the potential for significant positive impact. It is important to note that AI researchers often study individual AI functions (such as learning, reasoning or planning) and are only starting to determine how to combine them in a principled way into overall AI systems. Thus, datasets and simulators should be developed and provided for both individual AI functions and complete applications.

(10) The role that "market shaping" approaches such as incentive prizes and Advanced Market Commitments can play in accelerating the development of applications of AI to address societal needs, such as accelerated training for low and moderate income workers

A larger number of competitions already exist in AI (see, for example, the competition report column in AI Magazine that, in the past 4 years, has covered 23 different competitions). Most of these competitions offer only small monetary prizes, if any. Larger incentive prizes could accelerate the deployment of AI technology and AI systems. It is important to understand, though, that competitions often attract larger organizations that can afford the expense of participating even when the chance of receiving prize money is low. It is therefore important to provide funding for the actual research as well, not just to reward the research groups with currently high-performing technologies.

Furthermore, many promising AI technologies are still far from being directly useful for applications. Researchers must be provided with incentives to develop these technologies as well, rather than concentrating only on those with short-term benefits. Research is often best advanced with well-endowed long-term funding programs for both basic and applied research that provide a steady funding stream to a variety of research groups, large and small, across the country.

Respondent 120

Michael Littman, Brown University

Fueled by advances in artificial intelligence (AI) technology, robotics is poised to have similarly transformative impacts to AI on society, labor, and safety. Brown University's Humanity Centered Robotics Initiative (HCRI) believes that OSTP should encourage the creation of an NSF-funded research center for studying and sharing the social, legal and security implications of robotics. Ideally, such a center would be well positioned to orient research efforts to societally beneficial problems and to inform the government of cutting edge results and their implications.

The research community has made steady progress on the basic problems of robot perception, decision making, and motion, and we expect increasing numbers of robots engaged in productive interactions with the general public. Robotics, as the physical instantiation of automation and AI, will ultimately have extensive impacts on social and economic structures. In addition, there will be significant legal, policy and social challenges to safely integrating robots into daily life. Creating systems that navigate the human physical and social world presents significant challenges in human robot interaction, ethical use, and overall safety.

The research team for this center would have to include a diverse group of investigators including social, behavioral, and cognitive researchers, engineers and computer scientists, and legal scholars and ethicists. The types of questions explored by such a center could include:

- Robots that benefit society: The impact of robots on health care, social care, and labor. Robots in eldercare. Robot use in environmental science. The use of co-robots in dangerous working conditions.

- Ethical and legal challenges in robotics: Ethical and legal frameworks for integrating robots into society. Designing robot decision mechanisms that obey laws and social norms and are intuitive to use for people without technical background.

- Safety and control of robotic systems: Cybersecurity implications for robots, impact of robots on social relationships, algorithms and approaches that produce verifiable guarantees and are aligned with human values, improving robot control systems with automation and perception.

There is a strong need for research to improve human-robot interaction, robot perception, learning, decision making, and safe and efficient operation. As robots in general society increase in numbers and the narrow AI that drives them becomes more complex, the role of AI in robotics will become more prominent. Understanding how those systems interface with people and social systems will be paramount in the safe and efficient deployment of this new technology for the benefit of all.

Respondent 121

Richard Greenblatt, Minsky Institute @ MIT (embryonically organized)

Myself and a group met with Marvin Minsky at his house on 111 Ivy Street, Brookline, Mass one night a week last fall.

We discussed what a "Minsky Institute" might consist of and what its mission statement

might be. Unfortunately, Marvin passed away before anything fully converged. One idea, which was discussed and some areas of agreement reached was for a central clearing house which would facilitate the exchange of data by AI programs. A bare start was made on a few technical standards which might eventually facilitate things. Another idea, barely connected to the previous one, was that the Minsky Institute might serve as moderated forum
where the capabilities and status (current and projected) of various systems could be discussed, both by people within a particular project and by knowledgable workers in the field. In such a manner, a certain degree of "vetting" might be achieved. I think there is clearly a need for such vetting and hope a way can be found for this to occur thru such an organization as the Minsky Institute might yet become.

Respondent 122

William Rinehart, American Action Forum

The techniques of artificial intelligence are quickly being adopted in medicine, life science, and for analysis of big data. Continuing progress will require the adoption of optimal legal and regulatory frameworks. Federal and local governments can foster these technologies by promoting policies that allow individuals to experiment to further the progress of AI. This can be achieved by considering the costs in applying liability rules, intervening only when there are demonstrable benefits, providing room to experiment, and allowing for trade in technology and ideas.

Defining AI
The term AI often conjures up an early 1990s image of Arnold Schwarzenegger and more recently, Samantha from Her. AI of this kind, often called strong AI, is far from our current technological capabilities and may never be achieved. While some fret over the risks from super intelligent agents with unclear objectives, task specific AI holds immediate promise. Narrow AI is a term for a collection of economic and computer models built using real world data to achieve specific objectives. These objectives might include translating languages, better predicting the weather, spotting tumors in chest scans and mammograms, and helping people identify caloric information just from pictures of food.

Understanding AI Risks
In the course of searching for solutions, AI will encounter negative events. Risk management pioneer Aaron Wildavsky rightly defines risk this way, situating it as a byproduct of the search for welfare enhancing economic activity. Indeed, there is an important and deep relationship between wealth and risk, which should be little surprise considering that risk and return are correlated and the bedrock of finance theory.

Researchers in AI have identified three concrete problems that could contribute risk to AI:

1. The objective was incorrectly specified, leading to negative side effects or cheating by the AI;
2. The designer might have an objective in mind, but the costs in evaluating how well the AI is performing on these objectives make harmful behavior possible, or
3. The objective is clearly specified but some unwanted behavior occurs because the data is poorly curated or the model isn't sufficiently expressive of the environment it is interacting with.

Because of the variety of domains where AI is being applied, how these problems manifest will vary. AI applications within medicine will see different kinds of issues arise as compared to weather prediction models. How legal systems should be designed will and should vary considerably, depending on the already existing institutional environment, which includes the legal and moral 'rules of the game' that guide individuals' behavior, and the industry specific institutional arrangements, which generally define how companies are organized.

Because of the varying contexts, there is no workable one size fits all regulatory framework. All of this should give pause to any sort of broad regulatory effort to limit the application of AI, much as the European Union is currently contemplating. Optimal levels of regulation must be discovered, so much of the work is likely to come from court cases.

Creating Rules for Liability
Since AI innovation is in its early stages, it is too soon to determine which liability rules, rules of evidence, and damage rules for various jurisdictions should apply.

Autonomous vehicles serve as a good example of the complexity of this question. While autonomous cars won't radically replace all cars in the next five years, they are likely to come into increasing contact with human drivers and cause accidents. Four basic kinds of rules apply for fault in car accidents, which states have adopted in varying degrees. Most jurisdictions have adopted comparative fault, so damages are apportioned based on their proportionate shares. Over time, courts will likely shift the allocation of burden to human drivers if driverless cars prove safe. However, one can also imagine cases where courts will assign some percentage of fault to the autonomous auto-manufacturer based on an adaptation of product liability law. How these cases play out will depend heavily on the degree of AI implemented and the control that manufacturers allow for drivers.

Product liability is a complex area of law and should be allowed to adapt to the challenges of AI. However, there should be more focus on the costs of the system. The available empirical evidence suggests that there isn't a measurable effect on the frequency of product accidents due to varying product liability regimes. In other words, the purported safety benefits of product liability might not exist when real world costs are considered. Moreover, given the current legal system, a significant portion of the compensation that is meant to pay damages to victims goes instead to transactions costs in the form of legal fees. In total, there are

reasons to believe that enterprise is less likely to engage in those kinds of activities, which could be a deterrent to AI development. Thus, researchers should work towards better understanding how economically efficient these rules are in practice and jurisdictions should be careful on how they apply old rules onto new technologies, especially given the costs.

Openness to experimentation

Nevada was the first state to allow for the operation of autonomous vehicles in 2011 and has since been joined by five others, including California, Florida, Michigan, North Dakota, and Tennessee, as well as Washington D.C. While it is not guaranteed, these states will likely lead the way in developing autonomous vehicles since they are creating zones where the technology can start pushing down the risks. The long term benefits will come in the shape of investment and jobs. For policymakers, removing barriers to experimentation should be the utmost priority. This can broadly be achieved by adopting a mindset of permissionless innovation.

As we currently see in the computer security field, tools will be devised that search for these problems and correct them, similar to how algorithms can search for bad code and cybersecurity threats. Creating zones of experimentation where the three types of risk can be worked out will lead to a greater level of safety. The benefits may come in the form of laws passed in those five states and the District of Columbia, or perhaps via limited liability. Experimental spaces will ensure incentives are aligned to research and develop AI.

Given how promising these technologies are, prescriptive federal regulation is hardly justifiable at this time. In applying the old regulatory regime to these new spaces, regulators should be mindful of the three-part test:

1. Prove the existence of market abuse or failure by documenting actual consumer harm;
2. Explain how current law or rules are inadequate, and show that no alternatives exist including market correctives, deregulatory efforts, or public/private partnerships to solve the market failure; and
3. Demonstrate how the benefits of regulation will outweigh the potential countervailing benefits, implementation costs, and other associated regulatory burdens.

Openness to trade

While the United States is at the forefront of AI development, there is no guarantee that advances will always be made here. Two basic principles flow from this. First, the US should maintain an openness to trade with other countries and ensure that there are not any trade related encumbrances, especially in data transfer. Second, we should be a leader in this space by encouraging our closest trading partners, including those in the EU, to abandon myopic views of AI and allow for more experimentation with the available tools. Research and development has globalized and only if we embrace that reality will the U.S. be able to reap the rewards.

Digital literacy
Digital literacy needs to be emphasized. As compared to media literacy and computer literacy, digital literacy focuses on imparting knowledge of complex network systems and big data, as well as critical thinking skills to understand how these systems relate to stand alone devices. For states and local government, this doesn't translate necessarily into a need for all students to be able to code, but to at least appreciate how technology works. While they are sure to involve educational institutions, strategies for digital literacy will likely better serve everyone if they originate from local communities and users of the technologies instead of the federal government via strict mandates.

Conclusion
Much like the beneficial uses for AI, the optimal legal and regulatory institutions for AI will have to be discovered. While many reflexively coil when hearing the term AI, the narrow version of AI might offer some real benefits. Federal and local governments can foster these technologies by being supportive but taking a hands off approach in helping to mitigate that risk and allowing the legal system to do its job. Progress in this space will depend on how comfortable we are with new machine-human partnerships. To accomplish this, we do not need more laws and institutions, but more trust in the ones that already exist.

Respondent 123

Daniel Castro, Center for Data Innovation

July 22, 2016

Attn: Terah Lyons
Office of Science and Technology Policy,
Eisenhower Executive Office Building, 1650 Pennsylvania Ave. NW,
Washington, DC 20504

On behalf of the Center for Data Innovation (datainnovation.org), we are pleased to submit these comments in response to the Office of Science and Technology Policy's (OSTP) request for information on the overarching questions in artificial intelligence (AI).1

The Center for Data Innovation is the leading think tank studying the intersection of data, technology, and public policy. With staff in Washington, DC and Brussels, the Center formulates and promotes pragmatic public policies designed to maximize the benefits of data-driven innovation in the public and private sectors. The Center is a non-profit, non-partisan research institute affiliated with the Information Technology and Innovation Foundation.

AI is a part of computer science devoted to creating computing machines and systems that

perform operations analogous to human learning and decision-making.1 Technological advancements over the past decade demonstrate that AI will become dramatically more powerful and effective at solving everything from mundane challenges, such as helping consumers figure out what to buy for the holidays, to the most pressing social and economic challenges, ranging from diagnosing and developing new treatments for devastating diseases to dramatically improving worker productivity. In this submission, we outline some of the most significant benefits and challenges of AI so that policymakers can take an active role in supporting the development of AI, as well as avoid succumbing to widespread yet unfounded alarmist narratives about how AI is a threat to economic and social well-being or even an existential threat to humanity.

Our responses to the relevant questions are in the attached document.

Sincerely,

Daniel Castro
Director
Center for Data Innovation
XXXXXXXXX

Joshua New
Policy Analyst
Center for Data Innovation
XXXXXXXXX

THE LEGAL AND GOVERNANCE IMPLICATIONS OF AI

It was relatively clear how traditional software systems made decisions, as parameters were built in and largely understandable. In contrast, many AI systems make decisions based on complex models developed by an algorithm that continually adjust and improve based on experience. Since these adjustments may involve obscure changes in how variables are weighted in computer models, some critics have labeled these systems "black boxes" that are likely to create "algorithmic bias" that enables government and corporate abuse. These critics generally fall into two camps: those that believe companies or governments will deliberately "hide behind their algorithm" as a cover to exploit, discriminate, or otherwise act unethically; and those who argue that opaque, complicated systems will allow "runaway data" to produce unintended and damaging results.2

But resistance to AI because of these concerns fails to recognize a key point: AI, like any technology, can be used unethically or irresponsibly. AI systems are not independent from their developers and, more importantly, from the organizations using them. If a government

or business wants to systematically discriminate against certain groups of persons, it does not need AI to do so. To put it simply, bad actors will do illegal things with or without computers.3

Nonetheless, many critics seem convinced that the complexity of these systems is responsible for any problems that emerge, and that mandating "algorithmic transparency" is necessary to ensure that the public can police against biased AI systems.4 Combatting bias and protecting against harmful outcomes is of course important, but mandating that companies open their propriety AI software to the public would not solve these problems, but would create new ones. Consumers and policymakers are ill-equipped to actually understand the complicated decision-making processes of an AI algorithm, and AI systems can learn and change over time, making it difficult to measure unfairness by examining their underlying mechanics.5 Moreover, the economic impact of such a mandate would be significant, as it would prevent companies from capitalizing on their intellectual property and future investment and research into AI would slow.

Fortunately, many have recognized that embedding ethical principles into AI systems is both possible and effective. The White House's framework for "equal opportunity by design" in algorithmic systems, as described by its report on the opportunities and challenges of big data and civil rights presents, is one promising method.6 This approach, described more generally by Federal Trade Commissioner Terrell McSweeny as "responsibility by design," rightly recognizes that algorithmic systems can produce unintended outcomes, but does not demand a company waive rights to keep its software proprietary.7 Instead, the principle of responsibility by design provides developers with a productive framework for solving the root problems of undesirable results in algorithmic systems: bad data as an input, such as incomplete data and selection bias, and poorly designed algorithms, such as conflating correlation with causation, and failing to account for historical bias.8 In particular, the federal government should help address the problem of data poverty, where a lack of high-quality data about certain groups of individuals puts them at a social or economic disadvantage.9

It also is important to note that some calls for algorithmic transparency are actually more in line with the principle of responsibility by design. For example, former chief technologist of the Federal Trade Commission (FTC) Ashkan Soltani said that although pursuing algorithmic transparency was one of the goals of the FTC, "accountability" rather than "transparency" would be a more appropriate way to describe the ideal approach, and that making companies surrender their source codes is "not necessarily what we need to do."10 Rather than make companies relinquish their intellectual property rights, encouraging adherence to the principle of responsibility by design would allow companies to better police themselves to prevent unintended outcomes and still ensure that regulators could intervene and audit these systems should there be evidence of bias or other harms.11 For example, policymakers could work with the private sector to develop a framework for predicting and accounting for disparate impact in AI systems. Companies would be more

willing to deploy AI if they could clearly demonstrate they are acting in good faith by actively considering how their systems could produce discriminatory outcomes and taking steps to address these concerns.12

Figuring out just how to define responsibility by design and encourage adherence to it warrants continued research and discussion, but it is crucial that policymakers understand that AI systems are valuable because of their complexity, not in spite of it. Attempting to pull
back the curtain on this complexity to protect against undesirable outcomes threatens the progress of AI.

THE USE OF AI FOR PUBLIC GOOD

AI systems can help organizations make better-informed, timelier decisions, as well as tackle complicated, data-intensive problems that humans are ill-equipped to solve. As such, the potential benefits of AI are will likely be quite large. There are already compelling examples of AI offering substantial benefits for civil rights, public health, conservation, energy efficiency, financial services, and healthcare.13 In addition, AI has the potential to make government substantially more efficient and citizen-friendly if government agencies adopt the technology.

THE SOCIAL AND ECONOMIC IMPLICATIONS OF AI

The criteria for measuring the success of AI should not be whether it is "perfect" but rather if it is an improvement over the status quo. AI offers an unprecedented opportunity to automate decision-making and reduce the influence of explicit and subconscious human bias that permeate every aspect of society and the economy.14 AI systems make decisions based on data, and quantifying and analyzing the decision-making process can both expose the underlying bias exhibited by human-made decisions, as well as prevent subjective and potentially discriminatory human decision-making from ever entering the equation.15

Regarding economic implications, one of the most widely-repeated warnings about AI is that it will lead to mass unemployment, as smart machines become increasingly adept at performing work normally carried out by humans in both blue and white collar jobs.16 However, the "AI will destroy jobs" argument is incorrect, for several reasons. First, most AI applications will not be able to fully replace human workers, but rather will automate particular tasks and allow a human worker to spend their time in more valuable ways.17 Very few jobs could conceivably be automated in the short or medium term, and automation will instead transform the function of many existing jobs rather than eliminate them.18

Second, in instances where AI does eliminate jobs, the jobs lost will be offset by the resulting productivity growth that leads to the creation of new jobs. When a business replaces a human worker with an AI system, it does so because the AI increases the business's productivity by doing the job more effectively for lower cost. If jobs in one firm or industry are reduced or eliminated through higher productivity, then by definition production costs go down. These savings are passed on, often in though lower prices or

higher wages. This money is then spent, which creates jobs in whatever industries supply the goods and services on which people spend their increased savings or earnings.19

To be sure, there are winners and losers in the process of productivity improvement: Some workers will lose their jobs, and it is appropriate for policymakers to help those workers quickly transition to new employment. But there is simply no merit in the belief that productivity growth will reduce the overall number of jobs.20

THE MOST IMPORTANT RESEARCH GAPS IN AI THAT MUST BE ADDRESSED TO ADVANCE THIS FIELD AND BENEFIT THE PUBLIC

Ensuring that AI systems produce fair, unbiased, and safe results without mandating algorithmic transparency can pose complicated technical challenges that warrants further research. For example, Carnegie Mellon researchers have developed a method for determining why an AI system makes particular decisions without having to divulge the underlying workings of the system or code.21 This research will be useful to addressing regulators' concerns about discrimination, as well as helping companies that want to ensure they are acting ethically.

Federal research programs should avoid working in social engineering into AI research. For example, the National Science Foundation's National Robotics Initiative focuses on accelerating the development of robotic systems, but only systems that work beside or cooperatively with humans.22 While AI-powered worker assistance applications will be beneficial, limiting the focus of this research in such a manner precludes opportunities to develop AI systems that could replace workers, which history has shown to produce greater economic benefits in the moderate and long run.

CONCLUSION:

AI has the potential to generate substantial benefits to the economy, society, and overall quality of life, and it is encouraging to see OSTP proactively working to better understand the technology, promote its research and development, and set the record straight about the potential opportunities associated with the technology. OSTP should also play an active role in dispelling the prevalent alarmist myths about AI, particularly concerns that AI will lead to higher rates or unemployment and even eradicate the human race, which, besides being wrong, threaten the acceptance and advancement of this technology.

1 "Request for Information on Artificial Intelligence," Federal Register, June 27, 2016, https://www.federalregister.gov/articles/2016/06/27/2016-15082/request-for-information-on-artificial-intelligence.

1 Robert Atkinson, "'It's Going to Kill Us!' and other Myths About the Future of Artificial Intelligence," Information Technology and Innovation Foundation, June 2016, http://www2.itif.org/2016-myths-machine-learning.pdf.

2 Tim Hwang and Madeleine Clare Elish, "The Mirage of the Marketplace," Slate, August 9, 2015,

http://www.slate.com/articles/technology/future_tense/2015/07/uber_s_algorithm_and_the_mirage_of_t he_marketplace.html and David Auerbach, “The Code We Can’t Control,” Slate, January 14, 2015, http://www.slate.com/articles/technology/bitwise/2015/01/black_box_society_by_frank_pasquale_a_chil ling_vision_of_how_big_data_has.html.

3 Katherine Noyes, “The FTC Is Worried About Algorithmic Transparency, and You Should Be Too,” PC World, April 9, 2015, http://www.pcworld.com/article/2908372/the-ftc-is-worried-about-algorithmictransparency-and-you-should-be-too.html

4 For example, the Electronic Privacy Information Center argues that “Entities that collect personal information should be transparent about what information they collect, how they collect it, who will have access to it, and how it is intended to be used. Furthermore, the algorithms employed in big data should be made available to the public.” See Mark Rotenburg, “Comments of The Electronic Privacy Information Center to The Office of Science and Technology Policy, Request for Information: Big Data and the Future of Privacy” (Electronic Privacy Information Center, April 4, 2014), https://epic.org/privacy/big-data/EPIC-OSTP-Big-Data.pdf; See also David Auerbach, “The Code We Can’t Control,” Slate, January 14, 2015, http://www.slate.com/articles/technology/bitwise/2015/01/black_box_society_by_frank_pasquale_a_chilling_vision_of_how_big_data_has.html.

5 Lauren Smith, “Algorithmic Transparency: Examining from Within and Without,” IApp, January 28, 2016, https://iapp.org/news/a/algorithmic-transparency-examining-from-within-and-without/.

6 The White House Executive Office of the President, “Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights,” May 2016, https://www.whitehouse.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf.

7 Terrell McSweeny, "Tech for Good: Data for Social Empowerment" (keynote remarks of Commissioner Terrell McSweeny at Google DC Tech Talk Series, Washington District of Columbia, September 10, 2015), https://www.ftc.gov/system/files/documents/public_statements/800981/150909googletechroundtable.pdf.

8 The White House Executive Office of the President, “Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights,” May 2016, https://www.whitehouse.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf.

9 Daniel Castro, “The Rise of Data Poverty in America,” Center for Data Innovation, September 10, 2014, http://www2.datainnovation.org/2014-data-poverty.pdf.

10 Christopher Zara, “FTC Chief Technologist Ashkan Soltani On Algorithmic Transparency and the Fight Against Biased Bots,” International Business Times, April 9, 2015, http://www.ibtimes.com/ftc-chieftechnologist-ashkan-soltani-algorithmic-transparency-fight-against-biased-1876177.

11 Katherine Noyes, “The FTC Is Worried About Algorithmic Transparency, and You Should Be Too,” PC World, April 9, 2015, http://www.pcworld.com/article/2908372/the-ftc-is-

worried-about-algorithmictransparency-and-you-should-be-too.html.
12 Travis Korte and Daniel Castro, "Disparate Impact Analysis is Key to Ensuring Fairness in the Age of the Algorithm," Center for Data Innovation, January 20, 2015, https://www.datainnovation.org/2015/01/disparate-impact-analysis-is-key-to-ensuring-fairness-in-the-age-of-the-algorithm/.
13 Hope Reese, "AI App Uses Social media to Spot Public Health Outbreaks," TechRepublic, March 9, 2016, http://www.techrepublic.com/article/ai-app-uses-social-media-to-spot-public-health-outbreaks/, "Energy Savings from the Nest Learning Thermostat: Energy Bill Analysis Results," Nest Labs, February 2015, https://nest.com/downloads/press/documents/energy-savings-white-paper.pdf, Derrick Harris, "How PayPal Uses Deep Learning and Detective Work to Fight Fraud," Gigaom, March 6, 2015, https://gigaom.com/2015/03/06/how-paypal-uses-deep-learning-and-detective-work-to-fight-fraud/, and Alexander Kostura, "Artificial Intelligence is Key to Using Data for Social Good," Center for Data Innovation, July 11, 2016, https://www.datainnovation.org/2016/07/artificial-intelligence-is-key-to-using-data-for-social-good/.
14 Joshua New, "It's Humans, Not Algorithms, That Have a Bias Problem," Center for Data Innovation, November 16, 2015, https://www.datainnovation.org/2015/11/its-humans-not-algorithms-that-have-a-bias-problem/.
15 Joshua New, "The White House is Starting to Get it Right on Big Data," Center for Data innovation, June 13, 2016, https://www.datainnovation.org/2016/06/the-white-house-is-starting-to-get-it-right-on-big-data/.
16 Robert Atkinson, "'It's Going to Kill Us!' and other Myths About the Future of Artificial Intelligence," Information Technology and Innovation Foundation, June 2016, http://www2.itif.org/2016-myths-machine-learning.pdf.
17 Ibid.
18 Michael Chu, James Manyika, and Mehdi Miremadi, "Four Fundamentals of Workplace Automation," McKinsey Quarterly, November 2015, http://www.mckinsey.com/business-functions/business-technology/our-insights/four-fundamentals-of-workplace-automation.
19 Robert Atkinson, "'It's Going to Kill Us!' and other Myths About the Future of Artificial Intelligence," Information Technology and Innovation Foundation, June 2016, http://www2.itif.org/2016-myths-machine-learning.pdf.
20 Ibid.
21 "Carnegie Mellon Transparency Reports Make AI Decision-Making Accountable," Carnegie Mellon University, May 26, 2016, https://www.ecnmag.com/news/2016/05/carnegie-mellon-transparency-reports-make-ai-decision-making-accountable.
22 "National Robotics Initiative (NRI)," National Science Foundation, accessed July 20, 2016, http://www.nsf.gov/pubs/2016/nsf16517/nsf16517.htm.

Respondent 124

Huw Price, University of Cambridge, UK

Written evidence submitted in a personal capacity by Professor Huw Price (Bertrand Russell Professor of Philosophy, Cambridge)

Executive Summary

• There are good reasons to take seriously the possibility that artificial intelligence (AI) will eventually outstrip human intelligence, perhaps greatly so.
• This may happen in the lifetimes of our children and grandchildren. Its impacts, for better or worse, are likely to be immense.
• Our best prospect of ensuring that this development is beneficial is to tackle it cooperatively, and as early as reasonable foresight allows.
• The Government can play an important role in fostering the academic, technological and policy-level coordination this process is likely to require, both nationally and internationally.

Introduction

1. I am a Co-Founder, with Baron Rees of Ludlow and Mr Jaan Tallinn (Skype), of the Centre for the Study of Existential Risk, Cambridge. I am also Director of the new Leverhulme Centre for the Future of Intelligence, which is to be based in Cambridge, with partners at Oxford, Imperial College, and UC Berkeley. In these roles, I have been involved in recent discussions about the long term future of AI. I am writing with respect to the second issue on which the Committee seeks submissions: "The use of AI for the public good." I focus exclusively on long-term issues, and offer my personal recommendation concerning the role that the Government can most usefully play in the short and medium terms, with its eye on the long term.

The prospect of superintelligence

2. I J Good was a Cambridge-trained mathematician, who worked with Alan Turing at Bletchley Park, and at Manchester after the War. In their free time, Good and Turing often talked about the future of machine intelligence. Both were convinced that machines would one day be smarter than us. In the 1960s, when Good emerged from a decade at GCHQ, he began to write about the topic.

3. In his first paper[1] Good tries to estimate the economic value of an ultra-intelligent machine. Looking for a benchmark for productive brainpower, he settles impishly on John Maynard Keynes. He notes that Keynes' value to the economy had been estimated at 100 thousand million British pounds, and suggests that the machine might be good for a million times that – a mega-Keynes, as he puts it.

4. But there's a catch. "The sign is uncertain" – in other words, it is not clear whether

this huge impact would be negative or positive: "The machines will create social problems, but they might also be able to solve them, in addition to those that have been created by microbes and men." Most of all, Good insists that these questions need serious thought: "These remarks might appear fanciful to some readers, but to me they seem real and urgent, and worthy of emphasis outside science fiction."

5. In one sense, the prospect that concerned Good remains the same, fifty years later. It boils down to four key points:

• We have no strong reason to think that high-level intelligence is less possible in non-biological hardware than it is in our skulls.

• We have no reason to suppose that "human level" marks an interesting limit, in non-biological systems, freed of constraints (e.g., of size, energy consumption, access to memory, and slow biochemical processing speeds) that apply in our own case.

• So we should take seriously the possibility that AI will reach places inaccessible to us – kinds and levels of intelligence that are difficult for us to understand or map, and likely to involve capabilities far beyond our own, in many of the tasks to which we apply our own intelligence.

• Finally, there's a prospect that AI systems will themselves contribute to improvements in AI technology, at some point. As Good saw clearly, there's then a potential for an exponential rate of development.

6. The big change since 1965 is that the incentives that will lead us in this direction are now much more obvious. We don't know how long the path to high-level machine intelligence is, but we can be certain that its individual steps will be of huge commercial value, and immensely important in other ways – for security purposes, for example. So we can be sure that these pressures will take us in that direction, by default. AI is already worth trillions of dollars – perhaps not yet a mega-Keynes, but well on the way.

7. At present, however, AI is very good at (some) narrowly-defined tasks, but lacks the generality of human intelligence. The term artificial general intelligence (AGI) is used to characterise a (hypothetical) machine that could perform any intellectual task that a human being can, including tasks not tied to specific set of goals. The term artifical superintelligence (ASI) refers to an AGI that greatly exceeds human capacities, in these respects.

When is AGI or ASI likely to arrive, and what would it mean for us?

8. A time-line for the development of AGI is difficult to predict, in part because it may depend on an unknown number of future conceptual advances. A recent survey of AI researchers reported that most regarded AGI as more likely than not, well within this century.[2] It does not seem alarmist to say that while it is not on our doorsteps, it may be only "decades away" (as a leading AI researcher puts it recently, intending to dispell the popular impression that it is just around the corner).[3]

9. Concerning the impact of AGI or ASI, we know little more than Good. It does not seem controversial that its impact is likely to be very big indeed. The world-leading AI researcher Professor Stuart Russell (UC Berkeley) is convinced that – for better or worse – it would be "the biggest event in human history."[4] But the sign is still uncertain, as Good put it. The potential benefits are immense, not least in the light of AGI's potential to solve many other problems. But there's also a risk. As Turing himself put it, "It seems probable that once the machine thinking method has started, it would not take long to outstrip our feeble powers. ... At some stage therefore we should have to expect the machines to take control."[5]

What can we do now?

10. These issues are going to be with us for a long time, and are likely to become more pressing as AI develops. In the short term, the obvious strategy is to attempt to foster the level of interest, expertise, and cooperation that the task is likely to require in the future. In effect, we should be trying to steer some of the best of human intelligence to the job of making the best of artificial intelligence. Most of all, in my view, we should avoid the mistake of putting off the issue to another decade or generation, on the grounds that it seems too hard or too much like science fiction, or because other issues in the same area simply seem more pressing.

11. There are encouraging recent signs of rapidly growing interest in these issues, for example in an open letter now signed by many AI professionals and others, following an international meeting in Puerto Rico in January 2015.[6] In particular, there is a growing sense of the desirability of cooperation between technology, policy, and academic partners. Some of this cooperation will necessarily be pre-competitive sharing, for commercial and other reasons – but all the more reason to engineer the kind of trust and cooperation that make such sharing possible.

12. The US is playing a leading role in this recent collaborative effort. It is home to the world's leading developers of artificial intelligence, such as Google and OpenAI, and amongst the most dynamic and impactful research teams.

What useful role is there for government, given the long time horizon?

13. The likely time-scale of these developments, and their dependence on ongoing research and progress in the field, makes a decisive intervention at one point in time impractical. More than in most cases, we are bound to be scanning a moving horizon. Nevertheless, there is a clear role for government that is likely to be beneficial, no matter how the field develops. It can foster, promote, and add its voice to a cooperative effort, both nationally and internationally, to monitor developments in the field, to flag opportunities and challenges as they arise, and generally to try to ensure the community of technologists,

academics and policy-makers is as well prepared as possible to deal with both.

14. What the government can most usefully add to this mix, in my view, is to maintain or adapt a standing body of some kind, to play a monitoring, consultative and coordinating role for the foreseeable future (and hopefully well beyond it). By ensuring from the beginning that the focus of this body is explicitly on long-term issues, there is an opportunity to lessen the risk that long-term issues will always be pushed aside in favour of short-term concerns.

July 2016

[1] 'Speculations concerning the first ultraintelligent machine'. In Advances in Computers, ed. F. L. Alt and M. Rubinoff (Academic Pres, 1965). http//bit.ly/IJGood1965
[2] Müller, V. and Bostrom, N., 'Future progress in artificial intelligence: A Survey of Expert Opinion', in Vincent C. Müller (ed.), Fundamental Issues of Artificial Intelligence (Synthese Library; Berlin: Springer), 2014. http://bit.ly/AIsurvey
[3] 'Brave new world? Sci-fi fears "hold back progress of AI", warns expert', The Guardian, 12.04.2016. http://bit.ly/Bishop2016
[4] 'This AI pioneer has a few concerns', Wired, 23.05.2015. http://bit.ly/Russell2015
[5] 'Intelligent machinery, a heretical theory', Manchester, 1951. http://bit.ly/Turing1951
[6] 'Open letter – research priorities for robust and beneficial artificial intelligence', Future of Life Institute, 2015. http://bit.ly/FLI-letter

Respondent 125

Karl Rauscher, Global Information Infrastructure Commission (GIIC)

Global Information Infrastructure Commission (GIIC) Input to
The White House Office of Science and Technology Policy (OSTP)
Request for Information on Preparing for the Future of Artificial Intelligence (AI)

1. The legal and governance implications of AI

Introducing AI into society will likely bring many shocks to the existing paradigms for how we govern ourselves. Indeed, allowing non-human machines to perform human functions, and often even super-human functions, has profound implications for existing policy structures. Existing policy structures have been built around a basic set of entities, namely governments, commercial or non-commercial organizations and individuals in a world

where machines have existed alongside humans for a long time. However, having an AI-enhanced smart machine replicate the functions of humans and make critical decisions and perform high-consequence tasks is not something legal and governance systems are prepared for at this time.

As governments, commercial enterprises and civil society prepare for the coming AI-enabled revolution, we would all do well to recognize the distinct mechanisms for influencing and being able to predict, the behavior of the assortment of entities that will be engaging in new ways. Some situations can be managed by two or more entities entering into a voluntary agreement where no government interference is required. Others situations will call for the need for a standard of reference so entities are speaking the same language so to say. Other times, a principle for smart machine conduct or a best practice for their operation may be what is needed, and can often be developed by industry consensus. The final mechanism is when government steps in to assert authority because some behaviors must be controlled. The key is that the last option should be considered just that, the last option, as it will typically be the most constraining.

In addition, the need for these issues to be addressed at an international level is necessary, given the technology supply chains and commercial markets that are anticipated to develop.

2. The use of AI for public good

AI offers very big upside potential for improvements in the dimensions of quality, speed and cost - key differentiators for many products and services. The public will benefit from better quality intelligence - just imagine being able to rely on a smart machine that will have access to the entire Internet and arrive at a decision in a split second. The same smart machine may perform menial tasks that free up enormous time and resources for their owners.

In order to achieve the desired benefits for as many people as possible, it is important that policymakers do not introduce unnecessary regulations that drive up the cost for people to purchase AI-based products and services. It is therefore critical that governments, in consultations with stakeholders, determine what is essential.

3. The safety and control issues for AI

The fact that AI and AI-enabled smart machines will be associated with unintended harm to people is an uncomfortable subject. Yet this is an unfortunate reality. A reality not unlike the introduction of other technologies that we have accepted, such as airplanes and automobiles.

As no technology will work perfectly, there is a need for benchmarks that identify acceptable performance standards. How will such benchmarks be established? As a starting point, it is reasonable that one may start by considering the safety benchmarks arrived at with societally-accepted uses of technology. Many industries have examples of regularly occurring failures that affect the safety of the public, such as car accidents in the transportation sector, success rates with technology-enhanced medical procedures in the healthcare sector, and network failures for Public Safety Answering Points (PSAP or 911) in the communications sector. However, given the promise and potential of AI to usher in improvements, it is not unreasonable to set higher standards, perhaps much higher standards, for the safety of AI-enable smart machines. The introduction of higher safety standards may also ease the acceptance rate for AI on a public that still looks on such technology with a degree of skepticism.

Advances in higher safety performance standards will be achieved with the technologies that AI enterprises will bring to the market. However, given the interplay of AI, AI-enabled smart machines and their connectivity fabric, collaboration across the industry may be needed. Such collaboration would be well served if government and the public could bring to the table its expectations for reasonable safety performance standards. Thus an existing industry forum, or a new one, should step forward and take on the challenge of addressing this important emerging concern for public safety. It seems certain much benefit could be gleaned if such a forum if it captured lessons learned from AI failures in such a way that countermeasures in the form of best practices could be developed and shared, promoting a culture of continuous improvement.

4. The social and economic implications of AI

The greatest social and economic benefits of AI can only extend to those that are online. This is because the most advanced, and therefore most useful, smart machines will be those connected to the AI resources in the cloud. However, according to International Telecommunications Union (ITU) statistics, more than 60% of the world's population is still not yet online. Thus the introduction of AI will further widen the economic impact of the digital divide. The new "AI divide" will have a compounding effect on economies relative to their position on the digital divide, i.e. more positive for those online and more negative for those offline.

As has been seen in places where the digital divide was overcome, reducing the cost to participate is often a key factor. Thus considerations should be given to policies that promote a competitive environment that benefits consumers on a global scale. Other barriers are policies that have a dampening effect on market growth. These include over-regulation.

It is therefore imperative that policymakers avoid causing unnecessary increased costs in the marketplace and unwarranted regulatory impediments.

5. The most pressing, fundamental questions in AI research, common to most or all scientific fields

In the not-to-distance future, AI is anticipated to enable smart machines to perform tasks of great consequence to humans. Examples include operating dangerous equipment, performing surgery, and making other life-affecting decisions. For some of these operations, connectivity to the cloud will be vital for the safety of those involved. Because the criticality and consequences of AI are so much higher than anything we have seen thus far, the security and reliability of networked devices must be at much higher levels than the current Internet experience. Ultra-high reliable and ultra-high secure connectivity fabric will be essential if the AI application dreams of many fields can be realized, including medical, transportation and services, to mention a few.

6. The most important research gaps in AI that must be addressed to advance this field and benefit the public

There is far more research on AI engines and the smart machines that will make use of AI, than on the practical ways these two types of components will be connected. Yet the fabric of connectivity will have critical influence - both in enabling and in limiting - key aspects for consumers, including interoperability, privacy and security. There are many stakeholders who have interest in seeing this fabric develop well, despite many of them not realizing it yet. Governments, commercial enterprises, civil society and others will have interests that need to be regarded as this fabric is envisioned and implemented.

Thus the question arises from this pressing concern: How can this fabric of connectivity be developed in a way that protects these and other interests? The development of a fabric that meets the needs of the many stakeholders requires the participation of the same. Given the discussion above, this should include international interests. It is therefore imperative that an existing industry forum, or a new one, demonstrate its ability to engage key stakeholders from every corner of the emerging AI landscape, and then step forward to facilitate the AI–to–smart machine connectivity fabric so the interests of stakeholders are best advanced going forward.

7. The scientific and technical training that will be needed to take advantage of harnessing the potential of AI technology

In order for AI-enabled smart machines to be able to take on more and more responsibilities in society, there is a critical need for improvements in the core competencies required to produce privacy, safety reliability and security across all ingredients of cyberspace (e.g., operating environment, electric power, hardware, software, networks, signaling and data payload, human interfaces and policies).

The observation that the supply chain for cyberspace is quite international suggests that such best practices would be best utilized if implemented on an international scale. Thus international collaboration will be a key part of such research endeavors.

8. The specific steps that could be taken by the federal government, research institutes, universities, and philanthropies to encourage multi-disciplinary AI research

Given its expertise and the drive of its commercial interests, the private sector is likely to be leading the AI revolution, as it has for other technologies. It is therefore imperative that government understands how to effectively engage the industry so that its interest, and the interest of the public good, are well represented

To this end, government should develop a clear and concise inventory of concerns and interests and bring the same to private sector-led initiatives that are shown to be effective in convening a well-balanced field of stakeholders. In such fora, the government should seek to identify gaps in the existing research and collaboration landscape where interests of the public good may be neglected.

9. Any additional information related to AI research or policymaking, not requested above, that you believe OSTP should consider

In assessing where to engage the rapidly developing AI research and development community, governments around the world should prioritize their focus on addressing those aspects of policy not addressable elsewhere. Having this mindset at the start is important as a practical matter, so that limited resources can be used most beneficially for all involved. Beyond this practical matter, this focus will also allow the few, important gaps to be filled that only governments can fill.

Respondent 126

Geoff Lane, Application Developers Alliance

Public Comments of the Application Developers Alliance on Preparing
for the Future of Artificial Intelligence

Background and Introduction

The Application Developers Alliance ("Alliance") is a global industry organization that includes nearly 200 companies and more than 60,000 individual software developers.

- Alliance corporate membership includes small and large app publishers, infrastructure and service providers, and software industry stakeholders.
- The individual developers in the Alliance network are the workforce of the future — creators and builders whose forward-thinking products and services are improving our world.
- All Alliance members are invested in a forward-looking digital future that embraces cutting-edge technologies. Innovation and data are the lifeblood for every software developer.

The Alliance was formed to promote and support the interests of developers as creators, entrepreneurs, and innovators. Developers build the apps and software that enable products and systems that in turn support consumers, power businesses, and connect industries. All of our members' activities are interconnected, and it is equally in all our members' interests to ensure technologies like artificial intelligence ("A.I.") are developed and leveraged to deliver new and innovative products and services to consumers. In this regard, the Alliance is pleased to comment on preparing for the future of artificial intelligence.

Artificial intelligence and machine learning are already increasing efficiencies, creating economic growth, and improving lives. To ensure that A.I.'s potential is realized, policymakers must acknowledge that A.I. is an extension of the digital industry that has thus far benefited consumers, and that established 'light-touch' regulation of burgeoning technologies enables innovation.

There is tremendous potential in artificial intelligence. Developers are creating new and exciting products and services that comply with existing regulatory regimes, will trigger job and economic growth, and most importantly, provide benefits to users.

Question #4: Existing and Potential Benefits Created by Artificial Intelligence

Artificial intelligence is already helping to create economic opportunity, and the futures of other important sectors are tied to its success. Its progress is being felt in the way of driverless cars that give increased mobility to seniors and people with disabilities, song and book recommendations that improve consumer experiences, and interactive education programs that help address students' specific needs. But A.I.'s benefits extend far beyond the products consumers are currently taking advantage of.

Artificial intelligence has a positive effect on jobs and economic growth. A 2016 report

found that by 2020, the artificial intelligence market will grow to over $5 billion (Markets and Markets, 2016). Further, robotics, which may use A.I., are a net positive for the job market. In five of the six countries studied, unemployment rates either fell or remained the same as robotics usage increased (International Federation of Robotics, 2011). The same study found that between two and three jobs were created as a direct result of each robot in use, leaving aside indirect job creation that will blossom as a result of increased competition and lower costs for consumers.

Increasingly, artificial intelligence is becoming a critical component to the growth of the very promising Internet of Things ("IoT") sector. By 2020, the IoT market will approach $2 trillion (Atlantic Council, 2016) and include as many as 200 billion connected devices (Intel, 2013). Each of these devices will collect data that will need to be interpreted and analyzed. A.I. integration into IoT will be important to ensure that the data from these devices is optimized to create new products or improve existing ones. And as the IoT market grows, so too will the demand for new software developers. It is believed that over the next four years 10 million software developers will be needed to meet the increasing demand for IoT products (Assay, 2016). Enhancing IoT through A.I. will lead to greater employment opportunities for software developers.

This progress is not only occurring at large organizations; many small businesses and individual developers, including Alliance members, are at the forefront of A.I. innovation. The new products and services these companies and developers are creating have taken root, and are providing societal benefits with the capacity to drive even more transformation.

Question #1: Current Laws and Regulations are Sufficient to Govern Artificial Intelligence

Policymakers should practice restraint, and ensure any new laws or regulations are measured and the result of stakeholder collaboration. Prematurely enacting burdensome regulations or laws will stifle promising technologies before they fully develop, fall heaviest on small- or medium-sized enterprises or new market entrants, and enable only the largest and most well-established innovators to compete in the space. Additionally, onerous policies that curb A.I. research, development, and deployment may motivate domestic innovators to move their operations abroad in search of more favorable ecosystems.

The United States has long been the world's innovative hub, thanks in part to the government's 'light-touch' regulatory approach. Existing regulatory frameworks for A.I. technologies are sufficient, balancing innovation and consumer protection. A.I. systems are new iterations of products and services users have long taken advantage of, and as such, A.I. systems already comply with consumer protection regulations. For example, healthcare systems incorporating A.I. must be HIPAA compliant, and autonomous vehicles must comply with automobile safety regulations.

The consequences of restrictive A.I. and machine learning regulations are already being felt in other parts of the world. Though it has yet to take effect, the EU's 'General Data Protection Regulation,' which will restrict "automated individual decision-making," is already threatening innovation in Europe (Metz, 2016). This vague and overly-broad regulation strikes at the heart of the digital future, and threatens online recommendation engines, early detection credit card fraud software, national security analyses, individualized learning programs, and so much more.

Any laws or regulations relating to artificial intelligence should mirror the 'light-touch' approach that has allowed innovation to flourish, must take into account the challenges of regulating a burgeoning technology, and be sensitive to additional compliance burdens placed on small- and medium-sized enterprises. Policies that limit A.I. growth will slow innovation and economic growth, and ultimately jeopardize A.I.'s important consumer benefits.

Question #9: Open and Collaborative Approaches Help to Grow Artificial Intelligence

Technology companies are working together to open-source and collaborate on artificial intelligence projects. Some of the many examples of industry leaders collaborating to accelerate the growth of artificial intelligence include:

- Google's TensorFlow, an open-source software library for machine learning (Bohn, 2015).
- Yahoo's CaffeOnSpark, open-sourced for distributed deep learning on big data clusters (Novet, 2016).
- IBM's open-sourced SystemML, a large-scale machine learning project (Wheatley, 2015).
- Facebook's use of Torch, an open-source development environment, for its machine learning and A.I. research (Yegulalp, 2016).

By sharing tools, methods, and research, developers are enabled to stimulate growth and bring innovative products and services to market faster. Open-sourcing also helps to enhance opportunities for resource-constrained small- and medium-sized enterprises participating in A.I. development. Industry's continued open and collaborative approach to A.I. demonstrates its commitment to technological advancement, and companies should be applauded for their efforts to work together.

Conclusion

While still developing, artificial intelligence is already benefiting users and the economy. A.I. continues to make us more efficient, propel the development and deployment of innovative products and services, provide new employment opportunities, and contribute to the overall health of our economy. As the government considers what, if any, role it will play in

A.I., it is critical that it continues to enable 'permission-less' innovation, and carefully consider the effects new laws and regulations in the space will have on consumers, innovators, and the economy.

The Alliance is available to the White House Office of Science and Technology Policy or any federal agency to discuss this submission, or any other matter of interest to our industry. Thank you for the opportunity to make this submission.

Respectfully submitted,

Geoff Lane
Director, U.S. Policy and Government Relations
Application Developers Alliance
1015 7th Street NW, 2nd Floor
Washington, D.C. 20001

Endnotes

"A Guide to the Internet of Things." Intel, IDC, United Nations. October 2013.

"Artificial Intelligence (AI) Market by Technology (Machine Learning, Natural Language Processing (NLP), Image Processing, and Speech Recognition), Application & Geography - Global Forecast to 2020." Markets and Markets. February 2016.

Asay, Matt. "Why 10 Million Developers are Lining Up for the Internet of Things." TechRepublic. January 29, 2016.

Bohn, Dieter. "Google Is Giving Away a Big Part of Its Machine Learning Software." The Verge, November 9, 2015.

Lindsay, Greg; Beau Woods; and Joshua Corman. "Smart Homes and the Internet of Things." Atlantic Council: Brent Scowcroft Center on International Security. March 2016.

Metz, Cade. "Artificial Intelligence Is Setting Up the Internet for a Huge Clash with Europe." Wired Business. July 11, 2016.

Novet, Jordan. "Yahoo Open-Sources CaffeOnSpark Deep Learning Framework for Hadoop." VentureBeat. February 24, 2016.

"Positive Impact of Industrial Robots on Employment." International Federation of Robotics. February 21, 2011.

Wheatley, Mike. "IBM Open-Sources Its SystemML Machine Learning Tech." Silicon Angle, November 24, 2015.

Yegulalp, Serdar. "Facebook Open-Sources Its Machine Learning Magic." InfoWorld. January 16, 2016.

Respondent 127

Frank Pasquale, Professor of Law, University of Maryland

I have 5 main points:

1) Balancing Complementary and Substitutive AI

We have always shaped technology through law, and will continue to do so. The question is not whether, but how. Continuing to substitute AI for human work could reproduce much of the economy we now have, more cheaply. But a better aim is to create a better world—and that will take a commitment to enabling future human-machine cooperation, with escalating skill levels and autonomy for workers. Bigger and better data sets should mean that two crucial two roles--applying expertise and developing it—should be integrated in more settings.

The distinction between substitutive AI (which replaces human labor with software or robots) and complementary AI (which deploys technology to assist, accelerate, or improve humans' work) is critical. In my talk at AI Now at NYU, I discussed three cases in health care where complementary automation ought to be preferred:

--where it produces better outcomes;
--in sensitive areas like targeting persons for mental health interventions;
--and where it can improve data gathering (by, for example, complementing computerized data entry with scribes).

Law and policy (ranging from licensure rules to reimbursement regulations) could help assure that the health care sector pursued complementary automation where appropriate, rather than chasing the well-hyped narrative of robot doctors and nurses.

There is a rival vision, commonly rooted in "disruption theory." According to this view, high technology competitors should replace established firms and providers by first developing cheap, poor quality products for the bottom end of the market, and gradually improving

quality. While such an approach may work for consumer items, policymakers should be wary of promoting it in professions like medicine, education, and law, lest they exacerbate extant inequalities.

Even elementary medical apps can fail patients. The FTC has settled lawsuits against firms who claimed their software could aid in the detection of skin cancer by evaluating photographs of the user's moles. The FTC argued that there was insufficient evidence to support such claims. The companies were prohibited from making any "health or disease claims" about the impact of the apps on the health of users unless they provide "reliable scientific evidence" grounded in clinical tests. If algorithms designed merely to inform patients about their options aren't ready for prime time, why presume diagnostic robots are imminent?

As health records are digitized and more genomic information becomes available, teams of doctors and informaticists at "learning health care systems" (LHCSs) are fundamentally reshaping the way we think about complex diseases. A hospital might seek to identify and apply the "standard of care" to patients. An LHCS aims to develop personalized comparisons of treatment effectiveness, using records of past interventions to determine what worked bests for patients of similar age, sex, genetic makeup, and other variables. Those seeking treatment are both patients (in the present) and a new kind of research subject (helping future doctors learn what, out of a range of good treatments, is optimal).

2) Taking Data Collection Seriously

We might once have categorized a melanoma simply as a type of "skin cancer." But that is beginning to seem as outdated as calling pneumonia, bronchitis, and hay fever "cough." Personalized medicine will help more oncologists gain a more sophisticated understanding of a given cancer as, say, one of a number of BRAF-omas (referring to the exact genetic abnormality that helped cause it). Digitized records (first from individual hospitals, then health systems, and finally regional and national interoperable databases) will help indicate which combination of chemotherapy, radioimmunotherapy, surgery, and radiation has had the best results.

For those who dream of a "SuperWatson" moving from conquering Jeopardy to running hospitals, each of these advances may seem like steps toward cookbook medicine, implemented by machine. And who knows what's in the offing 80 years hence? In our lifetime, what matters is how all these data streams are integrated, how much effort is put into that aim, how the participants are compensated, and who has access to the results. These are all difficult questions, but no one should doubt that juggling all the data will take skilled and careful human intervention. Insuring such training for professionals generally should be part of graduate schools' agenda.

To dig a bit deeper in radiology: the lighting of bodily tissue is rapidly advancing. We've

seen the advances from x-rays and Doppler scans to single-photon emission computed tomography (SPECT) and Positron Emission Tomography (PET) scans. Now scientists are developing ever more ways of imaging the inside of the body. There are already ingestible pill-cams; imagine injectable versions of the same. The watching need not be invasive: new ways of sensing heat, changes in chemical or biological composition, and much more are both sensing data and better visualized over time.

The resulting data streams are far richer than what came before. Integrating them into a judgment, about how to tweak or change entirely patterns of treatment, will take creative, unsystematizable thought. As James Thrall has argued, "the data in our EMR, PACS, and radiology information system databases are "dumb" data. The data are typically accessed one image or one fact at a time, and it is left to the individual user to integrate the data and extract conceptual or operational value from them. The focus of the next 20 years will be turning dumb data from large and disparate data sources into knowledge and also using the ability to rapidly mobilize and analyze data to improve the efficiency of our work processes."

Richer results from the lab, new and better forms of imaging, genetic analysis, and other sources will need to be integrated into a coherent picture of a patient's state of illness. In Simon Head's thoughtful distinction, it will be a matter of practice, not predetermined process, to optimize medical responses to the new volumes and varieties of data. Both diagnostic and interventional radiologists will need to take up each case anew, not as a simple sorting exercise.

Moreover, there is important work to be done among health record keepers as well. In reliable big data science, researchers invest a great deal of time and effort in cleaning up and integrating data, assuring that it is actually accurate and verifiable. In medical contexts where lives are at stake, the case for assuring data integrity and creatively, carefully integrating data streams applies a fortiori.

3) Professionalizing and Better Valuing Care Work

Journalists frequently cite a 2013 paper from Oxford academics Carl Benedikt Frey and Michael Osborne to predict the imminent "disruption" of many jobs. They believe computerisation could "make nearly half of jobs redundant within 10 to 20 years." To the extent health workers are presently doing rather simple tasks, computation may well replace them. But we should also ask why some tasks are deemed "simple," too. In the case of home health care workers, if we define the role as simply cooking, bathing, and cleaning bedpans, this job (one of the fastest-growing occupational categories in the US) may well be robotizable. But if we closely observe growing evidence that isolation is a critical social determinant of health (overwhelming many other risk factors in predicting future morbidity and mortality), we may begin to professionalize aides to work on delicate, complex tasks of improving psychosocial well-being.

A pattern of cheap purchasing mandates leads to services that include only adequately helpful responses to elderly or disabled persons' situations. That, of course, is easy to automate. But if caregivers' roles were to include the suite of competences discussed by Robert Kuttner in his proposal to professionalize the service professions, replacing them with a Roomba, Baxter, and "Paro" robotic seal would be far less thinkable than it is now. And one need only read sensitive accounts of good hospice care to realize that, in so many areas of care work, complementary rather than substitutive automation will be critical. Indeed, we would do well to follow Lucy Suchman's advice to question whether the term "robotic carer" itself is an oxymoron.

As Frey & Osborne observe, there are critical barriers to successful automation: creative intelligence, and the social intelligence involved in negotiation, persuasion and care, are very hard to program. Each of those capacities will be in high demand in the learning health care systems and elder care of the future—if patients have the resources to demand them.

4) Harmonizing Macroeconomic Approaches to Automation and Health and Education Reform

This question of resources leads to a politico-economic point: the importance of harmonizing macroeconomic approaches to key sectors, and US automation policy.

There is a troubling tension at the heart of US labor policy on health care and automation. Numerous high-level officials express grave concerns about the "rise of the robots," since software is taking over more jobs once done by humans. They also tend to lament growth in health care jobs as a problem. In an economy where automation is pervasive, one would think they would be thankful for new positions at hospitals, nursing homes, and EHR vendors. But they remain conflicted, anxious about maintaining some arbitrary cap on health spending.

As Princeton/NYU economist William J. Baumol observed in his 2012 book The Cost Disease, arbitrary caps on health spending are unwise for a country with the GDP of the US. The aging of the baby boomers will create extraordinary demand for care. This is hard work that society should fairly compensate. At the same time, automation threatens to replace millions of extant jobs for those making less than $20 an hour, especially in transportation and logistics.

The situation suggests a natural match: between distressed or underemployed workers (now being replaced by self-driving cars, self-check-out kiosks, and other robotics), and emerging jobs in the health sector (for home health aides, health coaches, hospice nurses, and many other positions). Those jobs in health care can only emerge if policymakers value the hard work now done (and remaining to be done) for the sick and disabled.

As Princeton/NYU economist William J. Baumol observed in 2012: "[I]f improvements to health care . . . are hindered by the illusion that we cannot afford them, we will all be forced

to suffer from self-inflicted wounds. The very definition of rising productivity ensures that the future will offer us a cornucopia of desirable services and abundant products. The main threat to this happy prospect is the illusion that society cannot afford them, with resulting political developments–such as calls for reduced governmental revenues entwined with demands that budgets always be in balance–that deny these benefits to our descendants."

Some health economists contribute to the "illusion that we cannot afford" progress by smuggling an ideology of austerity into ostensibly neutral discussions about the size of the health care sector. Before proposing to cut more from the sector, they need to offer a positive industrial policy on where health spending should be going—and how cutting funds for medical care will lead to better spending elsewhere.

5) More AI? Or Better AI?

We do not simply need "more AI;" we need better AI. Retarding automation that controls, stigmatizes, and cheats innocent people is a vital role for 21st century regulators, as I show in my book The Black Box Society. We should also stop arms races with zero productive gains, especially in the military and finance sectors. Our future quality of life will hinge on dynamics barely remarked in contemporary debates on robotics. In what sectors will automation take on the character of an arms race, where one side's investment in better software provokes its competitors to try to invest more?

Automation policy must be built on twin foundations: making the increasingly algorithmic processes behind our daily experience accountable, and limiting zero-sum arms races in automation. While this may seem commonsensical, it will actually require us to embrace something I call the "paradox of cost"—that certain productive sectors of the economy should actually take a growing share of GDP, over time, rather than constantly respond to demands for cost-cutting. When a productive sector of the economy costs more, that can actually be a net gain--especially if it diverts resources away from another sector prone to provoking unproductive arms race.

It will not be cheap to integrate artificial intelligence into our economy in a way that enhances the value (and work experience) of professionals like teachers, doctors, engineers, and nurses. Ideally, rich societies would fund those endeavors by taxing or otherwise cutting back on the "arms race" automation so prominent in the finance and military sectors. As workers grapple with new forms of advice and support based on software and robots, they deserve laws and policies designed to democratize opportunities for productivity, autonomy, and professional status.

Select References

Nicholas Carr, The Glass Cage (2013).

Steven E. Dilsizian & Eliot L. Siegel, Artificial Intelligence in Medicine and Cardiac Imaging: Harnessing Big Data and Advanced Computing to Provide Personalized Medical Diagnosis and Treatment, 16 CURRENT CARDIOLOGY REP. 441, 445 (2014).

Eisenberg and Price, Promoting Health Care Innovation on the Demand Side, at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2766707.

Simon Head, Mindless: Why Smarter Machines are Making Dumber Humans.

Sharona Hoffman, Big, Bad Data, at https://www.aslme.org/media/downloadable/files/links/j/l/jlme-41_1-hoffman-supp.pdf.

For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights http://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html

Frank Pasquale, The University of Nowhere and the False Promise of Disruption, at https://lareviewofbooks.org/article/the-university-of-nowhere-the-false-promise-of-disruption#!

Frank Pasquale, Automating the Professions?, at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2775397

Frank Pasquale, To Replace or Respect, at https://www.boundary2.org/2015/01/to-replace-or-respect-futurology-as-if-people-mattered/

Frank Pasquale, The Black Box Society (2015).
Frank Pasquale, The Hidden Costs of Health Care Cost Cutting (2014), at http://digitalcommons.law.umaryland.edu/cgi/viewcontent.cgi?article=2516&context=fac_pubs.

James Thrall, Data Mining, Knowledge Creation, and Work Process Enhancement in the Second Generation of Radiology's Digital Age.

Respondent 128

Mark Lee, N/A

Realizing Ada Lovelace's dream of human-machine symbiosis with
"Symbiotic Genius: the fusion of Human Imagination with Machine Intelligence"

This document identifies the most important research gaps in AI (item 6 of the RFI), and proposes a solution that can benefit the public.

Artificial Intelligence (AI) is the realization of Alan Turing's approach towards the relationship between humans and machines, of "machines that can think on their own, that can learn and do everything that the human mind can do." However, over the next few decades, AI can drive widespread technological unemployment that leads to stark wealth disparities between the elite and the rest. (1)

Overlooked is the contrarian Ada Lovelace approach where "machines will never truly think, and that humans will always provide the creativity and intentionality." The goal of Lovelace's approach is "a partnership between humans and machines, a symbiosis where each side does what it does best. Machines augment rather than replicate and replace human intelligence." (2)

This document explores how the US can develop Lovelace's contrarian vision of human-machine symbiosis by fusing human imagination with machine intelligence. The document will explore:

(a) the minimal correlation between creativity and intelligence
(b) the evolution of genius: from solo to group, and finally to symbiotic genius (fusion of human imagination with machine intelligence)
(c) the mathematics of ideas
(d) the funding of a US creativity research initiative driven by the "mathematics of ideas"
(e) the applications of the mathematics of ideas
(f) symbiotic genius's drive to pervasive prosperity

(a) The minimal correlation between creativity and intelligence

Psychology research highlights that human intelligence and human creativity are distinct human abilities. Intelligence is largely inheritable - from 40% before entering elementary school to 80% by mid-adulthood. (3)

Creativity is less inheritable. Studies of identical twins reveal that only about 25% to 40% of creativity stem from genetics. (4)

A 2013 National Institutes of Health study reports "Meta-analytic findings suggest that the correlation between creative potential and intelligence generally is around r = .20" (5) Thus scientific evidence gleaned from multiple studies conclude that creativity and intelligence are distinct capabilities.

However creativity does decline over time. Sir Ken Robinson reports of a NASA study that shows 98% of a 3 to 5 year-old cohort were measured as being creative; five years later, only 32% of the same cohort were creative, and another five years later, only 10% of that

cohort were measured as being creative. Only 2% of young adults were measured as being creative. (6)

A paper released at the 2016 International Society of Intelligence Research indicates " imagination occupies a construct space that is relatively independent from cognitive ability and is more closely associated with personality," which frees people to re-awaken their creativity, which were gradually suppressed during their childhood (7)

(b) the evolution of genius: from solo to group, and finally to symbiotic genius
(fusion of human imagination with machine intelligence)

The term genius has traditionally been associated with a solo genius such as Albert Einstein who discovered the laws of relativity essentially by himself, albeit with the intermittent mathematical assistance of his associates.

Over time, group genius has emerged in which collaboration between humans of complementary abilities drives creativity; such was the discovery of DNA with the complementary skills of Watson and Crick.

Steve Jobs believed that creativity is "connecting things".

With solo genius, creativity abounds from the mental connections formed within one person. Group genius emerges from "idea connections" formed between two humans.

Walter Isaacson writes in his book "The Innovators" that Lovelace "glimpse a future in which machines would become partners of the human imagination".

Symbiotic genius, the fusion of human imagination with machine intelligence, emerges from the "idea connections" formed between humans and machines.

The only common medium that humans and machines share is mathematics. If the message is "ideas", and the medium is "mathematics", then symbiotic genius requires the development of the "mathematics of ideas".

(c) Mathematics of ideas

It is by logic that we prove, but by intuition that we discover.
Henri Poincare

Creativity drives mathematics forwards. Can mathematics drive creativity, and (scientific) innovation, which leads to economic growth?

Mathematics has become the language of science, and the mathematization of science drove

the Scientific Revolution. In his book 'The Invention of Science: A New History of the Scientific Revolution", York University (UK) professor David Wootton writes that "a revolution in ideas requires a revolution in language". To revolutionize creativity, we need a new medium for creativity; not words expressed through language, but mathematics.

2002 Fields Medalist Vladimir Voevodsky is attempting to reset the foundations of mathematics. He believes that his Univalent Foundations initiative will allow computers to verify mathematical theorem proving.

Scientific American writes about Voevodsky at the Heidelberg Laureate Forum (8)

"Voevodsky told mathematicians ... they're going to find themselves doing mathematics at the computer, with the aid of computer proof assistants. Soon, they won't consider a theorem proven until a computer has verified it. Soon, they'll be able to collaborate freely, even with mathematicians whose skills they don't have confidence in. And soon, they'll understand the foundations of mathematics very differently."

The US "Committee on the Mathematical Sciences in 2025; Board on Mathematical Sciences and Their Applications" defined a mathematical structure as "a mental construct that satisfies a collection of explicit formal rules on which mathematical reasoning can be carried out." (9)
This definition of a "mathematical structure" can be modified to define an idea: "a mental construct on which reasoning can be carried out." Thus ideas are a super-set of mathematical structures.

The downward Lowenheim-Skolem mathematical theorem states that statements expressed via a language cannot be more complex than the language's complexity.

Consider language as a container, in which its contents cannot be more complex than the container's (i.e. language) carrying capacity.

Ideas are traditionally expressed via language, which are conveyed through 2-dimensional media such as paper. Thus ideas conveyed through languages cannot exceed 2 dimensions.

Fortunately, languages can be considered as a subset of mathematical structures, which can be multi-dimensional, starting from 2 dimension onwards. Complex problems are multi-dimensional; thus the questions should be expressed in multi-dimensional mathematical structures, with corresponding multi-dimensional solutions.

(d) the funding of a US creativity research initiative driven by the "mathematics of ideas"

While AI is attracting hundreds of millions of research funding, creativity is greatly overlooked. Immediately after Google subsidiary DeepMind Alpha Go's stunning 4-1 Go

victory over South Korean Go grandmaster (9th Dan) Lee Sedol, the government of Korea announced a US$863 million investment in artificial intelligence.

While hundreds of millions of dollars flood into AI research, little has been dedicated to creativity research, even though creativity drives innovation and economic growth. A 2014 study lamented the dearth of creativity research funding by the US government:

"The amount of research money spent on creativity studies was only 2.1% and 1.3% of the total in government funding provided by the Department of Education and National Science Foundation in the United States of America when compared with studies on academic achievement, self-concept, memory, critical thinking, motivation, and intelligence. " (10)

While the EU's 2013-2016 EU $43 million euros' Prosecco initiative recognizes the difference between the two distinct mental capabilities of intelligence versus creativity, it focus primarily upon developing computer-based equivalent of creativity, thus leading to "computational creativity". http://prosecco-network.eu/
Prosecco overlooks the symbiotic possiblities of human imagination with machine intelligence.

The National Science Foundation is funding a joint $25M MIT-Harvard Center for Brains, Minds + Machines research initiative. Its complement should be a $25M research initiative that explores and discovers the "mathematics of ideas".

(e) the applications of the mathematics of ideas

Creativity professor Keith Sawyer (author of "Group Genius") originated a zig-zag process, which includes the following steps: ask, learn, look, play, think, fuse, choose and make. These steps are not necessarily sequential, and can loop back upon each other until successful completion. While his process are human-driven, the mathematics of ideas allow symbiotic fusion of human imagination with machine intelligence.

Human agency and imagination will lead the ask and learn steps, while machine intelligence will support the look, play, think and fuse steps. Human agency will choose and (if necessary) make the final results of the creativity and innovation process.

In physical ("atoms") reality, ideas are intangible. In the virtual (or augmented) reality, ideas can be made tangible, and subject to virtual manipulation.

Mathematical structures can be represented as "idea blocks", which are the "ideas" equivalent of lego blocks. In virtual reality, digital natives can easily manipulate and make new ideas using "idea blocks". Minecraft experience will allow younger workers to rapidly understand and apply "idea blocks" in their works tasks.

Science challenges such as protein-folding and quantum physics have been gamified to allow the general public to solve these challenges. However these challenges require custom approaches that are not easily transferable to other challenges.

As the language of science, mathematics provide a common language across multiple scientific disciplines. Thus the "mathematics of ideas" provides a common platform, i.e. a general purpose technology to allow for rapid and easy gamification of science problems across multiple scientific disciplines.

Exposing these science challenges as visual and spatial challenges formulated as idea blocks allow the general public to address these challenges.

Instead of merely using augmented reality for entertainment (a la Pokemon Go), the "mathematics of ideas" enable all to be creative via their manipulation of ideas rendered tangible in augmented/virtual reality.

A 2009 Vanderbilt University study found that "70% of the top 1% in spatial ability did not make the cut for the top 1% on either the math or the verbal composite". (11)

By analogy, 70% of the general population are better in visual and spatial skills than verbal and mathematical skills. By using "mathematics of ideas" to incorporate ideas as mathematical structures within virtual/augmented reality, then humans can apply their visual and spatial skills to manipulate and create new ideas.

(f) symbiotic genius's drive to pervasive prosperity

If at least 70% of the population who are stronger in visual and spatial skills can use the "idea blocks" enabled by the "mathematics of ideas" to explore, manipulate, and create new ideas, they can contribute far higher value-added creative tasks, and thus earn more.

Instead of being left stranded and unemployed by AI's growth, the fusion of human imagination with AI will allow them to be far more creative, which grants them greater prosperity. If AI grows in power, so too will the collective symbiotic genius of the American population grow in strength, and lead to pervasive prosperity.

Lovelace's symbiotic genius will ultimately enable human imagination to soar with the help of machine (intelligence)! As machine intelligence increases, human imagination will soar even higher. Therein lies the future of humankind!

References

(1) The Innovators : How a Group of Hackers, Geniuses, and Geeks Created the Digital Revolution - Walter Isaacson

(2) ibid
(3) https://www1.udel.edu/educ/gottfredson/reprints/2004socialconsequences.pdf
(4) http://hbswk.hbs.edu/item/five-discovery-skills-that-distinguish-great-innovators
(5) http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3682183/
(6) https://www.psychologytoday.com/blog/creative-synthesis/201203/be-more-creative-today
(7) http://www.isironline.org/wp-content/uploads/2016/07/ISIR-2016_full_programme.pdf
(8) http://blogs.scientificamerican.com/guest-blog/voevodskye28099s-mathematical-revolution/
(9) http://www.nap.edu/catalog/15269/the-mathematical-sciences-in-2025
(10) http://www.tandfonline.com/doi/abs/10.1080/10400419.2014.901100?journalCode=hcrj20
(11) https://my.vanderbilt.edu/smpy/files/2013/02/Wai2009SpatialAbility.pdf

Respondent 129

Mark Lee, N/A

Realizing Ada Lovelace's dream of human-machine symbiosis with
"Symbiotic Genius: the fusion of Human Imagination with Machine Intelligence"

Author: Mark Lee (XXXXXXXXX)
Date: July 22nd, 2016

This document identifies the most important research gaps in AI (item 6 of the RFI), and proposes a solution that can benefit the public.

Artificial Intelligence (AI) is the realization of Alan Turing's approach towards the relationship between humans and machines, of "machines that can think on their own, that can learn and do everything that the human mind can do." However, over the next few decades, AI can drive widespread technological unemployment that leads to stark wealth disparities between the elite and the rest. (1)

Overlooked is the contrarian Ada Lovelace approach where "machines will never truly think, and that humans will always provide the creativity and intentionality." The goal of Lovelace's approach is "a partnership between humans and machines, a symbiosis where each side does what it does best. Machines augment rather than replicate and replace human intelligence." (2)

This document explores how the US can develop Lovelace's contrarian vision of human-machine symbiosis by fusing human imagination with machine intelligence. The document

will explore:

(a) the minimal correlation between creativity and intelligence
(b) the evolution of genius: from solo to group, and finally to symbiotic genius (fusion of human imagination with machine intelligence)
(c) the mathematics of ideas
(d) the funding of a US creativity research initiative driven by the "mathematics of ideas"
(e) the applications of the mathematics of ideas
(f) symbiotic genius's drive to pervasive prosperity

(a) The minimal correlation between creativity and intelligence

Psychology research highlights that human intelligence and human creativity are distinct human abilities. Intelligence is largely inheritable - from 40% before entering elementary school to 80% by mid-adulthood. (3)

Creativity is less inheritable. Studies of identical twins reveal that only about 25% to 40% of creativity stem from genetics. (4)

A 2013 National Institutes of Health study reports "Meta-analytic findings suggest that the correlation between creative potential and intelligence generally is around r = .20" (5) Thus scientific evidence gleaned from multiple studies conclude that creativity and intelligence are distinct capabilities.

However creativity does decline over time. Sir Ken Robinson reports of a NASA study that shows 98% of a 3 to 5 year-old cohort were measured as being creative; five years later, only 32% of the same cohort were creative, and another five years later, only 10% of that cohort were measured as being creative. Only 2% of young adults were measured as being creative. (6)

A paper released at the 2016 International Society of Intelligence Research indicates " imagination occupies a construct space that is relatively independent from cognitive ability and is more closely associated with personality," which frees people to re-awaken their creativity, which were gradually suppressed during their childhood (7)

(b) the evolution of genius: from solo to group, and finally to symbiotic genius
(fusion of human imagination with machine intelligence)

The term genius has traditionally been associated with a solo genius such as Albert Einstein who discovered the laws of relativity essentially by himself, albeit with the intermittent mathematical assistance of his associates.

Over time, group genius has emerged in which collaboration between humans of complementary abilities drives creativity; such was the discovery of DNA with the complementary skills of Watson and Crick.

Steve Jobs believed that creativity is "connecting things".

With solo genius, creativity abounds from the mental connections formed within one person. Group genius emerges from "idea connections" formed between two humans.

Walter Isaacson writes in his book "The Innovators" that Lovelace "glimpse a future in which machines would become partners of the human imagination".

Symbiotic genius, the fusion of human imagination with machine intelligence, emerges from the "idea connections" formed between humans and machines.

The only common medium that humans and machines share is mathematics. If the message is "ideas", and the medium is "mathematics", then symbiotic genius requires the development of the "mathematics of ideas".

(c) Mathematics of ideas

It is by logic that we prove, but by intuition that we discover.
Henri Poincare

Creativity drives mathematics forwards. Can mathematics drive creativity, and (scientific) innovation, which leads to economic growth?

Mathematics has become the language of science, and the mathematization of science drove the Scientific Revolution. In his book 'The Invention of Science: A New History of the Scientific Revolution", York University (UK) professor David Wootton writes that "a revolution in ideas requires a revolution in language". To revolutionize creativity, we need a new medium for creativity; not words expressed through language, but mathematics.

2002 Fields Medalist Vladimir Voevodsky is attempting to reset the foundations of mathematics. He believes that his Univalent Foundations initiative will allow computers to verify mathematical theorem proving.

Scientific American writes about Voevodsky at the Heidelberg Laureate Forum (8)

"Voevodsky told mathematicians ... they're going to find themselves doing mathematics at the computer, with the aid of computer proof assistants. Soon, they won't consider a theorem proven until a computer has verified it. Soon, they'll be able to collaborate freely, even with mathematicians whose skills they don't have confidence in. And soon, they'll

understand the foundations of mathematics very differently."

The US "Committee on the Mathematical Sciences in 2025; Board on Mathematical Sciences and Their Applications" defined a mathematical structure as "a mental construct that satisfies a collection of explicit formal rules on which mathematical reasoning can be carried out." (9)
This definition of a "mathematical structure" can be modified to define an idea: "a mental construct on which reasoning can be carried out." Thus ideas are a super-set of mathematical structures.

The downward Lowenheim-Skolem mathematical theorem states that statements expressed via a language cannot be more complex than the language's complexity.

Consider language as a container, in which its contents cannot be more complex than the container's (i.e. language) carrying capacity.

Ideas are traditionally expressed via language, which are conveyed through 2-dimensional media such as paper. Thus ideas conveyed through languages cannot exceed 2 dimensions.

Fortunately, languages can be considered as a subset of mathematical structures, which can be multi-dimensional, starting from 2 dimension onwards. Complex problems are multi-dimensional; thus the questions should be expressed in multi-dimensional mathematical structures, with corresponding multi-dimensional solutions.

(d) the funding of a US creativity research initiative driven by the "mathematics of ideas"

While AI is attracting hundreds of millions of research funding, creativity is greatly overlooked. Immediately after Google subsidiary DeepMind Alpha Go's stunning 4-1 Go victory over South Korean Go grandmaster (9th Dan) Lee Sedol, the government of Korea announced a US$863 million investment in artificial intelligence.

While hundreds of millions of dollars flood into AI research, little has been dedicated to creativity research, even though creativity drives innovation and economic growth. A 2014 study lamented the dearth of creativity research funding by the US government:

"The amount of research money spent on creativity studies was only 2.1% and 1.3% of the total in government funding provided by the Department of Education and National Science Foundation in the United States of America when compared with studies on academic achievement, self-concept, memory, critical thinking, motivation, and intelligence. " (10)

While the EU's 2013-2016 EU $43 million euros' Prosecco initiative recognizes the difference between the two distinct mental capabilities of intelligence versus creativity, it focus primarily upon developing computer-based equivalent of creativity, thus leading to

"computational creativity". http://prosecco-network.eu/
Prosecco overlooks the symbiotic possiblities of human imagination with machine intelligence.

The National Science Foundation is funding a joint $25M MIT-Harvard Center for Brains, Minds + Machines research initiative. Its complement should be a $25M research initiative that explores and discovers the "mathematics of ideas".

(e) the applications of the mathematics of ideas

Creativity professor Keith Sawyer (author of "Group Genius") originated a zig-zag process, which includes the following steps: ask, learn, look, play, think, fuse, choose and make. These steps are not necessarily sequential, and can loop back upon each other until successful completion. While his process are human-driven, the mathematics of ideas allow symbiotic fusion of human imagination with machine intelligence.

Human agency and imagination will lead the ask and learn steps, while machine intelligence will support the look, play, think and fuse steps. Human agency will choose and (if necessary) make the final results of the creativity and innovation process.

In physical ("atoms") reality, ideas are intangible. In the virtual (or augmented) reality, ideas can be made tangible, and subject to virtual manipulation.

Mathematical structures can be represented as "idea blocks", which are the "ideas" equivalent of lego blocks. In virtual reality, digital natives can easily manipulate and make new ideas using "idea blocks". Minecraft experience will allow younger workers to rapidly understand and apply "idea blocks" in their works tasks.

Science challenges such as protein-folding and quantum physics have been gamified to allow the general public to solve these challenges. However these challenges require custom approaches that are not easily transferable to other challenges.

As the language of science, mathematics provide a common language across multiple scientific disciplines. Thus the "mathematics of ideas" provides a common platform, i.e. a general purpose technology to allow for rapid and easy gamification of science problems across multiple scientific disciplines.

Exposing these science challenges as visual and spatial challenges formulated as idea blocks allow the general public to address these challenges.

Instead of merely using augmented reality for entertainment (a la Pokemon Go), the "mathematics of ideas" enable all to be creative via their manipulation of ideas rendered tangible in augmented/virtual reality.

A 2009 Vanderbilt University study found that "70% of the top 1% in spatial ability did not make the cut for the top 1% on either the math or the verbal composite". (11)

By analogy, 70% of the general population are better in visual and spatial skills than verbal and mathematical skills. By using "mathematics of ideas" to incorporate ideas as mathematical structures within virtual/augmented reality, then humans can apply their visual and spatial skills to manipulate and create new ideas.

(f) symbiotic genius's drive to pervasive prosperity

If at least 70% of the population who are stronger in visual and spatial skills can use the "idea blocks" enabled by the "mathematics of ideas" to explore, manipulate, and create new ideas, they can contribute far higher value-added creative tasks, and thus earn more.

Instead of being left stranded and unemployed by AI's growth, the fusion of human imagination with AI will allow them to be far more creative, which grants them greater prosperity. If AI grows in power, so too will the collective symbiotic genius of the American population grow in strength, and lead to pervasive prosperity.

Lovelace's symbiotic genius will ultimately enable human imagination to soar with the help of machine (intelligence)! As machine intelligence increases, human imagination will soar even higher. Therein lies the future of humankind!

References

(1) The Innovators : How a Group of Hackers, Geniuses, and Geeks Created the Digital Revolution - Walter Isaacson
(2) ibid
(3) https://www1.udel.edu/educ/gottfredson/reprints/2004socialconsequences.pdf
(4) http://hbswk.hbs.edu/item/five-discovery-skills-that-distinguish-great-innovators
(5) http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3682183/
(6) https://www.psychologytoday.com/blog/creative-synthesis/201203/be-more-creative-today
(7) http://www.isironline.org/wp-content/uploads/2016/07/ISIR-2016_full_programme.pdf
(8) http://blogs.scientificamerican.com/guest-blog/voevodskye28099s-mathematical-revolution/
(9) http://www.nap.edu/catalog/15269/the-mathematical-sciences-in-2025
(10) http://www.tandfonline.com/doi/abs/10.1080/10400419.2014.901100?journalCode=hcrj20
(11) https://my.vanderbilt.edu/smpy/files/2013/02/Wai2009SpatialAbility.pdf

Respondent 130

Kyle Bogosian, Tulane University

I am a senior with personal and professional interests in both government policy and emerging technology. I'll address two issues: AI ethics and AI safety, mostly relating to criteria 1-4 in the provided description.

Regarding ethics: I spoke with Andrew Moore and a few others at SafArtInt about the role of ethics in determining objective functions and constraints for AI systems. I am a reader and enthusiast of moral philosophy and I believe there are several issues from academic "ivory-tower" moral philosophy which are important to the AI community.

The first is that moral disagreement is surmountable. Many people, both individuals I spoke to at SafArtInt as well as researchers (Bello and Bringsjord 2013, Shulman et al 2009) believe that moral disagreement between philosophers is a paralyzing problem for those of us who wish to build moral systems in AIs. However, we can still be justified in building AI systems grounded in classical ethics. The reason for this is that methods for comparing and discriminating between moral theories have been developed. Lokhorst 2011 discusses a logic-based method of having AIs switch between different moral systems based on the appropriate context, and MacAskill 2014 provides a thorough foundation of uncertainty and comparison between moral claims. In particular, regulating AI systems with a moral uncertainty framework would let us take into account the disagreements we have over ethical matters while providing reasonable and balanced outputs. It is worth noting that, as discussed at SafArtInt, one of the main ways of ensuring AI safety in non-moral domains consists of better providing models of uncertainty as well.

The second issue rises if someone asks why we don't just design AI morality as a copy or as a model of human thinking, as some (Bello and Brinsjord 2013) do, or why we don't just let NGOs, policymakers and politicians hammer things out rule by rule, as Andrew Moore implied. I won't tackle that issue directly here, but I will just note that there are large and systematic differences between the rules of classical morality and the beliefs of average people - they aren't coincidental methods of reaching the same conclusions. For instance, moral philosophers are much more likely than politicians and other voters to say that it is wrong to kill animals for food (http://leiterreports.typepad.com/blog/2012/10/philosophers-eating-ethics-a-discussion-of-the-poll-results.html, Alastair Norcross, Peter Singer, Christine Korsgaard). Other examples include philosophers being much more likely to take structural racism seriously, much less likely to believe in restricting immigration, and much less likely to derive moral rules from religious texts. I don't have sources for these sweeping generalizations because they are based on experience and discussion - but any philosopher or philosophy graduate student will likely tell you that they are correct. While I don't claim that philosophers are always in the right or vice versa, I simply wish to point out that the gap between the

experts' view on morality and the common view on morality is significant, so our AI systems' moral decisions will depend on which approach we choose to take. For this reason, we will have to have a serious discussion about the role of moral philosophy as a field in helping us with the design of AI ethical systems.

Regarding AI safety: I have read research papers written by computer scientists and physicists who are concerned about the long-term possibility of AI systems which recursively self-improve in order to maximize imperfectly designed objective functions. At SafArtInt, many people including Dr. Ed Felton were concerned that the popular fears and news media related to this issue could lead to an undesirable loss in support for AI research. I would like to make a plea for unity and cooperation on this sort of issue and would like to prevent "battle lines" from being drawn. Those who are interested in long term AI safety should not try to reduce AI funding and they should not claim that artificial general intelligence is just around the corner, although to my knowledge they generally do not do these things. Just like the White House, they are strong opponents of bad press and public misinformation, which is the main enemy of everyone involved. (Please note that I am not referring to popular culture figures like Elon Musk or Stephen Hawking - I'm referring to those who actually conduct professional research and analysis of long term AI safety.)

I believe the right approach for the AI community is to remain open to new ideas in AI forecasting and safety, and to calmly and rationally address issues scientifically and straightforwardly. It would be unfortunate if AI debates became reminiscent of climate change debates - rampant polarization, media sensationalism, and optimistic denialism should all be avoided. It's a difficult question - how do we analyze the issues of AGI and ASI without triggering popular paranoia and misinformation? The difficulty of doing so makes it all the more important for the OSTP and AI researchers to conduct honest, straightforward dialogue with organizations such as the Machine Intelligence Research Institute and the Future of Humanity Institute. (Note: I have no professional affiliation with the above organizations.)

Thank you for the conferences and RFI, I believe they were great ideas and I look forward to further activities in the future.

Respondent 131

David A. Heiner, Microsoft Corporation

Microsoft Response to Office of Science and Technology Policy
Request for Information: Preparing for the Future of Artificial Intelligence

David A. Heiner
Vice President and Deputy General Counsel, Regulatory Affairs
Microsoft Corporation

1. INTRODUCTION

Microsoft appreciates the opportunity to provide input to the White House Office of Science and Technology Policy (OSTP) Request for Information on artificial intelligence (AI). We applaud OSTP's leadership on its timely series of workshops to nurture public discussion on the opportunities and challenges ahead for AI, in a way that is holistic and inclusive of diverse perspectives. We appreciate the opportunities we have had to actively participate in these dialogues through our researchers in keynotes, panels, group discussions, and program committees, including leadership in organizing the workshop "AI Now: The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term".

"AI" is used to refer to a constellation of computing technologies that perform perception, learning, reasoning, and decision making, aimed at endowing machines with the ability to solve the kinds of problems now undertaken by humans. AI encompasses the sub-discipline of machine learning, where we have seen great strides in principles and applications over the last fifteen years, spurred by the growing availability of data, computational power, and innovative algorithms.

Microsoft has made deep investments in research and development on AI opportunities, and is a major contributor to the advancement of AI. The company is deeply enthusiastic about the promise of AI technologies to help understand and address some of society's greatest challenges.

We envision a promising future where machines augment and extend human abilities and experiences, empowering every individual to realize their full potential, thus enabling new socio-economic opportunities. This will present new challenges but, by working together and engaging on these challenges directly, industry, government, civil society, and the research community can help bring about a rewarding and positive future. We believe that the dialogue on these technologies is timely given the intensive efforts to develop and field applications more widely, including those used in governance and public policy, and those in safety-critical domains such as transportation and healthcare.

This response proposes three areas for further consideration in shaping and realizing this vision for AI. They span a number of questions in the RFI, corresponding most closely to numbers 2 through 6, and number 9 regarding general policy making:

- Development of shared AI principles;
- Select areas of research to support these principles;
- A collaborative approach to AI policy development.

2. SHAPING A HUMANISTIC APPROACH TO AI WITH PRINCIPLES

AI technology development is at an inflection point, where the vast amounts of data available, combined with computing power available in the cloud, has facilitated greater advances in machine learning as well as much broader uses of AI. The promise of AI is that the knowledge gained from applying analytics and machine learning to the great amount of data available will enhance any decision-making process with new insights and intelligence, leading to better outcomes.

Microsoft sees the development of usable technologies and development platforms as critical in the democratization and inclusiveness of AI advances. We are enthusiastic about the development of usable tools, languages, components, and platforms that empower people to harness the best technologies available.

However, we understand that there are many who are concerned about the economic disruptions that may come with the fast-paced automation and the displacement of different kinds of jobs. Such disruptions could initially most impact those who are struggling to survive. We also understand and share concerns that AI technologies could amplify and entrench biases that already exist in society, or may create new biases, based on the use of biased data sets and algorithms. We are aware of concerns relating to the unsupervised and indiscriminate development and application of AI technologies, and how such advances might place great power in the hands of a few, exacerbating power asymmetries between individuals and large organizations.

We believe it will be useful to continue to invest in better understanding the challenges and to collaborate with other organizations to address concerns and to develop and share best practices. And it will be best to build a collaborative vision for AI, so that as machines play a greater role in human work and decision-making, we can achieve greater societal progress and equality than ever before.

We believe that there is a strong future ahead in designing AI technologies to augment human capabilities and experiences. We have focused a great deal of research and development on advances in this area, including methods aimed at augmenting human cognition, via building systems with knowledge about human goals. Importantly, this also includes research into the limitations of human cognition in the realms of attention, memory, and judgment. We have also invested in methods that enable fluidity in coordinating a mix of human and machine contributions, and in a constellation of "complementary computing" methods, where machines reason about how they can best complement human problem solving in collaborating with people.

It is important that AI technologies empowered to perform new kinds of automation and decision making be deemed trustworthy by individuals and society at large. People will expect AI systems and those who run them to be fair, transparent, privacy-protecting, secure and accountable. This vision of empowerment, ethics and inclusiveness should guide the future development of AI. It will be important that the technology be applied in a way

that allows for due process, particularly for systems that play a role in significant social institutions like criminal justice, education, health, employment and housing. As technologies play an increasing role in mediating people's lives online and offline, it is essential that appropriate design, economic, and social choices be made to ensure that those technologies are respectful and inclusive, and help society progress by empowering all peoples and organizations. We also believe that we will need to couple the computational power and learning capabilities of machines with the sensitivity and emotional intelligence of humans. Simply put, technologies should be people-centered by design.

This humanistic approach to AI can be realized if relevant stakeholders from industry, government, civil society, and the research community come together to collaborate on shared principles and ethical frameworks. We have published our reflections on what these may be in order to start this much needed dialogue. We believe that AI should:

1. Be designed to assist humanity
2. Be transparent
3. Maximize efficiencies without destroying the dignity of people
4. Be designed for privacy
5. Have algorithmic accountability so that humans can undo unintended harm
6. Guard against bias

Complementing the above are considerations for the humans who are developing, deploying, and using these technologies:

1. Empathy
2. Education (knowledge and skills)
3. Creativity
4. Judgment and accountability

A common vision, with shared principles, will enable us to shape the future of AI. This is an essential step that will require all of us to work together to design and realize the future that we desire.

3. SELECT FOCUSED RESEARCH AREAS FOR AI

Eric Horvitz, a Technical Fellow and managing director at Microsoft Research, presented at an exploratory technical workshop prior to the event co-hosted with Carnegie Mellon University on "Safety and Control for Artificial Intelligence." During his framing talk, Eric spoke to the challenges and opportunities ahead with harnessing AI in valuable ways while minimizing the likelihood and costs of failures. He highlighted the opportunity to develop new technologies and also formulated a set of best practices. He described the importance of working to design fail-safe systems – "devices or practices that, in the event of a failure, respond or result in a way that will cause no harm, or at least minimizes harm." As it is neither mathematically possible nor computationally practical to model all failure

scenarios, AI systems must make decisions under uncertainty and with incomplete information in certain situations. Research can develop methodologies for enabling robust responses in addressing the failures that are not yet modeled – the "known unknown" cases, as well as the more challenging "unknown unknown" cases. Examples of such approaches include evaluating the risk (to core values) of different outcomes, and selecting a more conservative outcome that has lower risks—even if this means giving up on some potential value on the upside; monitoring performance to identify inconsistencies or anomalies, and engaging people for help or taking fail-safe action; employing a portfolio of models on the same problem, and developing methods for addressing model incompleteness, via developing methods for grappling with "unknown unknowns." Significantly more research is required to further develop these and similar approaches.

Consider Horvitz's question on fail-safe responses to "unknown unknowns" within the context of human-machine collaboration, and one realizes that there are many open questions. Does a general theory of safety need to be developed? Should safety issues include considering socio-economic harms and inequalities? How can machine learning and inference be used to identify the intersection between human cognition and machine intelligence in a given situation, and help to coordinate more effective actions between human and machine for fail-safe responses? A simple illustrative example involves a human driving an autonomous car, where AI can help nudge the human to pay more attention when a treacherous stretch of road is coming up, or the technology is detecting that the human is not paying adequate attention to the road under crowded conditions. There is a rich spectrum of autonomy, and many open questions on how to detect, define and arrive at "optimal" mixtures of human-machine initiatives.

Additional research is needed to develop best practices for the safe and ethical deployment of AI, including for example:

- Phases of study, testing and reporting for rolling out new capabilities in safety-critical domains (akin to FDA clinical trials);
- Algorithmic accountability, e.g., disclosure and control of parameters on failure rates and tradeoffs; system self-monitoring and reporting;
- Due process for inference and action, including transparency, explainability and redress;
- Standard protocols for human-machine collaborations, including ethical frameworks, standards for keeping people informed, passing the baton of control, and for revealing to others the presence of autonomous systems (e.g., identifying cars currently under autonomous control on roads)
- Allowing open access and study of data sets and algorithms used in governance and public policy decision making.

Kate Crawford, the co-chair of the NYU/White House event AI Now and Principal Researcher at Microsoft Research, spoke to the need to address the significant social and economic implications of AI. An important set of discussions at the AI Now experts'

workshop focused on the need to identify social inequity issues with the use of AI, including detection of data biases, discovery of systemic inequality that is being reflected in the algorithms, and unintended consequences or outcomes that can cause increased risk to one group over another. The event highlighted four critical domains of social inequality, ethics, labor and health, and the participants included a Carnegie Mellon University professor who found in one study that women were less likely than men to be shown search ads for highly paid jobs, and the investigative journalist who concluded that a commercial algorithm widely used by state court systems to predict repeat offenders is biased against black defendants. More research is needed to help detect these issues during development and deployment, and flag them for further intervention.

There is a broad spectrum of opportunities in the use of AI for social good. For example, use of AI to augment and enrich the world for the visually impaired; develop a patient-centric health-care approach that can help reduce medical errors (the third most common cause of death in the US); and new approaches to criminal justice, such as reducing bias in arrests, prosecutions, sentencing and diverting those in need of medical care away from incarceration. There are also many new areas of research, such as infodemiology, where diverse streams of digital information are analyzed to inform public health and policy, or use of search logs as large-scale sensing systems for drug safety or early detection of other potential issues. Ongoing work in these areas should be supported, and creative new projects encouraged.

A humanistic approach to AI requires an appreciation and understanding of social and cultural behaviors that are traditionally the domain of the social sciences. Whereas computer scientists are normally more concerned with building technology that provides the fastest, most efficient, and most accurate solutions, social scientists and humanists are more concerned with the impact of these technologies on people, our relationships, our lives, and our cultures. Work on AI will require technical disciplines to collaborate closely with social and humanistic disciplines throughout the development and deployment process. In a human-machine collaboration model, many types of insights are required, and multi-disciplinary research and education will produce better systems, and should be strongly encouraged. This will also support the types of understanding needed to ensure fairness, accountability and ethical practices in AI.

These research questions are just emerging, concurrent with increasingly broad deployment of AI systems into everyday life. They merit significant focus, as well as research funding from the government, academia, industry, and others, and are crucial to building a foundation for sustainable and people-centered AI.

4. AN EVOLVING AND COLLABORATIVE APPROACH TO POLICY DEVELOPMENT

The White House series provided a holistic exploration of emerging issues with AI, and demonstrated the value of including diverse perspectives from multiple disciplines

(computer scientists, data scientists, sociologists, economists, ethicists, subject-area experts, and others) and multiple stakeholders (industry, government, civil society, researchers). This has been a unique and proactive approach to policy development, and especially effective for emerging technologies that are not yet well understood.

Key areas that have been raised included labor impacts, bias and discrimination, safety and controllability of the technology, accountability and due process, amongst others. As AI is still in a nascent stage of commercial and technological development, it is timely to raise challenging questions, so that they are recognized and can be addressed, early on, intentionally and collaboratively, before widespread deployment.

For policy development, the convening of these dialogues should continue so that stakeholders, including federal agencies, can interact and learn from each other, prioritize issues of societal importance, and more importantly, work together to track challenges and develop workable solutions as new issues emerge. An inclusive approach that continues to value multi-disciplinary and multi-stakeholder contributions and actions can motivate a more open and collaborative model to policy development that would be appropriate for adapting to rapidly evolving technologies going forward. They also facilitate development of more principle- and evidence-based policy frameworks that can lead to more meaningful regulations, where desirable outcomes that are aligned with the vision of a more humanistic AI are encouraged.

5. CONCLUSION

There is an opportunity for government to collaborate with industry, civil society, and the research community to shape a future where AI and human are working together, where machines are augmenting human abilities and experiences to address societal challenges. We can start by developing a shared set of principles to realize this vision. We can foster and encourage specific research areas to enable the development of supportive technologies and enhance equality of opportunity. The White House Workshop series on AI provided an excellent starting point. Moving forward we need an even more collaborative and inclusive policy development process to identify, prioritize, adapt, and respond quickly to emerging issues. AI has the potential to help create a better world centered on humanistic principles, and we should continue to work together to actively realize this future.

6. FOR FURTHER READING

1. Angwin, J. et al, “Machine Bias,” ProPublica, May 23, 2016.
2. Crawford, K., “Artificial Intelligence’s White Guy Problem,” The New York Times, June 25, 2016.
3. Datta, A., Tschantz, M.C., and Datta, A., “Automated Experiments on Ad Privacy Settings,” Proceedings on Privacy Enhancing Technologies. Volume 2015, Issue 1, pages 92-

112, ISSN 2299-0984, April 2015.
4. Dietterich, T.G., and Horvitz, E.J., "Rise of Concerns about AI: Reflections and Directions," Communications of the ACM, Vol. 58 No. 10, pages 38-40, 10.1145/2770869.
5. Horvitz, E., "Reflections on Safety and Artificial Intelligence," presentation at CMU-OSTP meeting on "Safety and Control for Artificial Intelligence," June 27, 2016 (http://bit.ly/29hMvYI).
6. Horvitz, E., "AI in Support of People and Society," presentation at OCTP-CCC-AAAI meeting on "Artificial Intelligence for Social Good," June 7, 2016 (http://bit.ly/2ajn1cM).
7. Nadella, S., "The Partnership of the Future," Slate, June 28, 2016.
8. Primer for "AI Now: The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term" at https://artificialintelligencenow.com/.

Respondent 132

Richard Mallah, Future of Life Institute

Future of Life Institute Response to the White House RFI on AI

NOTE: REVISED VERSION

We thank the OSTP for providing this opportunity for stakeholders input into the OSTP's thinking about, and planning for, the potentially large impact AI will have in the coming decades. The Future of Life Institute, with its mission of increasing the odds of a good long-term future of Humanity, has focused a great deal on AI, and how we should endeavor to keep it robust (doing what we want it to do) and beneficial.

Regarding (2) the use of AI for public good;
Our view is that in the short term AI, like other information technologies, will serve as a powerful tool that can be used by corporations, governments, organizations, and individuals to accomplish their goals. As AI increases in capability, it should provide ever-stronger levers to enhance human capability in diverse fields including scientific research, engineering, data analytics, strategy and planning, legal analysis, etc., etc. This could enable accomplishment of many widely desirable goals, for example curing the majority of diseases, finding mutually beneficial paths in geostrategic analyses, developing clean energy, and finding ways of safely stopping deleterious anthropogenic climate change. In the longer term, we may well cross a threshold in which AIs transition from being tools for humans to accomplish their goals to agents that accomplish goals furnished to them by humans. In this case, as is discussed below, it is quite crucial that these goals are indeed for "the public good" and that they are accomplished by AIs in a manner that is also consistent with the public good.

Regarding (1) the legal and governance implications of AI;
A long-term issue issue that some governments have begun to address is what, if any, legal

rights robots (or machine intelligences) should be accorded (see for example Prodham 2016). Our view is that (contrary to the safety considerations discussed below), such discussion is premature and that extreme caution should be taken in setting any legal precedents bestowing such rights.

Regarding (3) the safety and control issues for AI;
Historically, practitioners in mainstream AI have focused on improving AI's pure capacity: its modeling capacity and its possible range of actions. As it becomes more powerful, society should broaden this focus to include building a clear understanding of how to make AI not just good at what it does, but reliably serve good aims. Societally beneficial values alignment of AI is not automatic. As AI pioneer Stuart Russell explains, "No matter how excellently an algorithm maximizes, and no matter how accurate its model of the world, a machine's decisions may be ineffably stupid, in the eyes of an ordinary human, if its utility function is not well aligned with human values." (2015).
Since humans rely heavily on shared tacit knowledge when talking about their values, it seems likely that attempts to represent human values formally will often leave out significant portions of what we think is important. This is what the classic stories of the genie in the lantern, the sorcerer's apprentice, and Midas' touch address. Fulfilling the letter of a goal with something far afield from the spirit of the goal like this is known as "perverse instantiation" (Bostrom 2011). This can occur because the system's programming or training lacks some relevant dimensions in which observations can vary, but that we really care about (Russell 2014). These are easy to miss because they are typically taken for granted by people; even trying with a lot of effort and a lot of training data, people cannot reliably think of what they've forgotten to think about. Trying to simply patch an ethical theory of explicit directives, a deontology, like Asimov's Laws, with a fourth or fifth additional rule would serve only to delay the serious deviations from what we'd want and encourage the system to find the next cheapest path to what it's understood it needs to do. The complexity of these systems will exceed human understanding quickly, yet we will have efficiency pressures to be increasingly dependent on them, ceding control to these systems. It becomes increasingly difficult to specify a values-robust set of rules as the domain approaches an open world model, in underconstrained cyberphysical contexts, and as tasks and environments get more complex and the capacity or scalability of human oversight is exceeded. Robustness includes interpretability, transparency, and the ability to produce valid explanations of decisions. Many of the prerequisites and artifacts created for for verification of machine learning also help its interpretability. Recognition of distributional shift, confidence in a trained model given the online data distribution, is also a prerequisite. Scalable human oversight, where the optimal amount of salient information is presented to and queried from a human, is an unsolved and critical challenge, not only for training phases, but in online modes as well. See Amodei et al.
In various architectures, information about system control signals can leak into the data these systems are trained on, leading to unexpected impairment of control or function. While privileging control information can help in the short term, more robust approaches such as the scalable oversight of corrigibility, will be required with more powerful systems.

See references Russell, Dewey, and Tegmark (2015) and Taylor (2016) for research threads that need to be worked on to address these issues.

Regarding (4) the social and economic implications of AI;
We are concerned that too little rigorous research has been done on the potential implications of AI for economics and employment. Although there is considerable controversy, we regard as compelling the research by, e.g. Erik Brynjolfsson and Andrew McAfee (http://secondmachineage.com), and by Frey and Osborne (2013) that AI and autonomous systems may replace humans in a large fraction of current jobs, on a timescale that faster than new jobs can be created or workers retrained. Indeed this process may already be underway. In the longer term, it is quite possible (though very contentious) that advanced and ubiquitous AI leads to an economic structure in which full employment is not a sensible expectation because a large fraction of the populace simply does not have (nor can easily be given) skills of significant economic value. Like other economic transitions, AI has the potential for a dramatic increase in prosperity. However, previous economic transitions may be poor guidance as to how this transition should be managed, and we encourage research into the likely effects of AI on the economy as well as potential policies that can ensure that this impact is an overall good for the vast majority of people.

Regarding (5) the most pressing, fundamental questions in AI research, common to most or all scientific fields;
Quantification of confidence rather than just probability, accounting of causality rather than correlations, and interpretability at multiple levels will be necessary for AI, in nearly any domain, to be robust. (See e.g. Amodei et al.)

Regarding (6) the most important research gaps in AI that must be addressed to advance this field and benefit the public;
Creating advanced AI responsibly requires value alignment. Approaching this does not require spelling out those values upfront, but rather, should initially be oriented around making sure that given values are actually able to be propagated and utilized reliably. To prevent deviation from the intent of those values, each of these subfields requires much more research: abstract reasoning about superior agents, ambiguity identification, anomaly explanation, computational humility or non-self-centered world models, computational respect or safe exploration, computational sympathy, concept geometry, corrigibility or scalable control, feature identification, formal verification of machine learning models and AI systems, interpretability, logical uncertainty modeling, metareasoning, ontology identification/ refactoring/alignment, robust induction, security in learning source provenance, user modeling, and values modeling.

Regarding (7) the scientific and technical training that will be needed to take advantage of harnessing the potential of AI technology;
To be able to use advanced AI systems effectively, both those developing AI and those deploying AI will need to understand the role of not only professional ethics, but the nature

of leverage, how to think about how their systems might interact with their deployment environments in methodical worst-case analyses, and how to identify and articulate stakeholder values.

Regarding (8) the specific steps that could be taken by the federal government, research institutes, universities, and philanthropies to encourage multi-disciplinary AI research; Research institutes and academia need to do more research on the topics mentioned in the answer to (6). Philanthropies and research institutes should organize and channel funds to grants for the aforementioned research to maximize societally beneficial impact. The federal government and philanthropies should channel more funds to research institutes and academia for the aforementioned research. As funding for AI increases, funding for AI safety, robustness, and beneficence should similarly increase. We recommend that a minimum of 5% of AI funding be put toward ensuring robustness, interpretability, values alignment, and safety of AI systems.
Parties should recognize that if scientists and technologists are worried about losing what they perceive as a single race to the finish, they will have more incentives to cut corners on safety and control, which would obviate the benefits of technical research that enables careful scientists to avoid the very real risks. For the long term, we recommend policies that will encourage the designers of transformative AI systems to work together cooperatively, perhaps through multinational and multicorporate collaborations, in order to discourage race dynamics.

Regarding (9) any additional information related to AI research or policymaking, not requested above, that you believe OSTP should consider.
Having no international agreements on restricting autonomous weapons could easily lead to quickly-spiraling arms races of destabilizing new WMDs that other countries with fewer inhibitions could win. The U.S. should therefore support multilateral, global, or international agreements to keep humans in the loop. If such agreements are adopted, even if enforcement guarantees are necessarily weaker than with NBC weapons, the spiraling race dynamic could be averted.
FLI helped coordinate, and supports, an open letter (http://futureoflife.org/open-letter-autonomous-weapons/) calling for an international agreement precluding the development of offensive fully autonomous weapons, supported by a very large number of AI researchers and other thinkers.
Globally allowing fully autonomous weapons could undermine key U.S. strategic advantages. A close analog is cyberwarfare: the U.S. likely has a significantly greater capability than other countries, but the power imbalance is much smaller than for conventional military weapons, and for a country to develop a strong cyber warfare capability would be dramatically cheaper and faster than developing a conventional weaponry capability that could seriously threaten the U.S. Allowing the frequent multidirectional incursions of cyber warfare into the kinetic sphere would be detrimental for all.

References

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, Dan Mané. 2016. “Concrete Problems in AI Safety.” arXiv:1606.06565 [cs.AI]. https://arxiv.org/pdf/1606.06565v1.pdf.
Bostrom, Nick. 2014. Superintelligence: Paths, Dangers, Strategies. Oxford: Oxford University Press.
Brogan, Jacob. 2016. "Digital Genies." Slate, April 22. http://www.slate.com/articles/technology/future_tense/2016/04/stuart_russell_interviewed_about_a_i_and_human_values.html.
Frey, Carl, and Michael Osborne. 2013. “The Future of Employment: How Susceptible Are Jobs to Computerisation.” http://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf.
MIRI Blog, May 4. https://intelligence.org/2016/05/04/announcing-a-new-research-program/.
Prodhan, Georgina. 2016. “Europe's robots to become 'electronic persons' under draft plan.” Reuters. http://www.reuters.com/article/us-europe-robotics-lawmaking-idUSKCN0Z72AY.
Russell, Stuart. 2015. "2015: What Do You Think About Machines That Think?" Edge. https://www.edge.org/response-detail/26157.
Russell, Stuart, Daniel Dewey, and Max Tegmark. 2015. “Research Priorities for Robust and Beneficial Artificial Intelligence”. AI Magazine 36:4.
Soares, Nate and Benja Fallenstein. 2014. “Agent Foundations for Aligning Machine Intelligence with Human Interests: A Technical Research Agenda.” MIRI. https://intelligence.org/files/TechnicalAgenda.pdf.
Taylor, Jessica. 2016. “A New MIRI Research Program with a Machine Learning Focus.”
Yudkowsky, Eliezer. 2008. “Artificial Intelligence as a Positive and Negative Factor in Global Risk.” In Global Catastrophic Risks, edited by Nick Bostrom and Milan M. Cirkovic, 308–345. New York: Oxford University Press.

Respondent 133

Bijan Madhani, Computer & Communications Industry Association

The Computer & Communications Industry Association commends the White House’s Office of Science and Technology Policy for its framing of this Request for Information on the future of Artificial Intelligence (AI). As the RFI notes, AI technologies offer great promise for new and innovative products and services, economic growth, and applications across society.

In discussing the potential benefits of artificial intelligence, it is helpful to sketch out the

contours of the field. Today, AI refers to the technical discipline of making machines intelligent—computational systems that can respond to complex factors in a particular context to achieve some goal.

Rather than a theoretical technology relegated to science fiction, AI is a tool currently used by academics, engineers, and scientists worldwide. Present research and development is focused on the practical application of AI to existing problems, rather than the development of an artificial general intelligence commonly portrayed in science fiction. Machine learning is a related discipline that has direct relevance to AI's use as a problem-solving tool, because it enables systems to make inferences from large samples of data.

While AI has been developing rapidly in recent years, its continued progress and impact cannot be taken for granted. AI has the potential to transform healthcare, transportation, security, education, and more—but only if stakeholders work collectively to encourage its innovative potential.

1. The legal and governance implications of AI

Like any new technology, artificial intelligence and its practical applications can raise regulatory and legal questions. As AI technologies develop, so too will society's ability to manage their use and to determine areas of possible concern. The ultimate goal should be to avoid actual harms and promote innovation in and the use of AI.

The range of potential uses for artificial intelligence is enormous and cuts across sectors. No single regulatory solution will appropriately respond to all possible concerns. But AI-enabled platforms are not emerging in a regulatory vacuum. The data and activities associated with the most sensitive applications of artificial intelligence are already subject to the protections of existing rules, which cover areas including privacy, data security, energy, finance, and transportation. Each of these sectors has an expert agency with knowledge and tools available to ensure that any harms AI might pose are appropriately addressed. The government should convene these agencies and stakeholders before considering new regulation to properly apply the protections of existing rules.

If specific new rules are deemed necessary to respond to concerns about AI, policymakers should look toward principles-based guidelines where possible. Best practices developed through stakeholder consensus can help drive innovation while providing protection where necessary. As appropriate, the government should convene stakeholders to aide the development of industry-wide best practices and self-regulatory regimes for the various applications of AI.

2. The use of AI for public good

Artificial intelligence and machine learning can be used for the public good in a variety of

fields, including healthcare, cybersecurity, and education.

The application of AI to healthcare problems will allow physicians to be more accurate, see more patients, and save more lives. AI can reduce human error by helping scientists and clinicians detect patterns in medical data, diagnose illnesses, and recommend treatments. Several startups around the country already use machine learning techniques and predictive analysis to provide personalized healthcare guidance to patients, improved follow-up care, and better identification of new pharmaceutical therapies. Using AI in healthcare improves the quality of care, lowers costs, and delivers better outcomes.

Cybersecurity, another data-driven field, is similarly primed for AI-enabled growth. Intelligent algorithms are beginning to form the core of real-time threat prediction, detection, and response on secure networks in the event of a cyberattack. For example, machine learning enables systems to understand normal user and network behaviors, for later use in identifying deviations that signal possible intrusions. Similar tools deployed by information sharing and analysis organizations can coordinate to detect fraud, combat breaches, and reduce identity theft across sectors and regions.

Significant social benefits will result from the application of AI to education. Smarter software can help teachers customize lesson plans based on individual students' needs and automate basic activities. Students will benefit from educational software that adapts to different learning styles and paces of study, which can also facilitate remote instruction.

4. The social and economic implications of AI

a. Economic implications

Artificial intelligence can lead to efficiencies and productivity improvements across the economy. AI-enabled modeling software can help analyze data, manage records, automate information acquisition, optimize logistics, and produce valuable insights about markets. The Analysis Group recently estimated that AI could have an aggregate economic impact of $1.49 trillion to $2.95 trillion over the next ten years.

The cumulative economic effects of advances in artificial intelligence and deep learning are likely to be positive, both in terms of labor participation and labor productivity, as proven by many prior technological innovations. Although the concern of "AI replacing humans" has received significant attention, AI does not mean automation. A more accurate representation of the effects of AI, particularly in the short and medium term, is a future in which deep learning augments human labor to increase workforce productivity and help create new jobs.

These productivity boons will be particularly important for small businesses. Smart platforms can boost economic activity by large numbers of small enterprises by allowing

them to intelligently scale their businesses and empower their employees through smarter tools.

b. Social implications: avoiding discrimination

AI systems that help make decisions based on complex factors and data sets can raise concerns about unfair or discriminatory outcomes. These outcomes might result from design choices or biases inherent in the data used to condition an intelligent system. If potential sources of bias are not unaccounted for, actual harms can result.

But well-designed AI systems can also help avoid discrimination in areas where it is unintentionally present. For example, present professional hiring practices can sometimes lead to unconsciously biased results. A number of new startups are helping to incorporate machine learning and automation into hiring processes. By using employers' own data and publicly available information to suggest candidates who might otherwise have been dismissed for reasons unrelated to qualification and fit, these startups are helping recruiters build more diverse and productive workplaces.

In seeking to avoid discrimination, policymakers should recognize that AI-enabled systems are simply tools. Existing laws that apply to sensitive areas like housing, finance, and employment already provide technology-neutral remedies for disparate impacts. It would be counterproductive to mandate human involvement in every AI system, since people often hold inherent biases. Regulators should instead aim to provide companies and consumers with tools to diagnose and prevent failures that might lead to discrimination.

Biased outcomes are also often the result a lack of quality data, which can negatively affect an otherwise well-intentioned machine learning protocol. To help rectify this, governments should facilitate the release of robust datasets that enable responsible analysis and use, especially in areas where AI systems are publicly deployed.

8. The specific steps that could be taken by the federal government, research institutes, universities, and philanthropies to encourage multi-disciplinary AI research

The government should enable policies that encourage research and development, foster the AI workforce, and promote public AI deployment.

a. Encourage diversity in all aspects of AI development

It is imperative for innovation and AI development that communities be diverse and represent a broad set of backgrounds and experiences. Key companies advancing AI, such as Nvidia, Google, and Enlitic, were founded by immigrants. Immigrant academics have become some of the leading voices and advocates for AI in the U.S. and help shape its future workforce to take advantage of their expertise. The U.S. should increase the availability of

H-1B visas to further capitalize on this worldwide talent pool.

b. Invest more resources in STEM education

The government, universities and research institutes should prioritize the value proposition and flexibility of STEM disciplines when recruiting individuals, as these skills directly translate to improved AI research and development. Examples include the White House's recently launched Computer Science for All initiative, which enables students to develop computational thinking skills early, and the Department of Labor's the TechHire program, which provides federal funding for accelerated talent pipelines in STEM-focused sectors.

c. Support internal government expertise in technology

The government should continue to expand its technical capabilities through programs like the U.S. Digital Service and 18F. Every agency will be better positioned to leverage AI technologies for complex problems in their respective domains if they house experts in computer science and technology.

Similarly, the Presidential Innovation Fellows program aims to connect innovative thinkers with relevant government agencies and civil servants. Fellows bring expert knowledge and practices into the government to address some of the nation's biggest challenges at the convergence of technology, policy, and process. This collaborative, user-centric approach will be essential for implementing AI-based solutions across the federal government.

d. Leverage global innovation networks

The U.S. should support pro-innovation legal regimes abroad. It should continue to partner with other innovative countries to share resources and advance areas for cooperative growth. Concurrently, these countries should also establish partnerships with developing nations, which have proven that innovation-driven growth is no longer the prerogative of high-income countries and have increasingly designed policies to increase their innovation capacity.

Progress in artificial intelligence is the product of international collaboration. Copyright is one field in which the U.S. can promote pro-innovation frameworks. Machine learning in particular is dependent on balanced copyright laws that promote innovation. Machine learning generally requires the analysis of large samples of data and information to condition an intelligent algorithm, the availability of which may be restricted by copyright regulations in certain countries. In the United States, established limitations and exceptions to copyright, such as fair use, enable access to non-expressive use of works for innovative purposes. However, U.S. companies, especially startups and small businesses, may face anticompetitive restrictions in other countries, which the U.S. should work to address with its partners.

e. Share data and support AI research

Private companies often underinvest in research and development since return on investment for experimentation is uncertain. In the past, federally-funded research has been the catalyst for many of today's AI technologies. Today, the government provides 60% of the funding for basic research in AI. NSF has spent $200 million thus far on AI, and DoD supports research with the Machine Reading and Mind's Eye projects as part of its annual $250 million budget on big data. As AI matures, the government must continue to budget for basic research on machine learning and emerging AI technologies.

A key part of the evolution of AI technology has been the development of technology-powered platforms and services that bring value to the mass market. Prioritizing research and AI projects within and in collaboration with government agencies can foster greater academic participation and industry growth. Many advances in machine learning have been products of research projects largely funded by DARPA, which offers cash prizes to innovators who successfully complete challenges in fields like robotics. Expert agencies could foster collaborations between machine learning experts and domain experts in fields like medicine, science, and business. This approach to research increases transparency and creates informed strategies that benefit future developments.

In addition, some research communities can have insular technical conferences. Community leaders, the federal government, and philanthropies could all provide support for collaborative venues through which machine learning and domain experts can interact.

Finally, the federal government should facilitate the responsible analysis and use of AI systems through the release of accurate and robust datasets. The ImageNet visual database and image classification challenge have helped spur the recent commercial deployment of deep learning algorithms. The Department of Commerce has been particularly active in advocating for open data initiatives and making datasets available to businesses. Encouraging the creation and curation of new or better datasets in a variety of application areas will help further drive machine learning to society's benefit.

Respondent 134

Maria Gini, University of Minnesota

(3) No robot should be allowed to kill people without human intervention. It is the Geneva convention, yet drones used for military operations have the capability to autonomously decide to kill people. The issue is not technical, nobody can prevent the development of robots capable of killing people, it is a policy issue. Progress on this front will go a long way to reassure the public that killer robots will have no legal place in our future.

(5) Computers need the ability to use common sense when making decisions or recommendations. Despite a lot of research, this still
remains a pressing and open question in AI. AI has made great progress in addressing specific sets of problems, such as playing Go or recognizing features in images or recognizing spoken language, but it is still lacking the ability to reason across problem domains and make connections among things that were not explicitly considered when developing systems.

(6) The use of data to learn predictive models has many valuable applications, yet it risks biasing the future by perpetuating the past. It is not just the example of loans, where minority people
are less likely to receive a loan because of past data on loan default. It is what to do when the data available are insufficient or collected
from a biased sample. Most machine learning algorithms assume independent and identically distributed random variables, but often
the data available are not. How to use past data to build models but then understand the limitations and assumptions of the models, and suggest different ways of making decisions, so that we learn from the past but do not limit the future.

(7) Computational thinking has to be taught widely. It is not more than STEM, STEM is too broad and includes disciplines like biology where there is an abundance of trained people compared to the jobs available. It is a way of thinking and formulating problems and solutions so precisely
that a computer can solve them. It is the foundation of AI (From John McCarthy, Dartmouth AI conference "The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it."). Computational thinking will help everyone to understand what AI can do, not just because their are shown specific examples but because they understand the power of precise reasoning. It will also help develop a new generation of citizens that are empowered to construct new solutions to world problems.

Respondent 135

Achutha Raman, 4PrivacyMatters

This brief response is respectfully submitted by Achutha Raman and 4Privacymatters to address the following areas identified by OSTP:
#2: The use of AI for the public Good
#11: additional information related to AI Research.
AI application research as is relates to Privacy has been to date in the area of anonymization and its inverse function i.e. de-identification. The public could benefit from research that can enable AI Agents to report on the data provenance associated with a data driven offer. This basic foundational capability can in turn enable deeper compliance checks that

guarantees individual privacy across the digital fabric. Areas that will need to be addressed include:

- Establishing provenance of datasets that could be embedded with any and all data so that they may be reconstructed on the fly – solutions could possibly lie in the intersection of Blockchain and AI technology
- Helping the Public easily enable AI agents that act on their behalf to notify privacy-breached use of data when encountering an offer made to them via a digital channel
- Developing AI Agents that can be used by compliance and regulatory testing agencies to check for Privacy non-compliance by taking up various personae and autonomously interrogating the "offer" generators

Respondent 136

Nate Soares, Machine Intelligence Research Institute

From Nate Soares, executive director of the Machine Intelligence Research Institute (XXXXXXXXX).

I. Review of safety and control concerns

AI experts largely agree that AI research will eventually lead to the development of AI systems that surpass humans in general reasoning and decision-making ability.[1] This is, after all, the goal of the field. However, there is widespread disagreement about how long it will take to cross that threshold, and what the relevant AI systems are likely to look like (autonomous agents, widely distributed decision support systems, human/AI teams, etc.).

Despite the uncertainty, a growing subset of the research community expects that advanced AI systems will give rise to a number of foreseeable safety and control difficulties, and that those difficulties can be preemptively addressed by technical research today. Stuart Russell, co-author of the leading undergraduate textbook in AI and professor at U.C. Berkeley, writes:

"The primary concern is not spooky emergent consciousness but simply the ability to make high-quality decisions. Here, quality refers to the expected outcome utility of actions taken, where the utility function is, presumably, specified by the human designer. Now we have a problem:

"1. The utility function may not be perfectly aligned with the values of the human race, which are (at best) very difficult to pin down.

"2. Any sufficiently capable intelligent system will prefer to ensure its own continued existence and to acquire physical and computational resources — not for their own sake, but to succeed in its assigned task.

“A system that is optimizing a function of n variables, where the objective depends on a subset of size k<n, will often set the remaining unconstrained variables to extreme values; if one of those unconstrained variables is actually something we care about, the solution found may be highly undesirable. This is essentially the old story of the genie in the lamp, or the sorcerer's apprentice, or King Midas: you get exactly what you ask for, not what you want.”[2]

Researchers’ worries about the impact of AI in the long term bear little relation to the doomsday scenarios most often depicted in Hollywood movies, in which “emergent consciousness” allows machines to throw off the shackles of their programmed goals and rebel. The concern is rather that such systems may pursue their programmed goals all too well, and that the programmed goals may not match the intended goals, or that the intended goals may have unintended negative consequences.

These challenges are not entirely novel. We can compare them to other principal-agent problems where incentive structures are designed with the hope that blind pursuit of those incentives promotes good outcomes. Historically, principal-agent problems have been difficult to solve even in domains where the people designing the incentive structures can rely on some amount of human goodwill and common sense. Consider the problem of designing tax codes to have reliably beneficial consequences, or the problem of designing regulations that reliably reduce corporate externalities. Advanced AI systems naively designed to optimize some objective function could result in unintended consequences that occur on digital timescales, but without goodwill and common sense to blunt the impact.

Given that researchers don’t know when breakthroughs will occur, and given that there are multiple lines of open technical research that can be pursued today to address these concerns, we believe it is prudent to begin serious work on those technical obstacles, to improve the community’s preparedness.

II. Technical research for safety and control

There are several promising lines of technical research that may help ensure that the AI systems of the future have a positive social impact. We divide this research into three broad categories:

- Value specification (VS): research that aids in the design of objective functions that capture the intentions of the operators, and/or that describe socially beneficial goals. Example: cooperative inverse reinforcement learning, a formal model of AI agents that inductively learn the goals of other agents (e.g., human operators).[3]

- High reliability (HR): research that aids in the design of AI systems that robustly, reliably, and verifiably pursue the given objectives. Example: the PAC learning framework, which

gives statistical guarantees about the correctness of solutions to certain types of classification problems.[4] This framework is a nice example of research done far in advance of the development of advanced AI systems that is nevertheless likely to aid in the design of systems that are robust and reliable.

- Error tolerance (ET): research that aids in the design of AI systems that are fail-safe and robust to design errors. Example: research into the design of objective functions that allow an agent to be shut down, but do not give that agent incentives to cause or prevent shutdown.[5]

Our "Agent foundations for aligning machine intelligence with human interests" report discusses these three targets in depth, and outlines some neglected technical research topics that are likely to be relevant to the future design of robustly beneficial AI systems regardless of their specific architecture.[6] Our "Alignment for advanced machine learning systems" report discusses technical research topics relevant to these questions under the stronger assumption that the advanced systems of the future will be qualitatively similar to modern-day machine learning (ML) systems.[7] We also recommend a research proposal led by Dario Amodei and Chris Olah of Google Brain, "Concrete problems in AI safety," for technical research problems that are applicable to near-future AI systems and are likely to also be applicable to more advanced systems down the road.[8] Actionable research directions discussed in these agendas include (among many other topics):

- robust inverse reinforcement learning: designing reward-based agents to learn human values in contexts where observed behavior may reveal biases or ignorance in place of genuine preferences. (VS)
- safe exploration: designing reinforcement learning agents to efficiently learn about their environments without performing high-risk experiments. (ET)
- low-impact agents: specifying decision-making systems that deliberately avoid having a large impact, good or bad, on their environment. (ET)

There are also a number of research areas that would likely aid in the development of safe AI systems, but which are not well-integrated into the existing AI community. As an example, many of the techniques in use by the program verification and high-assurance software communities cannot be applied to modern ML algorithms. Fostering more collaboration between these communities is likely to make it easier for us to design AI systems suitable for use in safety-critical situations. Actionable research directions for ML analysis and verification include:[9][10][11]

- algorithmic transparency: developing more formal tools for analyzing how and why ML algorithms perform as they do. (HR)
- type theory for program verification: developing high-assurance techniques for the re-use of verified code in new contexts. (HR)
- incremental re-verification: confirming the persistence of safety properties for adaptive systems. (HR)

Another category of important research for AI reliability is the development of basic theoretical tools for formally modeling intelligent agents. As an example, consider the interaction of probability theory (a theoretical tool for modeling uncertain reasoners) with modern machine learning algorithms. While modern ML systems do not strictly follow the axioms of probability theory, many of the theoretical guarantees that can be applied to them are probability-theoretic, taking the form "this agent will converge on a policy that is very close to the optimal policy, with very high probability." Probability theory is an example of basic research that was developed far in advance of present-day ML techniques, but has proven important for attaining strong (statistical) guarantees about the behavior of ML systems. We believe that more basic research of this kind can be done, and that it could prove to be similarly valuable.

There are a number of other aspects of good reasoning where analogous foundations are lacking, such as situations where AI systems have to allocate attention given limited computational resources, or predict the behavior of computations that are too expensive to run, or analyze the effects of potential alterations to their hardware or software. Further research into basic theoretical models of ideal reasoning (including research into bounded rationality) could yield tools that would help attain stronger theoretical guarantees about AI systems' behavior. Actionable research directions include:[6]

- decision theory: giving a formal account of reasoning in settings where an agent must engage in metacognition, reflection, self-modification, or reasoning about violations of the agent/environment boundary. (HR)
- logical uncertainty: generalizing Bayesian probability theory to settings where agents are uncertain about mathematical (e.g., computational) facts. (HR)

We believe that there are numerous promising avenues of foundational research which, if successful, could make it possible to get very strong guarantees about the behavior of advanced AI systems — stronger than many currently think is possible, in a time when the most successful machine learning techniques are often poorly understood. We believe that bringing together researchers in machine learning, program verification, and the mathematical study of formal agents would be a large step towards ensuring that highly advanced AI systems will have a robustly beneficial impact on society.

III. Coordination prospects

It is difficult to say much with confidence about the long-term impact of AI. For now, we believe that the lines of technical research outlined above are the best available tool for addressing concerns about advanced AI systems, and for learning more about what needs to be done.

Looking ahead, we expect the risks associated with transformative AI systems in the long term to be exacerbated if the designers of such systems (be they private-sector, public-sector, or part of some international collaboration) act under excessive time pressure. It is

our belief that any policy designed to ensure that the social impact of AI is beneficial should first and foremost ensure that transformative AI systems are deployed with careful consideration, rather than in fear or haste. If scientists and engineers are worried about losing a race to the finish, they will have more incentives to cut corners on safety and control, obviating the benefits of safety-conscious work.

In the long term, we recommend that policymakers make use of incentives to encourage designers of AI systems to work together cooperatively, perhaps through multinational and multicorporate collaborations, in order to discourage the development of race dynamics. In light of high levels of uncertainty about the future of AI among experts, and in light of the large potential of AI research to save lives, solve social problems, and serve the common good in the near future, we recommend against broad regulatory interventions in this space. We recommend that effort instead be put towards encouraging interdisciplinary technical research into the AI safety and control challenges that we have outlined above.

---

[1] Müller and Bostrom (2014). "Future progress in artificial intelligence: A survey of expert opinion." http://www.nickbostrom.com/papers/survey.pdf

[2] Russell (2014). “Of myths and moonshine.” https://www.edge.org/conversation/the-myth-of-ai#26015

[3] Hadfield-Mennell, et al. (2016). “Cooperative inverse reinforcement learning.” https://arxiv.org/abs/1606.03137

[4] Haussler (1990). "Probably approximately correct learning." http://aaaipress.org/Papers/AAAI/1990/AAAI90-163.pdf

[5] Soares, et al. (2015). “Corrigibility.” https://intelligence.org/files/Corrigibility.pdf

[6] Soares and Fallenstein (forthcoming). “Agent foundations for aligning machine intelligence with human interests.” https://intelligence.org/files/TechnicalAgenda.pdf

[7] Taylor, et al. (forthcoming). “Alignment for advanced machine learning systems.” https://intelligence.org/files/AlignmentMachineLearning.pdf

[8] Amodei, et al. (2016). “Concrete problems in AI safety.” https://arxiv.org/abs/1606.06565

[9] Muehlhauser (2014). “Diana Spears on the safety of adaptive agents.” https://intelligence.org/2014/04/09/diana-spears/

[10] Muehlhauser (2014). “Gerwin Klein on formal methods.” https://intelligence.org/2014/02/11/gerwin-klein-on-formal-methods/

[11] Muehlhauser (2014). “Robert Constable on correct-by-construction programming.” https://intelligence.org/2014/03/02/bob-constable/

Respondent 137

Ann Drobnis, Computing Community Consortium

Overview

On June 7, 2016, the Computing Community Consortium (CCC) and Association for the Advancement of Artificial Intelligence (AAAI) hosted a roundtable discussion on Artificial Intelligence for Social Good. The following response is a summary of the roundtable discussions. A more thorough report will be published by the CCC later this summer, and available at http://cra.org/ccc/resources/ccc-led-whitepapers/. The remainder of this document is organized into a brief recitation of discussions for the four areas discussed in the workshop, followed by some cross-cutting observations and recommendations.

AI for Urban Environments

The urban computing workshop session focused primarily on transportation networks, the goal being to use AI technology to improve mobility. However, transportation can be viewed as a concrete example of a service industry which is very likely to be transformed over the coming decade by AI technologies. Time spent commuting to school or to work is time not spent studying or with one’s family. Lack of transportation reduces access to preventative healthcare, easy access to supermarkets with healthful food is highly correlated with obesity (and hence heart disease, diabetes, etc.) and easy access for people to standard bank accounts is costly. AI technology has the potential to significantly improve mobility, and hence substantially reduce these and other inefficiencies in the market. Technology exists that can mobilize people who have been immobile; to increase flow/decrease congestion; and autonomous vehicles have the potential to decrease emissions. The easier it becomes for people to move about, the more vibrant our urban areas will be; likewise, the more fruitful the social and economic interactions that take place inside them will be.

Ubiquitous connectivity and instrumentation are enabling us to measure things that were previously unmeasurable. We can now collect information about individuals’ travel patterns, so that we can better understand how people move through cities, thereby improving our understanding of city life. AI technology can then be leveraged to move from descriptive models (data analytics) to predictive ones (machine learning) to prescriptive decisions (optimization, game theory, and mechanism design). The potential of this transformation is being demonstrated in pilot systems that optimize the flow of traffic

through cities, and in new on-demand, multi-modal transportation systems. It is now within the realm of AI technology to optimize traffic lights in real time, continuously adapting their behavior based on current traffic patterns (Smith, 2016); and to dispatch fleets of small vehicles to provide on-demand transportation, address the “first and last mile” problem that plagues many urban transit systems (Van Hentenryck, 2016). More pilot deployments are needed to fully understand the scope of the transformation that is under way in our cities.

In spite of the significant promise, many challenges lie ahead before these new opportunities can be fully realized. Transportation systems are complex, socio-technical systems that operate over multiple spatial and temporal scales. It is critical that we scale up existing pilots to multi-modal transportation models -- incorporating pedestrians, bicycles, cars, vans, and buses -- so that we can begin to understand how these models will impact big cities. Fundamental to this effort, it is crucial that we understand the human behavioral changes that new forms of mobility will induce, and the impact those behaviors will have on the efficacy of our system.

Sustainability

Sustainability can be interpreted narrowly as the conservation of endangered species and the sustainable management of ecosystems. It can also be interpreted broadly to include all aspects of sustainable biological, economic, and social systems that support human well-being. In this panel, the discussion focused primarily on the ecological component, but the larger issues of social and economic sustainability must be considered as well.
Automated data collection systems develop and deploy sensor networks (e.g. Trans-Africa Hyrdo-Meteorological Observatory; www.tahmo.org), camera traps to collect image or acoustic data, or unmanned aerial vehicles to obtain video imagery. AI algorithms are applied to optimize the locations of these sensors and traps. Crowd- sourcing and/or employing technically-trained people to collect data, such as the freshwater stream surveys conducted by the EMAP project (https://archive.epa.gov/emap/archive-emap/web/html/), are being married with computer vision methods as another hybrid method of data collection.
Techniques from data mining, statistics, and machine learning are used to discover trends and fit models. Such models can predict migration, dispersal, reproduction, and mortality of species. Virtually every ecosystem management problem combines an ecological model with an economic model of the economic costs and benefits of various policy outcomes. Examples include the design of a schedule for purchasing habitat parcels to support the spatial expansion of the Red Cockaded Woodpecker (Sheldon, et al., 2010; Sheldon, et al., 2015), and the use of detailed bird migration models developed by the Cornell Lab of Ornithology to rent rice fields in California (Nicol, et al., 2015). Algorithms for computing these policies combine ideas from network cascade analysis (maximizing spread in social networks) with techniques from machine learning, AI planning and decision-making, and Monte Carlo optimization. Finally, the PAWS project (Fang, et al., 2016) applies AI algorithms for game theory to optimize the patrol routes of game wardens in order to

maximize their deterrent effect while minimizing costs.
A major challenge for the medium term is to develop methods that can collect and model data encompassing a broad range of species at continental scales. A related challenge for current modeling efforts is that they generally assume stationary (steady-state) climate, land use, and species behavior whereas the real systems are experiencing climate change, rapid economic development, and continuing evolution, dispersal, and natural selection of species. Furthermore, as the scale of policy questions grows, it is no longer possible to focus only on the biological components of a system. Instead, one must incorporate models of social, cultural, and economic activity. Finally, sustainability hot spots are often located in developing countries. Issues that arise include poor networking infrastructure, little access to high-performance computing resources, lack of local personnel with sufficient education and training, and persistent corruption.

In the longer term, we must confront the fact that the long term behavior of ecological, economic, and social systems is radically uncertain. How can artificial intelligence methods deal with the uncertainty of these "unknown unknowns"? When formulating and optimizing management policies, we should adopt risk-sensitive methods. This is an active area of research (see, e.g., Chow, et al., 2015), and much more work is needed to understand how we can ensure that our models are robust to both the known unknowns (as in traditional risk management methods) and the unknown unknowns.

Healthcare

AI is well-positioned to have a broad and sustained impact on many aspects of healthcare. Social media analytics is emerging as an alternative or complementary approach for instantly measuring public health at large scale and with little or no cost. The nEmesis system, for example, helps health departments identify restaurants that are the source of food-borne illness (Sadilek et al. 2016). Decision support in a clinical environment is a second important area. The Surgical Critical Care Initiative (SC2i), a Department of Defense funded research program, has deployed two clinical decision support tools (CDSTs) to realize the promise of precision medicine for critical care (Belard et al. 2016). The invasive fungal infection CDST was deployed in 2014 to assist military providers with treatment decisions both near point of injury and at definitive treatment centers. The massive-transfusion protocol (MTP) CDST is currently being assessed under a two-year clinical trial at Emory-Grady, one of the two SC2i civilian hospitals. Automated real-time surveillance tools, operating from the electronic health record, identify individuals at risk for severe sepsis and septic shock at the early stages of decline, and much earlier than standard of care (Henry et al., 2015).
Opportunities in this space include:

Targeted therapy decisions: Many chronic diseases are difficult to treat because of high variation among affected individuals. Computational subtyping, for example, seeks to refine disease definition by identifying groups of individuals that manifest a disease similarly

(Saria & Goldenberg, 2015) (Collins, 2015). These subtypes can be used within a probabilistic framework to obtain individualized estimates of a patient's future disease course (Schulam & Saria, 2015).

New sensors, new healthcare delivery: AI can be used to analyze social media data and discover and suggest behavioral and environmental impacts on health -- e.g tracking influenza or quantifying alcohol and drug abuse in communities. Social networks can also be used to address the informational and psychosocial needs of individuals and the opportunity for cost-effective interventions for addressing mental health, addiction, and behavioral health issues using modern low cost sensing technologies. Low fidelity sensors, some of which are diagnostic, together with AI and internet technologies can enable low barrier telemedicine for example for chronic healthcare. Advances in natural language processing and machine reading can be used to synthesize, integrate and appropriately disseminate new medical knowledge (e.g., as reported in journal articles).

Pivoting from personalized medicine to personalized health will keep people from going to the hospital in the first place, and dealing with life issues and not just specific disease. For this, we need to move to modeling of the health of individuals and populations by using integrated data sets--- electronic health records data and other data gathered within the health system with genomic, socio-economic, demographic, environmental, social network and social media and other, non-traditional data sources, such as social service and law enforcement data.

Collaborative Decision-Making approaches that allow decision makers to reason with models of the health of individuals are needed. For example, can a healthcare provider ask how would a health trajectory change if the individual was being treated with two different drugs?

Challenges include: 1) addressing Bias that arises in fitting models from observational health data sources; 2) privacy and security methods that support work with data in a way that both sustains its utility while the decisions and outcomes of working with the data do not reveal information about individuals is essential; 3) incentive alignment to encourage various actors in the health ecosystem a reason to collect additional data and make their data available to the rest of the healthcare ecosystem; and 4) cloud-based data science platforms and common data models should be developed and promoted in order to reduce the barrier to entry for researchers and increase the likelihood of societally beneficial outcomes.

Public Welfare

AI has not had a lot of impact on fundamental issues our society faces today. However many opportunities exist. For example, the University of Chicago partnered with Chicago Department of Public Health to build a system to predict which children are at risk of lead

poisoning to allow CDPH to deploy inspectors and proactively address lead hazards. Over the past several years, several school districts around the US have been collaborating with universities to develop AI based systems to help them identify at-risk students who are unlikely to finish high school on time. Finally, the University of Chicago has been working, as part of the White House Police Data Initiative, to identify officers who are at risk of adverse incidents early and accurately so supervisors can effectively target interventions.

Work in this area requires deep and sustained interaction and efforts between the target community and AI researchers, but there isn't a ready supply of trained AI researchers (or practitioners) who are familiar with the unique aspects of working on public welfare problems. LIkewise, government and policymakers have little experience working directly with the research community. Finding funding mechanisms that bring both communities together to address local needs -- e.g. the NSF Data Hubs model -- is essential. Highlighting ongoing projects (and successes) to both raise awareness and to provide a roadmap is essential to growing this community. Platforms that are able to access, structure, and curate appropriate data sets do not exist.

Projects need to have a long-term structure, with appropriate intermediate goals, to avoid short-term fixes, or quick, but ephemeral, "feel-good" stories. Legal and regulatory hurdles including, access to data, and to populations to evaluate against, will require substantial investment of time, planning, and resources to effect. Creating a framework for ethical evaluation of costs and benefits must be established. Understanding the impact of innovations will require an understanding of the level of compliance, and possibly methods to manage or pivot solutions in response to perception, trust, and compliance of the target population.

There are several related technical challenges. Privacy issues, transparency and traceability of data collection and decision-making, and understanding of social context must be considered within the research context. Issues surrounding data bias and uncertainty have direct implications to fairness and the evaluation of the utility of possible decision paths. Related (government) organizational and (population) sociological constraints must also be considered. More technical problems include: 1) data analytics and machine learning models that are robust to systematic bias, missing data, and data heterogeneity; 2) the development of models or simulations that sufficiently predict to inform decision-making, and which also can then be adapted "closed-loop" as additional data is collected with time; 3) advanced models of decision-making and planning that incorporate social dynamics, resource constraints, and utility models for multiple actors; 4) consistent, cost-effective, and scalable models for measurement or data collection; and 5) methods for causal reasoning and explanation.

Some near term opportunities include: 1) tracking of location data and understanding how to better predict/deploy first responders, 2) using individual public transit and other transportation data (uber, bikeshare, etc.) to understand mobility patterns of people to

understand gaps in transit (where they live - where they work - what services they need) and also to assess impact of policy changes; 3) better detection of women who may be at risk of adverse births to target human services programs and resources; 4) better detection of adults in danger of becoming homeless/incarcerated; 5) increase the number of kids who are performing at grade level by creating interventions that would influence and change behavior; and 6) enhancing access to services/food/health.

Longer-term opportunities will build on the establishment of a platform for evidence-based decision making by government informed by more detailed and nuanced models. For example, is it possible to predict the acceptance or engagement of the population to a particular policy change. Also, such models could move toward a "systems of systems" analysis where information about welfare impacts education impacts law enforcement impacts health. Achieving these ends will require methods to integrate multiple AI systems, and monitor, detect, diagnose, and adapt to multi-faceted population behaviors.

Cross-cutting Observations and Recommendations

To date, AI has typically focused around deploying narrow wedges of technology in narrow application areas. However, as we look across application spaces, we see a common thread of needs and approaches that are necessary to scale these "niche" approaches to address broad socio-technical themes. Common themes in this report and our discussions include: 1) improving data quality and availability; 2) supporting technology and policies that ensure individual privacy and data security; 3) mechanisms to promote collaboration (at development time) and adoption (at deployment time) of innovations; 4) mechanisms to ensure fairness, transparency, accountability, reliability of decision; 5) methods to accurately measure and assess the effect of a technology intervention over varying time-scale; 6) long-term programs that train scientists in developing AI methods for complex socio-technical systems.

References

Belard A, Buchman T, Forsberg J, Potter BK, Dente CJ, Kirk A, Elster E. Precision diagnosis: a view of the clinical decision support systems(CDSS) landscape through the lens of critical care. J Clin Monit Comput. 2016 Feb 22.

Chow, Y., Tamar, A., Mannor, S., & Pavone, M. (2015). Risk-Sensitive and Robust Decision-Making: a CVaR Optimization Approach. In NIPS 2015.

Collins, F. Big Data Study Reveals Possible Subtypes of Type 2 Diabetes. NIH Director's Blog. 2015.
https://directorsblog.nih.gov/2015/11/10/big-data-reveals-possible-subtypes-of-type-2-diabetes/

Commuting: “The Stress that Doesn’t Pay”. Psychology Today. (2015) https://www.psychologytoday.com/blog/urban-survival/201501/commuting-the-stress-doesnt-pay

Fang, F., Nguyen, T. H., Pickles, R., Lam, W. Y., Clements, G. R., An, B., Singh, A., Tambe, M., Lemieux, A. (2016). Deploying PAWS : Field Optimization of the Protection Assistant for Wildlife Security. In Proceedings of the Twenty-Eighth Innovative Applications of Artificial Intelligence Conference. AAAI.

Henry, K.E., Hager, D.N., Pronovost, P.J. and Saria, S., 2015. A targeted real-time early warning score (TREWScore) for septic shock. Science Translational Medicine, 7(299), pp.299ra122-299ra122.

Nicol, S., Fuller, R. A., Iwamura, T., & Chadès, I. (2015). Adapting environmental management to uncertain but inevitable change. Proceedings Royal Society B, 282(1808), 20142984. http://doi.org/10.1098/rspb.2014.2984

Sadilek, A., Kautz, H., DiPrete, L., Labus, B., Portman, E., Teitel, J., Silenzio, V. (2016). Deploying nEmesis: Preventing Foodborne Illness by Data Mining Social Media. Association for the Advancement of Artificial Intelligence, 2016.

Saria, S., & Goldenberg, A. (2015). Subtyping: What it is and its role in precision medicine. Intelligent Systems, IEEE, 30(4), 70-75.

Schulam, P., & Saria, S. (2015). A Framework for Individualizing Predictions of Disease Trajectories by Exploiting Multi-Resolution Structure. In Advances in Neural Information Processing Systems (pp. 748-756).

Sheldon, D., Dilkina, B., Elmachtoub, A., Finseth, R., Sabharwal, A., Conrad, J., Finseth, R., Sabharwal, A., Conrad, J., Gomes, C., Shmoys, D., Allen, W., Amundsen, O. (2010). Maximizing the spread of cascades using network design. In Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (pp. 517–526).

Smith, S., Smart Infrastructure for Urban Mobility. Slides from the AI for Social Good Workshop. http://cra.org/ccc/wp-content/uploads/sites/2/2016/06/Stephen-Smith-AI-slides.pdf

Van Hentenryck, P. Reinventing Mobility with Artificial Intelligence. Slides from the AI for Social Good Workshop. http://cra.org/ccc/wp-content/uploads/sites/2/2016/06/Pascal-van-Hentenryck-AI-slides.pdf

Respondent 138

Sebastian Benthall, UC Berkeley School of Information

Sebastian Benthall
UC Berkeley School of Information

As a preface, it must be noted that "AI" is a dangerously underspecified term.

There is a real sense in which all significant use of computing since Alan Turing has been a form of "artificial intelligence". As an undergraduate taking Introduction to Artificial Intelligence in 2005, we learned that Marvin Minsky once defined AI as "anything humans are still better at than computers."

According to the very successful computational cognitive science paradigm, all cognition can be modeled as computational information processing. To a large extent the laws of rational thought have been codified in basic statistical and computational theory. (Anderson, 1991; Chater and Oaksford, 1999; Chater and Vitanyi. 2003; Russell et al., 2003; Griffiths, T. et al., 2008; Tenenbaum, J., et al., 2011) Contemporary machine learning and artificial intelligence is mainly just the implementation of these principles in machine systems designed to accomplish an expanding range of tasks.

What has changed recently is the expanding domain of data collected for use in systems implementing these rules and the consequent expanding domain of application of these systems. This is largely due to the ubiquity of computing through e.g. smartphones and its processing power through e.g. clustered and GPU processing. Of particular interests to governments is the use of AI as an actor exerting control on a social level, for example in the enforcement of public policy. Research has already shown that the subtle art of political prediction, a core competency for those involved in policy and governance, is better performed by simple statistical inference than by alleged "experts" (Tetlock, 2005). We should expect and encourage an expanding role of AI in governance.

(1) The legal and governance implications of AI;

There is recognition among legal scholars that software, especially in large-scale social systems, regulates society as does law (Lessig, 2009), that technological implementation of adminstrative policy challenges notions of due process (Citron, 2007), and that machine code (for example, DRM) can undermine law and bypass contract (Radin, 2004).

These ways that software challenge law are more extreme with respect to AI because the purpose of AI is to manage physical complexity (complexity due to data input size, or complexity due to computational need) that cannot be accomplished easily by human beings. To the extent that regulation of AI requires legal control, and legal control remains

an activity done primarily by human lawyers using natural language communication, AI presents an existential threat to legal authority. This has prompted some (Pasquale, 2015) to propose that systems that are too complex for people to understand should not be allowed to exist.

This view is tremendously short-cited. The growth and integration of society contribute to the need for expanding centralized information processing by the state. (Beniger, 2009). This information processing, whether it be bureaucratic or automated, is always beyond the grasp of an one person's understanding. Meanwhile, though the specific application of AI may be too complex for a person to understand because of the data used, the general principles of machine learning are within the grasp of a properly trained undergraduate.

A better approach to regulation of non-state AI (such as those systems used by industry) would be complementary state-run AI that has the same legal legitimacy as conventional state law and its enforcing institutions.

This raises pressing questions about the origin and legal status of the software, data collection systems, and actuators of such an AI.
AI systems can be purchased through standard government procurement mechanisms. It is important for for the logic of these systems to be available for review by lawmakers and the interested public. Requiring that these systems be open source is therefore necessary (but not sufficient) for addressing this need. Commercially supported open source software is used widely by the government in many domains; this is by no means a prohibitive restriction.
AI systems could be designed by lawmakers themselves. While the principles of engineering are today not considered part of legal training, that could change. Administrative law, which already provides ways for the state to employ technical specialists in the design of policy, is a promising area in which precedent for legitimate machine law could be established.

(2) the use of AI for public good;

The biggest difficulty facing the use of AI for the public good is the technical operationalization of “public good”. Present-day AI systems are excellent at optimizing the domains under their control according to well-defined goals. They are not excellent at learning those goals.

Recent research value learning is relevant to this problem. But even if an AI system is able to learn a person's values, whether or not these values can be aggregated into a measure of “the public good” is an open question. In any case, these operationalizations will be politically contestable.

In restricted domains, legitimate public policy may be a suitable proxy for “public good”. See response to prompt (1).

(4) the social and economic implications of AI;

* There is a prevalent concern that some uses of AI will “reproduce existing bias”. There is a problem with this framing in certain cases which is clear to those with statistical training: blame for an undesireable social outcome is placed on an AI system or product rather than the social context in which the product is embedded. The objectivity of an AI is challenged on the basis of its misuse. A good illustration of this problem is the controversial Northpointe recidivism risk assessment software (Angwin et al. 2016).

1. There is a noted problem, which is that Northpointe's risk assessments were differently poor for black defendants. Critics especially note that positive misclassifications rates (higher recidivism predicted than actual) were higher for black people, and conversely negative misclassification rates were lower for white people.

2. This difference in prediction meant that black people were treated less well--denied or higher bail, longer sentences.

3. Critics claim: that the Northpointe software reproduces existing bias because of these outcomes; that even those algorithms that implement sound statistical inference can be critiqued for their outcomes on the basis of the data used to train them; and that this software in particular is a case of biased software because of its outcomes.

* These criticisms are likely misguided.

1. There are two relevant kinds of statistical bias: sampling bias, where the data used to train statistical inference is not representative of the underly phenomenon, and inferential bias, bias in predicted values based on the algorithm for learning from that data.

2. If there is in fact a higher recidivism rate among black people than white people, then an algorithm that is unbiased in both the statistical and inferential sense will predict recidivism risk that is different between black people and white people as long as the training data it uses includes any information that correlates with race. This is a mathematical fact.

3. Prohibiting use of data that does not suffer from sampling bias in AI will reduce the quality of risk assessment.

4. If it is desirable to adjust the risk assessments so that they are neutral to race, it would be more effective to build this into the algorithm deliberately by taking account of race than by attempting to remove racially correlated information from the training data. (Dwork et al. 2012)

5. In this case and perhaps others, “reproducing existing bias” is a dysphemistic and misleading framing of what is in fact accurate, unbiased prediction, assuming recidivism rates are in fact unequal. Algorithms can be unbiased, and sometimes should be above criticism because of this.

6. Blaming the algorithm in this case obscures a more pervasive social and legal problem that can be solved without attempting to discredit sound inference and mathematical fact. A more productive approach to solving this problem than criticism statistical inference would be to address:

1. The causes of unequal recidivism rates. Social factors—such as lack of social support or unequal treatment by police--that are not the personal responsibility of defendants may contribute to the probability of recidivism.

2. The use of recidivism risk prediction in assigning outcomes such as bail and sentencing. The process that reproduces existing bias is the process that punishes those that already at higher risk of recidivism for reasons of social bias, not the algorithmic system that accurately predicts recidivism based on available information.

AI may be designed to address social inequality deliberately. Doing so might lead to AI that serves the public good, in the eyes of the progressively oriented public. But insisting that statistically unbiased predictive software is responsible for biased outcomes, as opposed to the uses of that software, obscures the true root of inequality, such as the causes of recidivism.

(7) the scientific and technical training that will be needed to take advantage of harnessing the potential of AI technology, and the challenges faced by institutions of higher education in retaining faculty and responding to explosive growth in student enrollment in AI-related courses and courses of study;

The core principles of AI are general and ubiquitous: probability theory, theory of computation, complexity theory. Since technology designed and limited by these principles affect every aspect of modern human life and especially democratic life, it is not adequate for education in these principles to be restricted to higher education.

* Basic probability theory, theory of computation, and complexity theory, as branches of mathematics, need to be considered part of data literacy and be part of the public education core curriculum.
* This curriculum should frame the importance of this mathematical understanding explicitly as a way of understanding intelligent systems used by e.g. governments.

References
==========

John Anderson. 1991. Is human cognition adaptive? Behavioral and Brain Sciences 14: 471–517.

Angwin, J., J. Larson, S. Mattu and L. Kirchner. (2016) Machine Bias. ProPublica May 23, 2016. Accessed 22 July 2016: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Beniger, J. (2009). The control revolution: Technological and economic origins of the information society. Harvard university press.

Chater, N., Oaksford, M. (1999) Ten years of the rational analysis of cognition. Trends in cognitive sciences 3 (2): 57–65. 10.1016/s1364- 6613(98)01273-x.

Chater, N., Vitanyi. P. (2003) Simplicity: a unifying principle in cognitive science? Trends in Cognitive Sciences, Volume 7, Issue 1, January 2003, Pages 19-22, ISSN 1364-6613, http://dx.doi.org/10.1016/S1364-6613(02)00005- 0.

Citron, D. K. (2007). Technological due process. Washington University Law Review, 85, 1249-1313.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O. and Zemel, R. (2012) Fairness through awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12). ACM, New York, NY, USA, 214-226. http://dx.doi.org/10.1145/2090236.2090255

Griffiths, T., Kemp, C., and Tenenbaum, J. (2008) Bayesian models of cognition. In Ron Sun (ed.), Cambridge Handbook of Computational Cognitive Modeling. Cambridge University Press.

Lessig, L. (2009) Code: And other laws of cyberspace. ReadHowYouWant.com

Pasquale, F. (2015). The black box society: The secret algorithms that control money and information. Harvard University Press.

Radin, M. J. (2004). Regulation by contract, regulation by machine. Journal of Institutional and Theoretical Economics (JITE)/Zeitschrift für die gesamte Staatswissenschaft, 142-156.

Russell, S. J., Norvig, P., Canny, J. F., Malik, J. M., & Edwards, D. D. (2003).Artificial intelligence: a modern approach (Vol. 2). Upper Saddle River: Prentice hall.

Tenenbaum, J., Kemp, C., Griffiths, T., and Goodman, N. (2011) How to Grow a Mind: Statistics, Structure, and Abstraction. Science 331 (6022), 1279-1285

Tetlock, P. (2005). Expert political judgment: How good is it? How can we know?. Princeton University Press.

Respondent 139

Susan Epstein, Hunter College / CUNY

By way of introduction, I am Professor of Computer Science at Hunter College and The Graduate Center of The City University of New York. I received my Ph.D. in computer science in 1983, and have published more than 130 papers on artificial intelligence (http://www.cs.hunter.cuny.edu/~epstein/). The National Science Foundation continues to be the principal funding agency for my work, which focuses on knowledge representation, machine learning, and cognitive architectures. I am known for my extensive interdisciplinary collaboration, and am a past chair of The Cognitive Science Society and a recent officer of ACM's SIGAI (Special Interest Group on Artificial Intelligence). My article on "Collaborative Intelligence" appeared in the archival journal Artificial Intelligence (http://www.cs.hunter.cuny.edu/~epstein/papers/AIJ%202015.pdf) in 2015.

I believe that artificial intelligence has significant potential for public good. The popular media, however, often attracts attention (and revenue) by frightening ill-informed citizens with monster-like robots, error-ridden systems, and economic turmoil. This document addresses several shortcomings that I believe must be addressed immediately, and over the long term, to support the American people.

The science of collaboration

AI's initial goal was to develop systems that would reason perfectly, that is, consider every option and prove carefully that the AI solution is in some sense optimal. When real-world problems proved intractable under this approach, the goal became systems that would be computationally rational, that is, select actions that maximize their expected reward. People do not need computers that play games better than they do; people need computers that help them retrieve and organize knowledge quickly and coherently, computers that identify and present anomalies and relationships that people may have overlooked in a sea of data, and computers that support and supplement human skills.

The best use of national resources, I believe, would support extensive research on how people and machines can best complement one another. People excel at some tasks, and computers at others. Collaborative intelligence is a framework in which people do what they do best, computers do what they do best, and they act jointly to achieve human goals. Commonsense reasoning requires vast knowledge, knowledge people derive from extensive

sensory exploration and observation in their rich, diverse world. No current computer is capable of such development, nor is there likely to be one in the foreseeable future. Moreover, it is difficult to extract commonsense knowledge from people on behalf machines. Instead, we should facilitate collaboration between people and machines.

A crucial issue here is the human's experience during such collaboration. People must trust computers if they are to work together with them. Thus computers must be able to explain how they reach their decisions, how they vet their knowledge, and why they select the actions they take or propose. Because increasing numbers of people will collaborate with computers, it is essential that communication from the computer take a human-like perspective. Communication should be both verbal and visual, through speech, text, sketches, and pictures. Imagine, for example, a robot and a person who are traveling together with an inadequate map. They both should be able to point at it, plan with it, and converse in language that, while occasionally stilted, is clear to and comfortable for the person.

Computers need to understand their human collaborators too. This would include the ability to sense and respond appropriately to human emotions, particularly frustration and bewilderment. Several architects recently explained to me that they had the facility to "run" simulated people through their building designs, but that they had been unable to find data on how people responded to signage, long-distance views, and novel environments. Human subject experiments belong in AI. Only recently, for example, have scientists discovered that less realistic avatars actually make people more comfortable. There is no reason to emulate the electro-chemical soup with which people reason; the computer should reason in ways that are efficient and effective for the task at hand.

People have little tolerance for collaborators that are fallible in ways they find "dumb". Examples of such collaboration-unready systems include picture taggers that are quickly befuddled by the same image when it is slightly shifted, and recommendation systems that fail to recognize that human users are quixotic and easily bored. Systems that offer a modicum of appropriate humor would also go far to smooth the initial awkwardness inherent in any new collaboration.

A science of collaboration requires a better understanding of what helps people be more productive when they partner with a machine. A science of collaboration would develop foundational methods for human-machine collaboration, including realistic environments outside a laboratory setting, introduction of the collaborators to one another in ways that facilitate their communication and understanding, preliminary discussion of a common vocabulary and the task before them, and statements about their individual strengths, weaknesses, and preferences. The ways someone best interacts with a computer will necessarily depend on the person, their common task, and the computational methods the machine relies on.

The description thus far has assumed a single person and a single computer. Additional and distinct challenges will arise when there are multiple computers and one person, multiple people and one computer, and a mixed human-robot team. Each of these areas requires, once again, human subject research.

Education to grow scientists

AI is firmly grounded in mathematics. AI practitioners need to understand why some methods will always be superior to others (e.g., algorithmic complexity) and how to evaluate the results of their work (e.g., experimental design and statistical analysis). As a computer science faculty member for more than 30 years, however, I have watched with increasing dismay as each wave of bright, hard-working, eager, US-born students arrives with less knowledge of mathematics than the contingent that preceded it. Meanwhile, no such degradation has been visible in my students whose early mathematics training was outside the US. American high schools and elementary schools and preschool programs are failing to prepare young scholars with the knowledge and respect for mathematics that they need to support AI development.

Because a diverse population will collaborate with AI systems, diversity among AI developers is also essential. Although AI needs their input and perspectives, women and minorities continue to be under-represented among both students and practitioners of AI. Negative attitudes toward computation and mathematics develop quickly in US female and minority children. I believe a crucial flaw in our educational system lies with early childhood educators who themselves dislike mathematics and science, and convey that implicitly. One way to address this issue is to train every teacher, particularly those who teach young children, in ways that will imbue them with enthusiasm for mathematics. Children model themselves on their heroes; they need to see their teachers, particularly women and minorities, enjoying mathematics and science.

Interdisciplinary education

In Spring 2016, I taught the inaugural version of SCI 10N01: Brains, Minds, and Machines, an interdisciplinary science course for first-year undergraduate students. It provides an integrated foundation in cognitive neuroscience, cognitive psychology and AI. The development of this course is supported by the Center for Brains, Minds, and Machines, sponsored by the National Science Foundation. SCI 10N01 will be a model for courses at several other schools in the next few years. To the best of my knowledge, this course is the first of its kind.

There are many challenges in interdisciplinary science courses, particularly at the introductory level. No single faculty member is adequately prepared to teach such a course, and institutional policy rarely facilitates inter-departmental team teaching. Most students, administrators, and pre-existing institutional software do not understand why the course

does not “belong” to a department. Although first-year students are the ideal target, graduation requirements and counselors discourage them from taking a course that does not ‘count,” so it is necessary to recruit students extensively. Nonetheless, all the students anonymously described the tremendous impact the course had, both on their perspective on intelligence and on their academic and career plans.

Although these students intend to major in a science (everything from physics to psychology), much of the material in the course is about perception and cognition. Every educated person needs to understand more about how her senses approximate the world and how her mind interferes with rational behavior. As they learned about how computers sense and reason, the students’ fear and discomfort with computers slipped away. They were primed to be collaborators with computers, and perhaps AI practitioners themselves. Further funding for such courses, and for institutional reform to support them, is essential. Science does not abide by walls; neither should science instruction.

Respondent 140

Mary Wareham, Human Rights Watch

Response to the White House Office on Science and Technology Policy regarding its Notice of Request for Information published on June 27, 2016

This submission is made by Human Rights Watch (www.hrw.org), a nonprofit, nongovernmental human rights organization pressing for changes in policy and practice that promote human rights and justice around the world. The submission is in regards to research and development of fully autonomous weapons. Human Rights Watch is the coordinator of the Campaign to Stop Killer Robots (www.stopkillerrobots.org).

Summary

Artificial intelligence and robotic autonomy have already had major impact on our lives, from simple processes like vacuuming to complex ones like self-driving cars and Google’s DeepMind project. However, no field of artificial intelligence raises urgent and serious human rights concerns more than the research and development of fully autonomous weapons. While none currently exist, these weapons raise serious moral and legal concerns because they would possess the ability to select and engage their targets without meaningful human control. Many people question whether the decision to kill a human being should be left to a machine. There are also grave doubts that fully autonomous weapons would ever be able to replicate human judgement and comply with international humanitarian law or international human rights law. These concerns are compounded by the obstacles to accountability that would exist for unlawful harm caused by fully autonomous weapons. For these reasons, Human Rights Watch has called for a preemptive international prohibition on the development, production, and use of fully autonomous

weapons.

International Humanitarian Law

The rule of distinction, which requires armed forces to distinguish between combatants and noncombatants, is a critical benchmark that is difficult for fully autonomous weapons to meet in order to comply with international humanitarian law. Fully autonomous weapons would not have the ability to sense or interpret the difference between soldiers and civilians, especially in contemporary combat environments. The requirement of distinction is arguably the bedrock principle of international humanitarian law. According to customary international law, articulated in Protocol I to the Geneva Conventions, combatants must "distinguish between the civilian population and combatants." Attacks that fail to distinguish are indiscriminate and unlawful.

International humanitarian law also prohibits disproportionate attacks, in which civilian harm outweighs military benefits. This requirement is one of the most complex rules of international humanitarian law, which requires human judgment that a fully autonomous weapon would not have. Determining the proportionality of a military operation depends heavily on context and it is highly unlikely that a robot could be pre-programmed to handle the infinite number of scenarios it might face. As a result, it would have to interpret a situation in real time, likely interfering with its ability to comply with the proportionality test.

Lastly, weapons should require meaningful human control when used in order to be compliant with the principles of distinction and proportionality in international humanitarian law. The ability to distinguish combatants from civilians or from wounded or surrendering soldiers, as well as the ability to weigh civilian harm against military advantage, require human qualities that would be difficult to replicate in machines, including fully autonomous weapons. Determining whether an individual is a legitimate target often depends on the capacity to detect and interpret subtle cues, such as tone of voice and body language. Humans usually understand such nuances because they can identify with other human beings and thus better gauge their intentions.

International Human Rights Law

Meaningful human control is also crucial to compliance with human rights law. In addition to undermining human dignity, lack of control would threaten the right not to be arbitrarily deprived of life. Whether in a law enforcement or an armed conflict situation, upholding that right depends on human qualities of perception and judgment that are difficult to replicate in machines yet essential to assessing the necessity of force.

In addition to the problems regarding the use of fully autonomous weapons on the battlefield, once available, this technology could be adapted to a range of other contexts that

can be grouped under the heading of law enforcement. For example, local police officers could potentially use such robots in crime fighting, the management of public protests, riot control, and other efforts to maintain law and order. State security forces could employ the weapons in attempts to control their opposition. Countries involved in international counter-terrorism could utilize them in scenarios that do not necessarily rise to the level of armed conflict as defined by international humanitarian law. Some law enforcement operations have legitimate ends, such as crime prevention; others, including violent suppression of peaceful protests, are inherently illegitimate. Fully autonomous weapons could be deployed in an operation regardless of its character and indeed, may encourage the use of force as they will replace the human officers that otherwise would have to be deployed.

The use of fully autonomous weapons in a law enforcement context would trigger the application of international human rights law. International human rights law applies in both peace and war, and it regulates the use of force in situations other than military operations and combat. In comparison to international humanitarian law, which governs military operations and combat and applies only during armed conflict, human rights law tends to have more stringent standards for regulating the use of lethal force, typically limiting it to where needed to defend human life and safety. Therefore, the challenges of developing a fully autonomous weapon that would comply with international law and still be useful are even greater when viewed through a human rights lens.

Fully autonomous weapons threaten to contravene foundational elements of human rights law. They could violate the right to life, a prerequisite for all other rights. Deficiencies in judgment, compassion, and capacity to identify with human beings could lead to unjustified killing of civilians during law enforcement or armed conflict operations. Fully autonomous weapons could also cause harm for which individuals could not be held accountable, thus undermining the right to a remedy. Robots could not be punished, and superior officers, programmers, and manufacturers would all be likely to escape liability. Finally, as machines, fully autonomous weapons could not comprehend or respect the inherent dignity of human beings. The inability to uphold this underlying principle of human rights raises serious moral questions about the prospect of allowing a robot to make the choice to take a human life.

Accountability

The hurdles to accountability for the production and use of fully autonomous weapons under current law are monumental. The weapons themselves could not be held accountable for their conduct because they could not act with criminal intent, would fall outside the jurisdiction of international tribunals, and could not be punished. There would also be insufficient direct responsibility for a human who deployed or operated a fully autonomous weapon that committed a criminal act, because fully autonomous weapons by definition would have the capacity to act autonomously and therefore could launch independently and

unforeseeably an indiscriminate attack against civilians or those hors de combat. Criminal liability would therefore likely apply only in situations where humans specifically intended to use the robots to violate the law. In the United States at least, civil liability would be virtually impossible due to the immunity granted by law to the military and its contractors and the evidentiary obstacles to products liability suits.

While proponents of fully autonomous weapons might imagine entirely new legal regimes that could provide compensation to victims, these regimes would not capture the elements of accountability under modern international humanitarian and human rights law. For example, a no-fault regime might provide compensation, but since it would not assign fault, it would not achieve adequate deterrence and retribution or place moral blame. Because these robots would be designed to kill, someone should be held legally and morally accountable for unlawful killings and other harms the weapons cause. The obstacles to assigning responsibility under the existing legal framework and the no-fault character of the proposed compensation scheme, however, would prevent this goal from being met.

Moral and Ethical Issues

Fully autonomous weapons also raise serious concerns under the Martens Clause. The clause, which encompasses rules beyond those found in treaties, requires that means of warfare be evaluated according to the "principles of humanity" and the "dictates of public conscience." Both experts and laypeople have expressed a range of strong opinions about whether or not fully autonomous machines should be given the power to deliver lethal force without human supervision. While there is no consensus, there is certainly a large number for whom the idea is shocking and unacceptable.

In May 2014, fourteen Nobel Peace Laureates called for a preemptive ban on fully autonomous weapons. The Laureates feared what might happen to essential protections to civilians if these weapons were deployed and the arms race that would likely spawn from their use. Furthermore, in November 2014, more than 120 religious leaders, including Archbishop Desmond Tutu, signed an interfaith declaration calling on states to work towards an international ban on fully autonomous weapons. Lastly, in July 2015, over 1,000 roboticists and more than 3,000 artificial intelligence searchers, scientists, and related professionals signed an open letter calling for a preemptive ban of "offensive autonomous weapons" in part because of the implications of their use. The list of signatories includes Elon Musk, Demis Hassabis, Barbara Grosz, Stephen Hawking, and Steve Wozniak.

Conclusion

Fully autonomous weapons have the potential to increase harm to civilians during armed conflict. They would likely be unable to meet basic principles of international humanitarian

law and international human rights law, would pose monumental challenges for accountability, and illicit vocal condemnations from moral, technological, and religious leaders. Although fully autonomous weapons do not exist yet, technology is rapidly moving in that direction. Therefore, as the Office of Science and Technology Policy forms its view on artificial intelligence, we urge you to recognize the multiple challenges with regards to the development and use of fully autonomous weapons and support a preemptive prohibition on their development, production, and use.

Respondent 141

Valerian Harris, Cagemini Americas Inc.

1. social and economic implications:
The immediate impact of AI is loss of jobs. With the current focus on automating repetitive tasks nearly all administrative jobs will be performed by an algorithm.
This loss of job will have implications on crime, human health, social security. This disruption of white collar jobs will re-define the middle-class within a very short time.

2. the most important research gaps in AI that must be addressed to advance this field and benefit the public;

Current set of designers and developers are focusing on developing algorithms to automate the known tasks. All these algorithms are nearly 50 years old. there is hardly any new thinking on the possibilities with such large amounts of data. Especially unlabeled data from social media. The focus should be on developing algorithms for un-supervised learning for preventive and predictive techniques for health-care, crime, traffic management etc.

Policy Making:
- All the current efforts on AI come from "FOR PROFIT" organizations. If Vision and Funding is provided, similar to a Space Program, then we will see the benefits of AI leading towards job creation and improving the society

Respondent 142

Henry Lieberman, MIT

First of all, we'd like to thank OSTP for making this request. In our 40 years in the AI field, this is the first time we've heard the government directly ask the scientific community for their opinion on this important topic. Since the call said you'd "appreciate brevity", we'll start with our most important messages:

• Recent dire warnings by well-known figures such as Elon Musk and Steven Hawking of the "dangers of runaway AI" are overblown. While research into AI safety makes sense, government should not view AI as an existential threat, in the same category as things like

climate change. (Question 3)

• We now have government and corporate structures that were invented around the time of the Industrial Revolution. AI, in conjunction with personal manufacturing, hold the promise of a radical transformation of the economy, for the better. We have some proposals for reinventing the economy for the Information Age. (Questions 1, 2, 4, 9)

• Will robots take our jobs? Yes. Is that something to fear, or to try to stop? No. Should we bring back manufacturing to the US? Yes. But retraining workers and calls to "bring back manufacturing *jobs*" are at best, short-term measures. The real solution is to rethink the notion of a "job", itself a creation of the Industrial Revolution era. (1, 2, 4, 9)

• Government should resist the temptation by the military and intelligence agencies to start a new arms race in cyberspace. Computer break-ins and malware can cause harm, but overzealous "defense" will make the problem worse, as well as consume resources that could be used for positive social good. (1, 2, 3, 4, 9)

• Make America less competitive. Yes, you heard that right. We said less, not more. In general, technology will increase the benefits and reduce the cost of cooperation. Conversely, it will reduce the benefits and increase the cost of competition. (1, 2, 3, 5, 9)

• AI can serve a "reasoning assistant" to guide decision makers, just as an accountant uses a spreadsheet as a math assistant. When a group is faced with a complex decision, managing the rationale for various solutions gets to be so daunting that many people just throw up their hands and go with their gut, leading to sub-optimal choices. AI can help guide decision makers through intricate solutions while mitigating unintended consequences. (1, 2, 3, 4, 7, 8)

Now, we'll present our arguments for these positions.

• Dangers of AI. A recent public letter by prominent scientific leaders has warned about the possibility of "runaway AI": a situation where AI becomes smarter than humans and uncontrollable, and fights against people. While this is a theoretical possibility, we think this is as unlikely as Mary Shelley's Frankenstein. The problem with the Frankenstein scenarios is that they assume there is technological progress sufficient to create an AI, but they don't assume any progress in controllability, or in psychology or social relations.

AI is not just about "intelligence", but also about computational understanding of other human traits like emotion, motivation, and cooperation with other agents. One of the benefits of AI is that we will learn more about these traits in people, helping us deal with problems like mental illness and crime. By the time we get to strong AI, we will also figure out how to program it so it won't have the insane aggressiveness of the James Bond villain. The AI's will be smarter than that.

Certainly we should have research into AI safety, just like we have research into automobile safety, gun safety, or aviation safety. We need considerably more research into software safety. Debugging is the science of how to detect and repair errors in computer programs. The best route to making progress in the neglected area of debugging, is actually to use AI techniques. So it is especially important that we don't let fear of AI prevent us from developing the very techniques that will assure a positive outcome for AI.

• AI and the economy.
AI holds the promise of being able to solve virtually all of our economic problems. Automation has always increased the average productivity per person, and we expect this to continue, to the point where almost any job could be automated, and everybody's material needs can be provided for. Personal manufacturing, starting with advanced 3-D printing, will obsolete most large-scale factories, exactly like personal computers replaced large-scale mainframes. Advances in AI programming environments will give individuals the ability to write complex software for themselves, rather than be passive recipients of industrial production of software.

This means that we have to start a long-term transition out of Industrial Revolution era capitalism, to what we call Makerism. Makerism puts the means of production in the hands of individuals, and small cooperative groups. AI can help with the coordination required, thereby disintermediating inefficient bureaucratic organizations.

We realize that this is a radical shift, and we don't expect it to happen quickly. But delays will likely lead to more human suffering. Government can encourage research and facilitate this transition. We also realize that many elements of the status quo will fight the transition to Makerism in the short term, because they perceive it to be disadvantageous to their own position. But in the medium term, Makerism will help everyone: the poor with more wealth, the rich with a healthier overall society.

Imagine you were in the feudal era, but you knew that the Industrial Revolution was going to happen soon. Could you have figured out a way to help it happen, without having the society have to go through the robber-baron and sweatshop era?

• Jobs.
A common fear people have is "Will I lose my job?"; "What will I do?". Let's remember that the idea of a "job" itself was only a creation of the Industrial Revolution era; it didn't really exist before that, and probably won't after. People will still engage in productive and meaningful life activity even if we don't have to structure it as a conventional job. It's only that we won't have the coercion, regimentation, and exploitation that now characerizes employment.

Let's divide the job issue into two questions. How will people provide for their needs? How

will they decide how to spend their lives in a meaningful way? Jobs today provide the first, and (only for a lucky few) the second.

To answer the first, remember we're positing that per-capita productivity will increase to the point that the sum total of productivity can provide for everybody's needs. We do have enough raw GDP to give everyone a middle class lifestyle now, but our political and economic structures favor inequality to the extent that we can't make it happen. The biggest benefit of Makerism might be its ability to reduce inequality. It's the answer to the question, "If AI is so great, how can it solve poverty?".

One idea could be a Universal Basic Income, an idea supported by many economists, including conservatives. Gradual reduction of work week hours, gradual lowering of retirement ages, etc. could be ways to phase it in. There might also be private-sector ways of accomplishing it. It would eliminate poverty. Poverty now has so many negative consequences, that we can definitively say: Poverty is so expensive, we can't afford it.

The issue of what people will do in the absence of jobs has more diverse answers and takes more discussion, which we address in our book, referenced below. What makes jobs personally meaningful is interaction with co-workers, a feeling of helping people, work as a means of self-expression; there are other ways to accomplish these aims.

We know this is hard to imagine in today's world of scrambling for scarce jobs, inadequate salaries, tight government budgets, etc. etc. But AI can help us do precisely the things that will get us there: increase productivity, eliminate war, invest in education and health.

• Cyberwar.
One of the biggest dangers we see for government at the moment is that military and intelligence agencies will use the fear of computer attacks and malware to start a new arms race in cyberspace. They have a vested interest in doing so, since they will be blamed for inability to counter an attack, but they have no responsibility for the negative consequences caused by their activities. Don't take the bait.

Because there is no perfect defense, the military will always advocate more defensive and aggressive measures. Others will interpret (or misinterpret) these measures as hostile actions, respond in kind, and it will spiral out of control. That's much more of a real threat than an independent AI spiraling out of control. Because defense is expensive, it sucks resources that could be used for positive developments, and again the security and defense people don't bear the opportunity cost.

The Internet relies on cooperation and trust, and overzealous security can easily destroy the trust that makes it work. A distrustful Internet makes the cooperation we need to develop AI's future, impossible.

Malware is indeed a problem. Government should support research in security, especially the usability problems with today's dismal security measures. One of the biggest reasons today's computer systems are vulnerable is because many systems are programmed with antique programming languages that have problems like "memory leaks" that make them crackable. The solution is to support research into better high level programming languages and AI techniques like "actor models of computation" that are not so vulnerable.

• Cooperation and Competition.
In our book, we present a scientific argument, combining elements of game theory and evolutionary theory, to say that technological advances, including and especially AI, affect the classic tradeoff between cooperation and competition. Traditional arguments like "competition brings out the best" or "competition motivates people" are becoming less and less true. Cooperation has nonlinear "network effects" that increase its benefits.

That's why it's especially important that we do everything possible to reduce the competitiveness of most aspects of society, from partisan gridlock in government, to commerical competition, to international trade and power disputes, to wars. Technology both makes cooperation easier, and increases the benefits, as we explain in the next section.

That's why we cringe when we hear, "Let's make America more competitive". Let's make it less.

• AI assistance for decision making.
Governments have two primary barriers to making optimal choices for the governed. One is that, even in democracies, they are fundamentally power based. Power might be due to number of votes gathered, seniority, money, or other forms of political capital. But even if we could decrease the corruption of power, governments face a second barrier: complexity. Pretty much any issue that's important enough for a Congress to handle is too complicated for individuals to understand. That lack of understanding drastically limits the solution space because legislators won't back a plan they don't understand (for good reason).

We hit both barriers head on with a new process we call "Reasonocracy". It displaces "he with the most power decides" with cooperative reasoning. The core issue with reasoning is that, although it can be broken down into a number of understandable steps, when there's a lot of steps, agrigating those steps into a coherent whole can't be held in your head. Just as the accountant *could* add up 100 numbers by hand, it would be time consuming and error prone, so they use a tool, a spreadsheet. We, and others, have produced tools that act as a spreadsheet for reasoning, a tool that can add up rationale into a well-reasoned solution.

An additional problem with "cooperative reasoning" is the "cooperative" part. We can hold a discussion or a debate, but at the end we're left with an overall impression, or the last things someone said or whichever idea was shouted loudest or is most "sound bitable".

Even if we are allowed to review the meeting minutes before deciding, chronological ordering of spoken ideas suffers from vaugue ideas and incoherent ordering of relevant ideas.

Reasonocracy needs AI tools at its core, but requires much more in terms of civic engagement process. This requires a sophisticated educational program, (another project that can benefit from AI, with Moocs and intelligent tutoring systems) and additional clever ways to manage the complexity of agreggating and disseminating the creativity of the nation.

About the Authors

Much of the material presented above is from our forthcoming book, entitled "Why Can't We All Just Get Along?", by Christopher Fry and Henry Lieberman. We would be happy to share a draft or relevant material from it if you have interest.

Henry Lieberman (XXXXXXXXX) is a Research Scientist at MIT, who is now launching the Marvin Minsky Insitute for AI at the Computer Science & Artificial Intelligence Lab (CSAIL). His academic area is Intelligent User Interfaces, combining AI and Human-Computer interaction. He also ran the Software Agents group at the MIT Media Lab. He served on the board of directors for AAAI (the professional organization for AI), and was twice chair of the ACM Intelligent User Interfaces Conference. He's published over 120 articles, and three books. His doctoral-level degree is from the University of Paris, where he was also a Visiting Professor.

Christopher Fry (XXXXXXXXX) is an independent researcher who has worked at the MIT Media Lab, MIT Sloan Business School, IBM, Bolt Beranek & Newman and other AI companies, and helped launch several AI-related startups. He attended Berklee School of Music, and did one of the seminal computer music composition systems. His design for future intelligent transportation systems was presented at the US House of Representatives and the United Nations.

Respondent 143

Rand Leeb-du Toit, EXOscalr

Thank you for the opportunity to provide input. I am a special adviser to numerous companies on future technologies and trends, including the field of AI and smart machines.

I would like to address question (6) - The most important research gaps in AI that must be addressed to advance this field and benefit the public good:

We are faced with a pandemic of fear and disengagement. People are living a fundamentally meaningless way of life. This is driven by a paucity of collective human empathy and wisdom. Causally, this pandemic is exacerbated by the fact that, as humans, we are affected by all our interactions. And yet the systems, tools and environments that we interact with and within on a daily basis are sterile of emotion, do not engender empathy and fail to progress our collective wisdom.

The most important research gap in AI revolves around tackling this pandemic. By focusing on creating an empathetic symbiosis between human and emotionally smart machine in a wisdom-generating system we increase the probability that AI will benefit the public good.

Respondent 144

Grant Soosalu, mBIT International

Topics 3 & 6: Currently AI is modeled on the human head brain alone, however recent neuroscience findings have uncovered that humans have complex adaptive and functional neural networks in the heart and enteric (gut) regions that are deeply involved in embodied cognition. Sociopaths are largely head brain driven and have functional disconnection from these heart and gut neural networks. It is therefore vitally important that AI's be designed and implemented with the modeling of heart and gut brains and their core competencies and prime functions so that we don't produce AI's that are sociopathic in their underlaying ontology. Their is emergent work on AI that is heart-based and compassion driven. This work should be given more funding and focus. For more information on the human heart and gut brains, on their core competencies and highest expressions, see the work of Soosalu and Oka, www.mbraining.com

Respondent 145

Andrew Kim, Google Inc.

We applaud the White House's Office of Science and Technology Policy (OSTP) for convening an important national dialogue on artificial intelligence and appreciate the opportunity to provide input based on our experience working with the technology.

Google's mission is to organize the world's information and make it universally accessible and useful. In pursuing this mission, our company has always been excited by the promise of artificial intelligence and the real-life benefits it can impart to society. Driven by significant advances in research and computing power, this technology - particularly in the application of machine learning - is increasingly becoming a key feature of our products.

We write to outline the opportunities that we perceive in the technology, some existing challenges and our approaches to them, and opportunities for government to be a catalyst for enabling this technology to reach its fullest potential.

Where We Are

Artificial intelligence is not a speculative science-fiction technology but a practical software engineering tool already being used to help millions of people around the world every day. Machine learning, a field within artificial intelligence that specifically studies algorithms that learn from data, is already benefiting many of Google's products. The field of robotics, often lumped into these discussions, deals with mechanical systems that sometimes, though not necessarily, incorporate techniques of machine learning or artificial intelligence. Our submission focuses on machine learning, though these points could be extrapolated to artificial intelligence more broadly.

Recent breakthroughs in machine learning have been decades in the making. Many contemporary developments in machine learning derive from the results of government-funded basic research in the 1980s and 1990s, which are only now becoming practical because of the availability of computational power, richer sources of information about the world, and a growing community of talent.

In our products, machine learning advances have improved efforts ranging from user protection (e.g., spam and malware filters) to accessibility (e.g., voice recognition). For example, using a method that learns language from patterns in bilingual text, Google Translate translates more than 100 billion words a day in 103 languages. With Google Photos, you can search for anything from "hugs" to "border collies" because the system uses our latest image recognition system to automatically categorize objects and concepts in images. We also recently announced that machine learning is helping to reduce the environmental impact of our data centers.

Beyond enhancing existing products, these technologies will drive efficiencies and may dramatically improve society's ability to tackle some of our most pressing global challenges in health, environment, transportation, and beyond. In research to be published soon, Google has shown how machine learning can make the diagnosis of diabetic retinopathy - one of the fastest growing causes of blindness worldwide - more broadly accessible. Researchers are exploring the use of machine learning on topics ranging from weather prediction and genetic diseases to conservation and economic forecasting. Technologists are just starting to build out practical use cases, and we are excited to see the scientific community study and deploy these technologies in other fields.

While the recent pace of development has been rapid, neither the advancement of machine learning nor its impact on society is preordained. We need to address several issues in order to maximize the benefits of this technology for everyone.

Three Near-Term Challenges for Machine Learning

Many discussions about the potential benefits and consequences of machine learning remain speculative and focused on potential long-term implications and theoretical edge cases. Many research questions need to be addressed before society comes to confront these hypothetical questions.

Google is focused on three near-term challenges that we believe need to be made a priority. We believe that positive work in these areas will help machine learning advance and see even more widespread application over the next five to ten years.

Preserve Open Research Norms and Practices

Machine learning has flourished in part because of a set of common norms that encourage research results to be published and shared openly. It is important to preserve these community principles towards openness that have proven important to past work in the space.

In this vein, Google has been an active and open contributor to the research community. We have published results and actively participated in conferences on a variety of topics including large-scale deep learning, computer vision, sequence-to-sequence modeling, and visualization of the internal processes of neural networks. We also recently open-sourced TensorFlow - Google's internal machine learning toolkit - to allow anyone to experiment in the space and advance the state of the art.

Focus Research on Tangible Problems

As with any technology, it is important to maximize the positives and minimize concrete harms. The rapid advance of the field of machine learning has raised concerns surrounding the safety of implementing these systems in a variety of different contexts. Alongside this has been the recognition that machine learning systems trained on biased data may themselves render biased or discriminatory outcomes, or that such systems, by focusing on more objective criteria, might help reduce or avoid discrimination.

No system is perfect, and errors will emerge. However, advances in our technical capabilities will expand our ability to meet these challenges.

To that end, we believe that solutions to these problems can and should be grounded in rigorous engineering research to provide the creators of these systems with approaches and tools they can use to tackle these problems. "Concrete Problems in AI Safety", a recent paper from our researchers and others, takes this approach in questions around safety. We also applaud the work of researchers who - along with researchers like Moritz Hardt at Google - are looking at short-term questions of bias and discrimination.

Potential harms are not just matters of research. As a recent ProPublica investigation into

machine learning used by the judicial system illustrated, partial or biased data can produce discriminatory results as machine learning algorithms draw incorrect inferences from the examples they are trained on. This points to the need for improved tools for diagnosing these failures, as well as the need to avoid data gaps where the dearth of good data can make the use of machine learning problematic. Particularly where these systems are deployed in public administration, we believe the federal government and the international community can and should promote the release of complete, high quality, and robust datasets that enable responsible analysis and use.

The cycle of innovation and improvement is a continuous process. Moreover, these issues are diverse and we should not expect that they will be resolved through a simplistic "one size fits all" solution.

Diversify the Community Working on Machine Learning

Machine learning can produce benefits that should be broadly shared throughout society. Having people from a variety of perspectives, backgrounds, and experiences working on and developing the technology will help us to identify potential issues.

Continued investment in computer science education will help support this goal. Google is working to expand computer science education in a way that engages and retains students from all backgrounds. This includes programs that increase access and exposure, including initiatives like CS First, Made with Code, Exploring Computational Thinking, and the Computer Science Summer Institute, and our support of innovative organizations through the RISE Awards. Broad dissemination of the know-how around machine learning is critical as well, since it provides resources for anyone interested in learning more. We're helping to further this goal through our support of massively open online courses with leading researchers like Geoff Hinton.

The Government as a Catalyst for Development

We commend the White House for its efforts to advance the public discourse around machine learning. A broader understanding of the technology is an important part of appropriately identifying opportunities where machine learning will have the biggest and most positive impact.

We believe that government has a role to play in catalyzing research and efforts that meet the challenges outlined above and in promoting the development and application of these technologies. Some examples:

Support Research: The federal government has traditionally played an important role in supporting long-term fundamental research. The government has and should continue to play that role with machine learning, supporting research into the novel application of these

technologies in meeting social challenges and addressing potential limits and shortfalls.

Convene Talent to Meet Social Challenges: Machine learning has proven to be an effective tool for making progress on complex problems at significant scale. The federal government faces these types of challenges in fields like energy, transportation, environment, urban planning, and public health. We believe that it can convene task forces to explore the use of these new technologies in these fields and improving the work of government. We also support efforts like those by the US Digital Service that seek to build technical capacity within government, and encourage institutionalizing such initiatives.

Leverage the Diplomatic Network: Machine learning is an international phenomenon, drawing on researchers and datasets from around the world. Successful development of machine learning depends on the continuation of pro-innovation legal regimes in fields such as copyright, where flexible, well-designed limitations and exceptions can spur the development of new technologies. We recommend expanding the Digital Trade and Digital Economy Officer programs to cover issues in machine learning, and would support the growth of these programs more generally. In addition, the government could press for existing international fora, including the G-20, to incorporate work to establish broad norms for research and development in this field as has been done in other areas.

Promote Education and Diversity: The government has encouraged public support for increasing access to and diversity in STEM education and careers. W should continue this effort to ensure that more students from more backgrounds have access to computer science education. We strongly support the Computer Science for All Initiative, and would encourage similar projects.

Ensure Flexibility: Despite many recent breakthroughs in machine learning, the field and its applications are still nascent. As these opportunities are still emerging, we encourage a cautious and nuanced regulatory approach that will allow innovative uses to flourish and reach their full potential. To the extent that new rules are needed to address safety, operations, or other product infrastructure areas, we support the approach of having expert agencies take the lead on regulation of specific uses in their areas. In consumer protection areas like privacy, it will be important for existing agencies to maintain a harmonized approach as they assess whether new rules are needed and, if so, how they should be integrated with existing approaches developed over time. We also believe consensus-driven best practices and self-regulatory bodies will play an important role in ensuring the flexibility necessary to drive innovation while simultaneously developing nuanced and appropriate safeguards.

Although we have focused on a number of near-term considerations, we believe that the opportunities and challenges are significant, and warrant continued discussion as research and development progresses.

We appreciate having the opportunity to submit our views to this request, and are available if there are any further questions or assistance we can provide.

Greg Corrado
Senior Research Scientist

Sarah Holland
Public Policy and Government Relations Manager

Google Inc. - July 2016

Respondent 146

Guruduth Banavar, IBM Corporation

Dear OSTP,

The IBM Response to this RFI is a 2000 word essay at the URL:

http://www.research.ibm.com/cognitive-computing/ostp/rfi-response.shtml

Please note that there are links at the bottom of each section in the essay for further information.

If you would like a PDF of the essay, or further information, please contact Guru Banavar at XXXXXXXXX

Thank you.

Sincerely,
Guru Banavar, IBM

Respondent 147

James Cooper, Program on Economics & Privacy, Antonin Scalia Law School, George Mason University

Comment
Office of Science and Technology Policy
Request for Information on Artificial Intelligence

July 22, 2016

James C. Cooper*
Associate Professor of Law and Director, Program on Economics & Privacy
Antonin Scalia Law School
George Mason University

Introduction

Artificial intelligence and machine learning (collectively, "AI") offer great promise in myriad fields. As AI increasingly is used to make decisions of consequence, its potential to impacts consumers' lives will grow.

This comment is directed at the policy considerations related to AI decision-making in the commercial sphere. It argues that when determining whether any government intervention is needed to address such potential impacts on consumers from AI decision-making, policymakers should keep in mind two considerations. First, interventions should focus on instances that have been shown to be harmful, or are likely to be harmful. Second, policymakers should consider the extent to which market forces are likely to ameliorate any concerns; private incentives exist to correct biases, and they may operate more quickly and more efficiently than government action.

A. Harms-Based Approach

When developing policy, a starting point should always be a benefit-cost framework, centered on actual or likely harm. Policy should not be based on hypotheticals or remote risks.

An approach that focuses on actual or likely harms offers at least three advantages over an ex ante regulatory approach, which proscribes or mandates certain practices. First, by focusing on harm, one can be sure that government action is actually providing consumers with some benefits. Requiring harm to trigger action at least guarantees that the necessary (but not sufficient) condition for intervention to provide net consumer benefits is met.

Second, heterogeneous consumer preferences and costs of precaution increase the social costs associated with a common standard. Third, a harm-based approach has the advantage of being more nimble than prescriptive rules, which would have to be reworked as technology or other market conditions change to alter the benefit-cost calculation with respect to certain AI practices. This consideration should weigh heavily, given the rapidly evolving nature of AI.

B. Identifying Harms

Policymakers should focus on conduct that causes, or is likely to cause, harm. One of the potential harms identified with respect to AI is biased decision-making. Before labeling the outcomes of AI decision-making harmful, however, one must be careful to take account of the context in which they occur.

One widely cited study, for example, found evidence that women were less likely than men to see an advertisement related to high paying jobs. The computers trained to appear as women in the study instead were shown a generic job posting service. The inference taken by the authors is that "discrimination in the normative sense of the word" was at work, and that it could work to "further the current gender pay gap." Such an inference, however, is based on incomplete information—one must consider the forces underlying the real-time auction to serve the particular ad. A key factor driving the paper's findings is likely to be the willingness of other advertisers to bid to show ads to the female visitors. As competition for the attention of females increases, so will the price per impression, which could be too high for the employment ad. That is, the difference in ad serving rates is likely to be an artifact of more bidders competing for women's attention than men, rather than evidence of underlying bias in the AI system.

Similarly, another widely cited study reports differential online pricing for office supplies based on zip codes as evidence of a problematic algorithm. In context, however, differential online pricing based on zip codes is common and much more likely related to heterogeneity in local costs and competition than underlying bias in an algorithm. A national chain has incentives to avoid having its ecommerce channel cannibalize its offline stores—which must respond to local supply and demand conditions. It accomplishes this procompetitive goal by equalizing offline and online prices to consumers across markets.

As these examples illustrate, policymakers should be hesitant to intervene when there is a beneficial (or benign) business explanation for the observed phenomena.

In addition to considering context, policymakers should focus on situations in which the output of classification from AI is likely to be harmful. Differences in online ad serving are unlikely to place severe constraints on opportunity sets, as they are only one source of information among many. For example, it is highly doubtful that a woman seeking employment will limit herself to opportunities presented in online advertisements on one particular web site. Similarly, differential pricing is unlikely to pose problems for consumers, and in many cases is likely to increase consumer welfare, especially for relatively economically disadvantaged populations.

C. Consider Market Forces

To the extent that firms are employing AI techniques that erroneously offer different commercial opportunities to different classes of consumers, policymakers must consider

the competitive environment when deciding whether intervention is appropriate. Private incentives exist to correct biases—biased AI decision-making that erroneously limits commercial options to certain disadvantaged populations represent a profit opportunity. For example, when a subprime automobile dealer was able to correctly distinguish systematic from transitory high-risk individuals, it was able to increase its profit while also increasing the amount of credit available to those who were relatively more credit worthy. And competitive forces are likely to operate more quickly and more efficiently than government action to ameliorate such problems.

Relatedly, it is also crucial to distinguish between commercial and government use of AI decision-making. Unlike commercial entities that may use AI to incorrectly classify certain populations, governments are not subject to correction in the marketplace. Thus, governmental use of biased AI to make decisions regarding liberty or other fundamental rights, such as in criminal sentencing, are much harder to detect and correct. Accordingly, governmental uses of AI should be the primary focus of policy concern.

Conclusion

AI offers great promise. As policymakers consider approaches to AI, they should focus on practices that are likely to be harmful to consumers, and ones that are unlikely to be corrected by market forces.

* This comment reflects the views of the author only. Affiliation is for identification purposes only.

See James C. Cooper, Separation, Pooling, and Big Data, at 41-43 (April 2016), at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2655794.

Amit Datta et al., Automated Experiments on Ad Privacy Settings, 1 PROCEEDINGS ON PRIVACY ENHANCING TECHNOLOGIES, 92 (2015), at https://www.andrew.cmu.edu/user/danupam/dtd-pets15.pdf.

Id. at 105.

Indeed, the authors recognize as much. See id. at 93 ("Without knowledge of the internal workings of the [Internet advertising] ecosystem, we cannot assign responsibility for our findings to any single player within it nor rule out that they are unintended consequences of interactions between players.").

Jennifer Valentino-Devries, Jeremy Singer-Vine, and Ashkan Soltani, Websites Vary Deals and Prices Based on Users' Information, WALL ST. J, (Dec. 12, 2012) (finding that differential online pricing based on zip code leads to those in relatively poorer zip codes to pay more).

Indeed, the miniscule clickthrough rates—averaging around .1% in he U.S.—demonstrates how unlikely this scenario is. See David Chaffey, Display Advertising Clickthrough Rates (Apr. 26, 2016), at http://www.smartinsights.com/internet-advertising/internet-advertising-analytics/display-advertising-clickthrough-rates/.

See Executive office of the President, Differential Pricing at 17 (Feb. 2015), at https://www.whitehouse.gov/sites/default/files/whitehouse_files/docs/Big_Data_Report_Nonembargo_v2.pdf. The Antitrust Division has not brought a Robinson-Patman case since the 1960s, and the FTC has brought only one Robinson-Patman case since 1992. See ANTITRUST MODERNIZATION COMMISSION, REPORT & RECOMMENDATIONS at 318 (2007). Further, the bi-partisan Antitrust Modernization Commission recommended the repeal of the Robinson-Patman Act, concluding:

[S]eventy years after passage of the Robinson-Patman Act, courts remain unable to reconcile the Act with the basic purpose of antitrust laws to protect competition and consumer welfare. . . There is no point in further efforts to reconcile the Act with the antitrust laws in general; the Robinson-Patman Act instead should be repealed.

Id. at 322.

See FTC, BIG DATA: TOOL OF INCLUSION, Statement of Commissioner Maureen K. Ohlhausen, at A1-A2, (Jan. 2016), at https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf.

See Liran Einav, Mark Jenkins & Jonathan Levin, The Impact of Credit Scoring on Consumer Lending, 44 RAND J. OF ECON. 249 (2013).

See Julia Angwin et al., Machine Bias, PROPUBLICA (May 23, 2016), at https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Respondent 148

Byron Galbraith, Talla

AI will have a major, transformative, and positive impact on Education.

This response addresses three of the topics posed in the White House's Request for Information on Preparing for the Future of Artificial Intelligence: (2) the use of AI for public good, (3) the safety and control issues for AI, and (7) the scientific and technical training that will be needed to take advantage of harnessing the potential of AI technology. These topics are all addressed in the context of Education, with a specific focus on the upcoming role of AI in K-12 classrooms.

Topic (2): AI for public good – AI will enable and empower students and teachers, especially in low-income and underserved areas.

AI embedded in learning software and accessible via natural language interfaces will simultaneously empower students in learning traditional topics through individualized custom learning curricula, but also drive focus on 21st century skills seen as vital to modern

education. AI will not, however, replace teachers in the classroom, just as B.F Skinner's teaching machine, radio, cable television, personal computers, the Internet, and smart devices have also all failed at replacing teachers. AI will augment teachers, allowing them to focus more on roles as instructional coaches and facilitators of learning.

We are already seeing examples of this in the form of Intelligent Tutoring Systems (ITS), which can be seen as simple AI that reside inside learning software. ITS monitor student behavior such as analyzing how they answer questions or watching what choices they make inside a game-based learning environment. This analysis is used to adaptively adjust the scaffolding around the intended learning objectives, e.g. presenting additional guidance prompts or increasing the level of question difficulty. It can also be relayed to teachers in order to track how students are doing and alerting them to when students need additional assistance.

While these are very promising, a significant limitation of these AI is that they require the usage of some form of specialized computer software, such a reading comprehension game. Access to computer and Internet resources by low-income families is frequently not guaranteed, either at home or at the schools predominantly serving them. Non-profit institutions are making excellent progress here, validating these methods, but as they largely rely on government and foundational grants, those solutions have difficulty scaling beyond the lifetime of the particular funded project.

An exciting potential solution to the problem of computer resource access is coming in the form of AI-powered chat bots – intelligent agents that you can communicate with via natural language. While these AI can also have great power embedded inside a domain-specific application, such as an interactive helper inside a physics simulation environment, they can also be accessed via standard messaging systems, such as SMS. Mobile phones and smart devices are much more prevalent than desktop or laptop computers. Having access to a tutor you can chat with at any time of day about various topics will have a much greater impact on the underserved populations, as access is not gated as much by privilege.

This potential boon to Education does come with some real concerns about security and privacy.

Topic (3): The safety and control issues for AI – AI-initiated decisions must be transparent to students, parents, and teachers, while administrators must be confident that proper student privacy regulations are followed.

AI are trained by practitioners and on existing data sources, both of which have inherent, systemic bias. Understanding those sources of bias and how they affect the decisions that AI make are important, especially when considering deploying those AI to a diverse set of socioeconomic and demographic populations. If an AI in a learning software application or chat bot interface collects data and makes decisions about students, it is very important that

those individuals affected by an AI-initiated decision have a way to understand what reasoning led to that outcome, especially if it is negative or deleterious. This can be a complex and difficult undertaking, as many of the algorithms used in AI are difficult for even practitioners and experts to fully elucidate.

On the other side, both school administrators and vendors of products targeting K-12 must have guidance on how interactive AI, especially chat bots, are exposed to child online protection regulations like the Children's Online Privacy Protection Act and California's Student Online Personal Information Protection Act. Vendors must ensure that if their AI products are collecting data from children, that data is stored, encrypted, and access restricted according to all necessary guidelines.

The future prospect of AI in the classroom is bright, but there is still quite a long way to go in developing the technologies needed to realize that potential.

Topic (7) The scientific and technical training that will be needed to take advantage of harnessing the potential of AI technology – The future of educational AI is massively multidisciplinary.

In order to fulfill the vision and potential of natural language-powered AI for Education, we must have greater expertise with the computational processing of written language, especially in conversational settings. This includes understanding and anticipating user intent, translating that intent into actions, and generating acceptable natural language responses in return.

Built largely from the domains of computational, mathematical, and linguistic training thus far, these fields are required more than ever, but by themselves are insufficient for developing advanced AI, as a tremendous amount of natural language interaction in conversation relies on multiple levels of context and assumed shared experience. Psychology, sociology, and anthropology, therefore, are also going to need representation in the development of AI. Finally, expertise in Education and Learning are needed to craft the roles AI will place in the classroom and ensure the affordances of AI are exploited to maximize their effectiveness.

Respondent 149

Anthony Aguirre, Future of Life Institute

Future of Life Institute Response to the White House RFI on AI

NOTE: REVISED VERSION

We thank the OSTP for providing this opportunity for stakeholder input into the OSTP's

thinking about, and planning for, the potentially large impact AI will have in the coming decades. The Future of Life Institute, with its mission of increasing the odds of a highly positive long-term future of humanity, has focused a great deal on AI, and how we should endeavor to keep it robust (doing what we want it to do) and beneficial.

Regarding (2) the use of AI for public good:

Our view is that in the short term AI, like other information technologies, will serve as a powerful tool that can be used by corporations, governments, organizations, and individuals to accomplish their goals. As AI increases in effectiveness, it should provide ever-stronger levers to enhance human capability in diverse fields including scientific research, engineering, data analytics, strategy and planning, legal analysis, etc., etc. This could enable accomplishment of many widely desirable goals (for example curing the majority of diseases, finding mutually beneficial paths in geostrategic analyses, developing clean energy, and finding ways of safely stopping deleterious anthropogenic climate change.) In the longer term, we may well cross a threshold in which AI systems transition from being tools for humans to accomplish their goals, to agents that accomplish goals furnished to them by humans. In this case, as is discussed below, it is quite crucial that these goals are indeed for "the public good" and that they are accomplished by AIs in a manner that is also consistent with the public good.

Regarding (3) the safety and control issues for AI:

Historically, practitioners in mainstream AI have rightly focused on improving AI's pure capacity: its modeling capacity and its possible range of actions. As AI becomes more powerful, if society wishes to benefit from AI, we must broaden this focus to include building a clear understanding of how to make AI not just good at what it does, but reliably serve good aims. This is because societally beneficial values alignment of AI is not automatic. Crucially, AI systems are designed not just to enact a set of rules, but rather to accomplish a goal in ways that the programmer does not explicitly specify in advance. This leads to an unpredictability that can allow to adverse consequences. As AI pioneer Stuart Russell explains, "No matter how excellently an algorithm maximizes, and no matter how accurate its model of the world, a machine's decisions may be ineffably stupid, in the eyes of an ordinary human, if its utility function is not well aligned with human values." (2015).

Since humans rely heavily on shared tacit knowledge when discussing their values, it seems likely that attempts to represent human values formally will often leave out significant portions of what we think is important. This is addressed by the classic stories of the genie in the lantern, the sorcerer's apprentice, and Midas' touch. Fulfilling the letter of a goal with something far afield from the spirit of the goal is known as "perverse instantiation" (Bostrom 2011). This can occur because the system's programming or training has not

explored relevant dimensions that we really care about (Russell 2014). These are easy to miss because they are typically taken for granted by people, and even with a lot of effort and training data, people cannot reliably think of what they've forgotten to think about.

The complexity of some AI systems in the future (and even now) is likely to exceed human understanding, yet as these systems become more effective we will have efficiency pressures to be increasingly dependent on them, and to cede control to them. It becomes increasingly difficult to specify a set of explicit rules that is robustly in accord with our values, as the domain approaches a complex open world model, operates in the (necessarily complex) real world, and/or as tasks and environments become so complex as to exceed the capacity or scalability of human oversight.. Thus more sophisticated approaches will be necessary to ensure that AI systems accomplish the goals they are given without adverse side effects. See references Russell, Dewey, and Tegmark (2015), Taylor (2016), and Amodei et al. for research threads addressing these issues.

We wish to emphasize that while in the short term these safety issues are important in the same sense that safety is important in any widely-deployed technology, their importance grows in proportion to the power and generality of the AI systems. If future AI systems become "superintelligent" (in the sense of exceeding most or all human mental capabilities), their capability may quickly exceed that of humans, so that safety and value alignment become questions of existential importance.

Parties should recognize that if scientists and technologists are worried about losing what they perceive as a race to reach a threshold power in an AI system, they will have more incentives to cut corners on safety and control, which would obviate the benefits of technical research that enables careful scientists to avoid the very real risks. For the long term, we recommend policies that will encourage the designers of transformative AI systems to work together cooperatively, perhaps through multinational and multicorporate collaborations, in order to discourage race dynamics.

Regarding (4) the social and economic implications of AI:

We are concerned that too little rigorous research has been done on the potential implications of AI for economics and employment. Although there is considerable controversy, we regard as compelling the research by, e.g. Erik Brynjolfsson and Andrew McAfee (http://secondmachineage.com), and by Frey and Osborne (2013) that AI and autonomous systems may replace humans in a large fraction of current jobs, on a timescale that faster than new jobs can be created or workers retrained. Indeed this process may already be underway. In the longer term, it is quite possible (though very contentious) that advanced and ubiquitous AI leads to an economic structure in which full employment is not a sensible expectation because a large fraction of the populace simply does not have (nor can easily be given) skills of significant economic value. Like other economic transitions, AI

has the potential for a dramatic increase in prosperity. However, previous economic transitions may be poor guidance as to how this transition should be managed, and we encourage research into the likely effects of AI on the economy as well as potential policies that can ensure that this impact is an overall good for the vast majority of people.

Regarding (6) the most important research gaps in AI that must be addressed to advance this field and benefit the public:

We would argue that a "virtuous cycle" has now taken hold in AI research, where both public and private R&D leads to systems of significant economic value, which underwrites and incentivizes further research. This cycle can leave insufficiently funded, however, research on the wider implications of, safety of, ethics of, and policy implications of, AI systems that are outside the focus of corporate or even many academic research groups, but have a compelling public interest. FLI helped to develop a set of suggested "Research Priorities for Robust and Beneficial Artificial Intelligence" along these lines (available at http://futureoflife.org/data/documents/research_priorities.pdf); we also support AI safety-relevant research agendas from MIRI (https://intelligence.org/files/TechnicalAgenda.pdf) and as suggested in Amodei et al. (2016). We would advocate for increased funding of research in the areas described by all of these agendas, which address problems in the following research topics: abstract reasoning about superior agents, ambiguity identification, anomaly explanation, computational humility or non-self-centered world models, computational respect or safe exploration, computational sympathy, concept geometry, corrigibility or scalable control, feature identification, formal verification of machine learning models and AI systems, interpretability, logical uncertainty modeling, metareasoning, ontology identification/ refactoring/alignment, robust induction, security in learning source provenance, user modeling, and values modeling.

Regarding (8) the specific steps that could be taken by the federal government, research institutes, universities, and philanthropies to encourage multi-disciplinary AI research:

We believe that research of the type described in answer (6) is crucial in ensuring that advances in AI lead to public good, and that additional federal and philanthropic research funding in these areas can potentially have very high positive impact, and should increase in pace with increased spending on general AI research. Just as safety is a major part of research in biotechnology and nuclear systems, we suggest that funding sources target minimum of 5% of total assessed AI funding be put toward ensuring robustness, interpretability, values alignment, and safety of AI systems.

Regarding (7) the scientific and technical training that will be needed to take advantage of harnessing the potential of AI technology:

In close analogy to the biotech or nuclear power communities, FLI believes that to use AI's

power safely and beneficially will require both those developing AI and those deploying AI to understand or coordinate with relevant fields of ethics and formal risk analysis, to understand how to identify and articulate stakeholder values, and to understand how to think about how their systems might interact with their deployment environments in methodical worst-case analyses.

Regarding (1) the legal and governance implications of AI:

A long-term issue issue that some governments have begun to address is what, if any, legal rights robots (or machine intelligences) should be accorded (see for example Prodham 2016). Our view is that (contrary to the safety considerations discussed above), such discussion is premature and that extreme caution should be taken in setting any legal precedents bestowing such rights.

Regarding (9) any additional information related to AI research or policymaking, not requested above, that [we] believe OSTP should consider:

FLI helped coordinate, and supports, an open letter (http://futureoflife.org/open-letter-autonomous-weapons/) calling for an international agreement precluding the development of offensive fully autonomous weapons. Supported by a very large number of AI researchers and other thinkers, this letter argues that without such an agreement we are likely to see a destabilizing AI weapons arms race that could potentially lead to a new type of WMD accessible to a wide variety of groups and governments. Autonomous weapons could also share some of the characteristics of that make cyberattacks/cyberweapons worrisome: they are potentially inexpensive to develop and deploy, difficult to track, and deployable from a distance at low risk (hence difficult to deter). We therefore believe the U.S. should support multilateral, global, or international agreements to keep humans in the loop. If such agreements are adopted, even if enforcement guarantees are necessarily weaker than with nuclear, biological, and chemical weapons, the spiraling race dynamic could be avoided or slowed.

References

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, Dan Mané. 2016. “Concrete Problems in AI Safety.” arXiv:1606.06565 [cs.AI]. https://arxiv.org/pdf/1606.06565v1.pdf.

Bostrom, Nick. 2014. Superintelligence: Paths, Dangers, Strategies. Oxford: Oxford University Press.

Brogan, Jacob. 2016. "Digital Genies." Slate, April 22. http://www.slate.com/articles/technology/future_tense/2016/04/stuart_russell_interviewed_about_a_i_and_human_values.html.

Frey, Carl, and Michael Osborne. 2013. “The Future of Employment: How Susceptible Are Jobs to Computerisation.” http://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf.

MIRI Blog, May 4. https://intelligence.org/2016/05/04/announcing-a-new-research-program/.
Prodhan, Georgina. 2016. “Europe's robots to become 'electronic persons' under draft plan.” Reuters. http://www.reuters.com/article/us-europe-robotics-lawmaking-idUSKCN0Z72AY.

Russell, Stuart. 2015. "2015: What Do You Think About Machines That Think?" Edge. https://www.edge.org/response-detail/26157.

Russell, Stuart, Daniel Dewey, and Max Tegmark. 2015. “Research Priorities for Robust and Beneficial Artificial Intelligence”. AI Magazine 36:4.

Soares, Nate and Benja Fallenstein. 2014. “Agent Foundations for Aligning Machine Intelligence with Human Interests: A Technical Research Agenda.” MIRI. https://intelligence.org/files/TechnicalAgenda.pdf.

Taylor, Jessica. 2016. “A New MIRI Research Program with a Machine Learning Focus.”

Yudkowsky, Eliezer. 2008. “Artificial Intelligence as a Positive and Negative Factor in Global Risk.” In Global Catastrophic Risks, edited by Nick Bostrom and Milan M. Cirkovic, 308–345. New York: Oxford University Press.

Respondent 150

Amy Magnus, Air Force Institute of Technology

As a research in strong artificial intelligence, I distinguish training from learning. Training that cumulates in compliance does not contribute significantly to intelligence. Intelligence involves from learning and learning from inquisitiveness. My work centers on the Modified Turing Test, a standard test for artificial intelligence. The Modified Turing Test places the computer in the role of interrogator. An intelligent computer must be able to distinguish pretenders from experts.

Consider The Imitation Game, the 2014 movie dramatizing Alan Turing’s role in the cracking the Enigma code during World War II. The film itself is based on the Turing Test. Police arrest Turing and suspect that he is a Soviet spy. In the dramatization, a police officer

conducts an interrogation that is a Turing Test. He must decide whether or not Turing is a spy, one form of pretender. Across the table, Turing is conducting a Modified Turing Test. He participates in the police officer's interrogation to test whether or not the officer can help him escape trouble with the law. At the conclusion of the interrogation—SPOILER ALERT—the policeman is unable to make the choice between patriot and spy. The officer fails to complete his administration of the Turing Test and thus fails the Modified Turing Test. However, Turing passes a Modified Turing Test of his own. The Alan Turing character picks a reasonable time to ask the police officer "Am I a spy or a patriot?". Turing's intelligence allows him to conclude with confidence that the policeman is a fool and of no use to him. [The movie script is brilliant writing, though I regret that the movie portrays Turing as autistic. Not every misunderstood genius is on the spectrum.]

The Modified Turing Test (ModTT) is a strong AI test as it gives the computer the responsibility to select what is being tested. As interrogator, the computer has control over the test questions.

Deep Learning achieves compliance-based training which is a start. We have the mathematics to crack open Deep Learning. Domain Theory, combinatorial geometry, and constraint program provide sufficient insight into the mathematical forms of deep learning to make reasoned assessments about what solutions represent appropriate expertise and which do not. A scientific rule that has informed pattern recognition and constraint programming, Ockham's razor states simple explanations should be preferred to more complicated ones and that the explanation of a new phenomenon should be based on what is already known. Constraint programming addresses the first and most familiar part of Ockham's razor: Simple Explanations are preferred to complicated ones. Regulation addresses the second part: Explanations should be based on what is already known. Let me recommend a third condition: Among explanations of similar complexity, the one that answers the most questions and the most relevant questions should be preferred to the others.

Violate the first part of Ockham's razor and we risk memorizing the solution. Violate the second part and we risk arrogance. If we adhere to the third condition to Ockham's razor, we move away from compliance based training into semantic symbol generation and toward artificial languages.

Artificial Intelligence should give us insight into our own intelligence. It should be able to deep the conversation not supplant it.

Respondent 151

Amy Magnus, Air Force Institute of Technology

As a researcher in strong artificial intelligence, I distinguish training from learning. Training that cumulates in compliance does not contribute significantly to intelligence. Intelligence evolves from learning and learning from inquisitiveness.

My work centers on the Modified Turing Test, a standard test for artificial intelligence. The Modified Turing Test places the computer in the role of interrogator. An intelligent computer must be able to distinguish pretenders from experts.

Consider The Imitation Game, the 2014 movie dramatizing Alan Turing's role in the cracking the Enigma code during World War II. The film itself is based on the Turing Test. Police arrest Turing and suspect that he is a Soviet spy. In the dramatization, a police officer conducts an interrogation that is a Turing Test. He must decide whether or not Turing is a spy, one form of pretender. Across the table, Turing is conducting a Modified Turing Test. He participates in the police officer's interrogation to test whether or not the officer can help him escape trouble with the law. At the conclusion of the interrogation—SPOILER ALERT—the policeman is unable to make the choice between patriot and spy. The officer fails to complete his administration of the Turing Test and thus fails the Modified Turing Test. However, Turing passes a Modified Turing Test of his own. The Alan Turing character picks a reasonable time to ask the police officer "Am I a spy or a patriot?". Turing's intelligence allows him to conclude with confidence that the policeman is a fool and of no use to him. [The movie script is brilliant writing, though I regret that the movie portrays Turing as autistic. Not every misunderstood genius is on the spectrum.]

The Modified Turing Test (ModTT) is a strong AI test as the computer in the role of interrogator has control over the test questions. To realize the Modified Turing Test, I need information science fitted to the role of interrogation. I concentrate on a symbolic discipline and a sub-symbolic discipline—respectively, information access and inquisitive pattern recognition.

First let's discuss information access: Information access is a sister discipline to information retrieval, the mathematics behind search engines. Information retrieval extracts the expert view of a document; that is, it parses the specifics of a document. Information Access parses the general information of a document and thus extracts the interrogator's view of the document. The distinction between expert and interrogator views is subtle but important. When we understand the difference between information access and information retrieval, we understand why taking an 8th grade science test is a hard artificial intelligence problem. Even though humans are raised in an environment steeped in call and response, query and assertion, most of our documents represent the expert view, not the interrogator's view. Documents mostly assert; less often, they question. Is that a problem? For computers, yes. The issue that artificial intelligence most overcome is that assertions are disproportionately represented compared to queries. For AI to answer questions, it has to recognize the semantics of the under-represented question, it must not get confused by the semantics of the over-represented assertion, and in fact needs to learn the rules for translating between

query and assertion. Attempting to use Information Retrieval in support of transfer learning is like using high frequency data to describe low frequency physics. You'll be looking in the wrong place and see only noise.

Next, consider inquisitive pattern recognition where machine learning culminates in the cultivation of artificial symbols and the population of artificial languages. The objective of artificial languages is to generate relevant queries for the interrogator. The query is the most common mechanism for initiating interaction between self and other. A query is a compact trove of information characterizing the work of its author, the perceived expertise of its receiver, and the value of its subject. Well posed, a question informs and structures work; ill-timed, it can upset the apple cart. Deep Learning could potentially contribute to pattern recognition as a discipline but first it must be more ambitious. Deep Learning achieves compliance-based training but training is only a start. We have the mathematics to crack open Deep Learning. Domain Theory, combinatorial geometry, and constraint program provide sufficient insight into the mathematical forms of deep learning to make reasoned assessments about what solutions represent appropriate expertise and which do not. A scientific rule that has informed pattern recognition and constraint programming, Ockham's razor states simple explanations should be preferred to more complicated ones and that the explanation of a new phenomenon should be based on what is already known. Constraint programming addresses the first and most familiar part of Ockham's razor: Simple Explanations are preferred to complicated ones. Regulation addresses the second part: Explanations should be based on what is already known. Let me recommend a third condition: Among explanations of similar complexity, the one that answers the most questions and the most relevant questions should be preferred to the others.

Violate the first part of Ockham's razor and we risk memorizing the solution. Violate the second part and we risk arrogance. If we adhere to the third condition to Ockham's razor, we move away from compliance based training into semantic symbol generation and toward artificial languages.

Artificial Intelligence should give us insight into our own intelligence. It should be able to deep the conversation not supplant it. Developing intelligence is not about discovering what is right and what is wrong. It's about mastery...discovering what is hard and what is easy and how to become proficient in both.

Respondent 152

Stephen Williams, OptimaLogic

AI of various forms is the single most powerful collection of technology: It can transform manufacturing, farming, transportation, safety and disaster recovery, personal care, training & education, sports, entertainment, space travel, etc. We have already solved an amazing amount, but we've only scratched the surface. We should focus on creating the next big

leaps, sharing methods widely academically and as open source when possible. The application of AI has just begun to play out: there are still many opportunities unexplored.

Responses: 1) The legal issues will take some attention, but don't seem overly difficult to resolve. The governance implications are huge: Not because control is needed, but because of the implications of a surplus economy, rapid change, and the resulting shift in jobs. This is good overall, but will need management, creatively inspiring people to bridge difficulties.

2) AI can increase the public good in many ways: Providing optimization of resources and funds, providing automated labor for civil construction, safety, coordination, awareness of trends, and predicting the effect of changes.

3) It is possible to create dangerous systems driven by AI. The main solution to that is better testing, theory, safety overwatch systems, themselves driven by AI, and more innovation around checks and balances in systems and environments. A key part of this will be driven by financial incentives and similar feedback.

4) See 1).

Respondent 153

John Watson, Columbus Research, Inc.

At Columbus Research, Inc. (San Diego, CA), we develop and use AI for two seemingly separate industries: Analysis of P-20 student data and robotics control systems. The connection between these is that much of our robotics control AI is used in custom curriculum for classroom robotics lessons. In order to efficiently respond to several of the OSTP topics, we provide the topic numbers and summary responses below.

Response to OSTP topic 2, the use of AI for public good:

Education has been a slow adopter of technology, including AI. Where general manufacturing and sales businesses have been using computer technology and databases for over 40 years to manage supply chains, inventory control and sales; the education industry only adopted widespread use of student information systems in the past two decades. Regarding AI specifically, transactional business data has been mined using AI algorithms since the early 1990's to detect patterns in data, from credit card fraud to purchasing behavior.

In education there are two applications of AI that have yet to emerge. The first is an optimization problem; the second, analytical analysis towards identifying patterns and anomalies. First, the U.S. has had a long-stated goal of increasing STEM (science, technology, engineering and math) graduates from high school and community college

student success. These actions could help respond to the need for a skilled workforce, and the health of the economy. Statistics show that there has been a loss of manufacturing jobs and that products made in America have "become more complex, requiring .. smarter workers. Automation is part of that trend...Another part of the problem is that the people who need work aren't properly trained for some of the available jobs. There is a basic mismatch between the skills the labor market has and the skills the labor market needs...California is a home to a swath of 'economic castaways' who didn't go to college and are now struggling to attract the attention of high-tech employers," (Christopher Thornberg, Beacon Economics; as found in Los Angeles Times, July 21, 2016). The current method for optimizing a student path planning begins in high school and is largely a counselor-base, manual process. Yet, there are significant data resources available to researchers and AI developers to develop optimization plans, even on an individual student level. These systems could provide additional guidance as students make key decisions.

Second, on a macro level, while the Department of Education has worked to foster strong SLDS's (state longitudinal data systems) holding educational data, and much progress has been made in the past decade, is has been a long-standing challenge for states to develop wide educational research agendas. Instead, states face fragmented data siloes and too numerous stakeholder groups, and after prioritization, end up only analyzing a small portion of their vast data stores to answer few research questions. The Department of Education has attempted to nudge states to produce actionable metrics output, partly through its CEDS (common educational data standards) definitions. In these SLDS data are patterns, outliers, early-warning signals, but the data is hardly touched. AI can serve states' goals by providing wide analysis of whole large data sets using similar techniques and tools use in other industries.

Response to OSTP topics 1) the legal and governance implications of AI, and 3) the safety and control issues for AI:

Since the advent of the Internet and World Wide Web, the software industry has changed dramatically. Popular iPhone apps appearing on Apple Computer's App Store can be developed and distributed quickly with a relatively minor investment and as few as a single programmer. Using conventional development tools and techniques prior to 2000 would require a team of developers, slow BBS-based connections to transfer code and data files between team members, and physical software delivery via CD or DVD. Software, and AI, development processes and tools have been turned upside-down.

As small boutique AI group, we have internally discussed the risks of AI development due to the frictionless software, data and Internet environments we find today. AI daemons can be written to run in the cloud. Even in a well-meaning situation, say, a researcher experimenting with AI on the cloud using Google Tensorflow, could unleash rogue systems that, unlike current malware being intentionally managed to maximize benefit to the

hacker, could run creating difficult to control denial of service attacks, resource hogs (taking down networks), or systems designed to locate certain patterns in data, traffic, behavior or recognize some state, and take action against the host or user. What happens if there is a mistake made? Who would know? Methods for managing this possibility include 1) adding an AI sensor layer to monitor Internet traffic for defined patterns of behavior, 2) registration of certain classes of AI development organizations to enable government oversight, or 3) voluntary programs for cloud hosts (Microsoft, Amazon and Google cloud services, along with regional research data centers -- supercomputer centers) to self-monitor for these classes of activities.

Thanks for the opportunity to submit this summary response.

John B. Watson, Ph.D.

Respondent 154

John Watson, Columbus Research, Inc.

At Columbus Research, Inc. (San Diego, CA), we develop and use AI for two seemingly separate industries: Analysis of P-20 student data and robotics control systems. The connection between these is that much of our robotics control AI is used in custom curriculum for classroom robotics lessons. In order to efficiently respond to several of the OSTP topics, we provide the topic numbers and summary responses below.

Response to OSTP topic 2, the use of AI for public good:

Education has been a slow adopter of technology, including AI. Where general manufacturing and sales businesses have been using computer technology and databases for over 40 years to manage supply chains, inventory control and sales; the education industry only adopted widespread use of student information systems in the past two decades. Regarding AI specifically, transactional business data has been mined using AI algorithms since the early 1990's to detect patterns in data, from credit card fraud to purchasing behavior.

In education there are two applications of AI that have yet to emerge. The first is an optimization problem; the second, analytical analysis towards identifying patterns and anomalies. First, the U.S. has had a long-stated goal of increasing STEM (science, technology, engineering and math) graduates from high school and community college student success. These actions could help respond to the need for a skilled workforce, and the health of the economy. Statistics show that there has been a loss of manufacturing jobs and that products made in America have "become more complex, requiring .. smarter workers. Automation is part of that trend...Another part of the problem is that the people who need work aren't properly trained for some of the available jobs. There is a basic

mismatch between the skills the labor market has and the skills the labor market needs...California is a home to a swath of 'economic castaways' who didn't go to college and are now struggling to attract the attention of high-tech employers," (Christopher Thornberg, Beacon Economics; as found in Los Angeles Times, July 21, 2016). The current method for optimizing a student path planning begins in high school and is largely a counselor-base, manual process. Yet, there are significant data resources available to researchers and AI developers to develop optimization plans, even on an individual student level. These systems could provide additional guidance as students make key decisions.

Second, on a macro level, while the Department of Education has worked to foster strong SLDS's (state longitudinal data systems) holding educational data, and much progress has been made in the past decade, is has been a long-standing challenge for states to develop wide educational research agendas. Instead, states face fragmented data siloes and too numerous stakeholder groups, and after prioritization, end up only analyzing a small portion of their vast data stores to answer few research questions. The Department of Education has attempted to nudge states to produce actionable metrics output, partly through its CEDS (common educational data standards) definitions. In these SLDS data are patterns, outliers, early-warning signals, but the data is hardly touched. AI can serve states' goals by providing wide analysis of whole large data sets using similar techniques and tools use in other industries.

Response to OSTP topics 1) the legal and governance implications of AI, and 3) the safety and control issues for AI:

Since the advent of the Internet and World Wide Web, the software industry has changed dramatically. Popular iPhone apps appearing on Apple Computer's App Store can be developed and distributed quickly with a relatively minor investment and as few as a single programmer. Using conventional development tools and techniques prior to 2000 would require a team of developers, slow BBS-based connections to transfer code and data files between team members, and physical software delivery via CD or DVD. Software, and AI, development processes and tools have been turned upside-down.

As small boutique AI group, we have internally discussed the risks of AI development due to the frictionless software, data and Internet environments we find today. AI daemons can be written to run in the cloud. Even in a well-meaning situation, say, a researcher experimenting with AI on the cloud using Google Tensorflow, could unleash rogue systems that, unlike current malware being intentionally managed to maximize benefit to the hacker, could run creating difficult to control denial of service attacks, resource hogs (taking down networks), or systems designed to locate certain patterns in data, traffic, behavior or recognize some state, and take action against the host or user. What happens if there is a mistake made? Who would know? Methods for managing this possibility include 1) adding an AI sensor layer to monitor Internet traffic for defined patterns of behavior, 2)

registration of certain classes of AI development organizations to enable government oversight, or 3) voluntary programs for cloud hosts (Microsoft, Amazon and Google cloud services, along with regional research data centers -- supercomputer centers) to self-monitor for these classes of activities.

Thanks for the opportunity to submit this summary response.

John B. Watson, Ph.D.

Respondent 155

Daniel Olsher, Integral Mind

Abstract from our paper at:
http://intmind.com/pubs/cogSolv.pdf

Truly understanding what others need and want, how they see the world, and how they feel are core prerequisites for
successful conflict resolution and humanitarian response. Today, however, human cognitive limitations, insufficient
expertise in the right hands, and difficulty in managing complex social, conflict, and real-world knowledge conspire to
prevent us from reaching our ultimate potential.

This paper introduces cogSolv, a highly novel Artificial Intelligence
system capable of understanding how people from other groups view the world, simulating their reactions, and
combining this with knowledge of the real world in order to persuade, find negotiation win-wins and enhance
outcomes, avoid offense, provide peacekeeping decision tools, and protect emergency responders' health.

Ready to go
today, portable, and requiring virtually no specialist expertise, cogSolv allows governments and local NGOs to use
expert culture and conflict resolution knowledge to accurately perform a wide range of humanitarian simulations.

cogSolv assists responders with training, managing complexity, centralizing and sharing knowledge, and, ultimately,
maximizing the potential for equitable conflict resolution and maximally effective humanitarian response.

Respondent 156

Andrew Porter, Center for Theology and the Natural Sciences

Have you talked to Hubert L. Dreyfus at UC Berkeley? I think he could put AI in a different perspective for you.

Respondent 157

Erin Hahn, Johns Hopkins University Applied Physics Laboratory

Response to OSTP Request for Information on Artificial Intelligence
Topic (1): The legal and governance implications of artificial intelligence
Ms. Erin Hahn, Ms. Emmy Probasco, Mr. Tan MacLeod
Johns Hopkins University Applied Physics Laboratory

The Johns Hopkins University Applied Physics Laboratory (JHU/APL) submits this response to topic (1): the legal and governance implications of artificial intelligence (Al). On this topic, we note two important issues: (1) in considering legal and governance issues, we must understand that law and policy reflect public attitudes and changes to either require an appreciation of public perceptions about the capabilities and limitation of the technologies, and (2) discussion of the legal and governance implications of AI must be informed by the technical community, particularly in recognition that Al has the potential to impact all facets of life.

Al will enable delegation of complex and nuanced decisions and actions from humans to systems at a level not previously experienced and difficult to understand. The sophistication enables such a degree of advanced autonomous operation by these systems that existing legal and govemru1ce concept about accountability and liability may be challenged. Moreover, autonomous systems will continue to challenge social norms in ways that raise novel issues that might require new policies.

The nexus of law, policy, and public sentiment towards technologies presents opportunities and challenges to the development and use of autonomous systems both commercially and by the government. Public sentiment toward Al --- currently an ambiguous mixture of optimism and fear --- shapes law and policy. The perceptions that influence public sentiments are informed by exposure and personal experiences. Developers have an opportunity to shape public perceptions (and law and policy) by thoughtfully developing systems that, where sensible, comport with established norms and gradually evolve the public's perception of these systems. The mainstream automotive industry's gradual evolution from automatic anti-lock brakes to cruise control to self-parking features may be a useful model. The primary public concern with autonomous systems is best summed up as an issue of trust. Initial research into the use of autonomous modes of operation in space

vehicles, as one example, indicates that trust is gained by experience, understanding, and often in situations of necessity.

It is common to apply existing legal or governance frameworks to emerging technology, and in many cases, it works; with any gaps addressed by minor adaptations to existing law. Indeed, this is currently the approach with systems utilizing aspects of AI---they are analyzed against current frameworks for safety and liability and, in the military context, under existing principles of international humanitarian law. Whether Al can meet our current legal and moral standards can be seen as a minimum baseline for acceptance of the technology, but advanced AI may require us to think quite differently about our current governance frameworks. For example, manufacturers must warn consumers of foreseeable risks a product may pose to avoid liability for negligence. However, development of AI is moving away from traditional algorithms, and systems may engage in behaviors not foreseen by their creators. If we cannot rely on foreseeability (and many developers warn that we cannot), then we cannot use traditional frameworks. We can perhaps use a strict liability standard, which requires no finding of fault or negligence, but still leaves the issue of compensation to the aggrieved unresolved. This issue is even more complicated in circumstances where accountability is not simply an issue of legal liability and recompense but one of normative accountability, as in the case of military leaders considering the use of autonomous military systems. Who is accountable if an autonomous system uses inadvertent damage or death in war?

This question underscores the importance that the discussion of the legal and governance implications or Al must be informed by the technical community to avoid hyperbole and focus on actual risks, constraints, and hypotheticals. We cannot adequately address whether, as the example highlighted, we have to shift toward a completely new legal paradigm or manner of overseeing Al without understanding what we can reasonably expect (or not) from the technology. While it may be difficult even for developers to know, we need to understand areas that, if thoughtfully addressed now, can reduce risks posed by AI over time. Rather than worry about the inevitability of Skynet (in which case, concern over legal and governance issues is meaningless), we need a more nuanced legal discussion about Al that those developing the technology can best inform. We have seen this discussion develop with technology such as driverless cars.

With so many potential instantiations of Al, the technical and policy communities should consider now the range of impacts to society. With this consideration and the benefit of technical analysis, we should begin the national conversation on emerging autonomous technologies early to help shape public perceptions and research and development choices---before laws and policies are made that counteract the potential benefits of the technology.

Ms. Erin Hahn (XXXXXXXXX)
Ms. Emmy Probasco (XXXXXXXXX)
Mr. Ian Macleod (XXXXXXXXX)

Response to OSTP Request for Information on Artificial Intelligence
Topic (2): On the public use of Al
Dave Scheidt
Johns Hopkins University Applied Physics Laboratory

What is artificial intelligence (AI)?

The romantic view of AI is the study of thinking machines, specifically machines that are equivalent to, and think like, humans. While it is true that many of the reasoning methods being pursued by the Al community are anthropomorphic, AI has a more pragmatic side, which is the study of algorithms that allow machines to produce answers to complex, unsolvable problems. To understand the difference between an AI system and "unintelligent" computer systems, we should first consider how software is normally developed. The software development process begins with the specification of requirements which state system goals and objectives. After requirements specification, software engineers painstakingly determine the appropriate machine response to each and every combination of inputs that might be encountered by the system's software during software design. Sometimes the appropriate response to system stimuli are defined mathematically by using control theory and sometimes software engineers use brute force by exhaustively enumerating all possible stimulus-response pairings. Traditional software engineering works well for the vast majority of software systems employed for man; however, when system requirements demand that machines make decisions in a world that is too complex or too uncertain for engineers to solve during software design, AI is required. The design process for AI, necessitated by the need to have machines address unsolvable problems during execution, is a radically different approach to developing control software. When developing AI software engineers do not program explicit responses to situations encountered by the machine; rather, software engineers write software that provides machine with an ability to solve problems during program execution allowing the Al software to produce a decision that was not explicitly encoded in the software. All major forms of AI research, which includes deductive, inductive, and abductive reasoning, search-based algorithms, and neural networks, exhibit the property that the machine's answers to specific problems are not explicitly encoded with in the AI software; rather, methods for devising answers to the problem are encoded. This subtle distinction between Al and unintelligent controller provides the power and promise of Al and AI's greatest risk. The promise of Al is an ability to solve important problems that cannot be solved through other means. The risk of Al is the potential to produce unvetted responses to situations that run counter to the designer's wishes.

On understanding the utility of embedding Al into machines:

For the last decade, the Johns Hopkins University Applied Physics Laboratory (JHU/APL)

has been studying the underlying principles on whether machines should rely upon a human operator or AI to make a decision. Our research frames the machine versus human intelligence question as a command and control (C2) problem. C2 research studies the management of information distribution and decision authority over a distributed set of actors, which, in the case of AI-enabled machines, includes actors that are humans and actors that are Al-enabled machines. Our C2 analysis has identified three key environmental characteristics that are key to defining which node in the system should be empowered to respond to changes in the environment. The three defining characteristics are: (l) the complexity (C) of the problem being addressed, (2) the rate at which information (T) can be exchanged and processed by remote controller, and (3) the rate at which the unexpected (U) occurs within the environment. A general formula that defines the difficulty (D) of a C2 scenario can be described as: D - C*U/I. The C2 difficulty equation is key to understanding the potential of AI-enabled systems, when considering whether or not an Al-enabled machine will be more effective than a human operator; the higher the value of D, the more likely the Al-enabled machine will outperform a human controlled machine.

General Use Cases of AI:

Now that we understand the practical definition of AI, we ask the question: when is it useful to have a machine use AI to make a decision? After all, after millions of years of evolution and roughly 10,000 years of civilization, humans are quite good at making decisions in complex, uncertain environments. Through our research in AI-enabled systems, JHU/APL identified three general use cases for AI: first, for some tasks, Al is more cost effective that humans; second, Al is better suited than humans al solving some, but not all, problems; third, Al allows us to develop machines that are capable of responding faster than humans to unexpected circumstances.

Humans possess diverse intelligence. A single human brain is capable of producing effective decisions for an astoundingly diverse set of problems. For example, that a single human would have the intelligence to perform CPR, cook Jambalaya, walk down a crowded street without colliding with other people, and change the oil in a Honda Civic is remarkable. Every person on the planet is capable of competently performing thousands of complicated tasks. By comparison, machines exhibiting practical AI are currently, and likely to be for the foreseeable future, dedicated specialists. For example, an autonomous car may, in the very near future, exhibit driving skills that are safer and more efficient than the driving skills of the average human driven driver; however, it is both unnecessary and unlikely that an autonomous car's Al will know how to groom a horse, make soup, or dance the jitterbug. The narrow scoping of AI within intelligent machines limits the risk associated with intelligent machines to the tasks the machines were designed to accomplish.

The greatest risk associated with AI is the risk of undesirable detrimental, consequences from decisions emerging from unintended combinations of legitimate rules and/or patterns. The potential for emergent consequences was central to lsaac Asimov's influential book, l

Robot, as well as Arthur C. Clark 's novel 2001: A Space Odyssey and Dennis Felltharn Jones' Colossus. This theme of powerful, runaway Al is also found in popular movies such as "The Terminator" and the "The Matrix." The narrow focus of practical Al cannot produce the catastrophic, existential threats from AI that are popularized in popular fiction. Quite simply, the scope of the reasoning embedded within a practical AI system is, and is likely to remain, both limited and entirely under the control of the human engineers that develop the system. For example, autonomous cars that use practical AI to safely drive from 4th and Main in Dover, Delaware to the 2000 block of Euclid Avenue in Cleveland will not be programmed to reason on the failings of mankind, consider humans impact of the environment, or reason about the global political order because the human designers will not be incentivized to spend the time and energy to expand the scope of Al in the car to include algorithms that consider those issues. This does not mean that failures from emergent unintended consequences within practical AI won't occur; they will. It means that failures within practical Al are constrained to the scope of the AI that human engineers place within the system and that these failures will be limited, controllable, and repairable.

Recognizing that practical AI is limited in scope and dedicated to addressing complex, uncertain problems that can't be solved by normal software engineering methods let us examine some examples of the three general use cases for AI.

Cost effective AI -- Al-enabled machines are able to perform jobs that are traditionally performed by humans at lower cost. While the loss of a job is individually traumatic, the net benefit to society when humans are replaced by machines has always been positive in the past and will produce a productivity gain as Al becomes mainstream as well. As mankind saw with the advent of the plow, the harness and domestication of animals, and the Jaquard loom, advancements that allow machines to take over human jobs inevitably introduce additional human capital into the market, freeing human intellectual capital to pursue endeavors that are still out of reach from machine. Some of the tasks that are expected to be performed by less expensive, practical AI-enabled machines include:

a. Transportation services, to include truck, bus, and automobile drivers, railroad engineers, and pilots;
b. Routine maintenance and cleaning tasks, to include trash pickup, cleaning, decontamination tasks; and
c. Logistics and fulfillment tasks.

Al that makes better decisions than humans - There exists a set of narrowly focused problem for which practical AI is already more effective than human decision-making. One example is route planning over public roads provided by the Google/Android and Apple Map utilities. While humans are capable of analyzing traffic data and roadmaps to produce a path between two locations, search-based AI algorithms provided on Apple and Android phones reliably produce equal or better quality substantially faster than their human counterparts. To date, tasks in which AI outperforms humans are limited to the solving of

problems that, while complex, are well understood and easily modeled within a computer. It is possible that this current limitation on practical Al may be overturned as recent advancements in structured neural network techniques such as Deep Learning are on the cusp of producing Al that enables machines to reason over unstructured, difficult-lo-model problems. Regardless of Al method, AI intellectual dominance over human is restricted to narrowly scoped problem sets and is likely to remain so. Some of the tasks that AI-enabled machines will perform better than humans include:

d. Routing tasks, to include traffic management and driving or all forms of vehicles as well as information routing through computer networks;
e. Logistics, to include fulfillment and warehouse management; and
f. Exploration of large, structured data sets, to include natural language and computer language data

Al is foster than human intelligence -- Thus far, we have discussed the benefits of using practical AI to improve the human condition by using AI to do a better job of tasks that are currently being performed by human. Our last general use case is the most compelling because it allows us to build machines that perform tasks that cannot be performed by humans at all. These AI-enabled machines use AI to solve complex problems faster than humans. The speed at which AI-enabled machines can react allows us to build thinking machines that are capable of diagnosing and repairing faults within complex high-speed systems. One such system is the national power grid, considered by many to be the most complex machine ever devised by man. Diagnosing and reconfiguring faults within this system faster than these faults can propagate throughout the system is beyond the capacity of man [see Northeast blackout of 2003 for an example]. Likewise, diagnosing and responding to damage or attacks in cyberspace is beyond the capacity of man, but not of AI. A particularly compel ling use case for AI is when a machine that cannot communicate with a human, or can only communicate slowly, must address an unanticipated problem. An excellent example is NASA's New Horizon's probe that visited Pluto in the summer of 2015. During the rendezvous with Pluto New Horizons was 4.5 light-hours from Earth; the 9-hour round trip to communicate with the probe prevented human operators from handling any unanticipated faults during the rendezvous. New Horizons was designed with limited, practical AI to diagnose and manage faults during the Pluto rendezvous. Fortunately for NA SA, no faults occurred; however, had an unanticipated fault occurred it would have been impossible for humans to diagnose and manage the fault within the rendezvous window. The only way to manage New Horizons would have been through AI. Examples of tasks that are can only be managed through AI due to the time required for humans to make decisions include:

g. Exploration of deep space by unmanned spacecraft;
h. Exploration of the deep ocean and underground by unmanned craft;
i. Timely fault management of the national power grid and the Internet;
j. Exploration and damage control of toxic sites with communications difficulty, to include

the Fukishima nuclear power plant; and
k. Unmanned military vehicles operating in denied (jammed) environments.

Mr. David Scheidt (XXXXXXXXX)

Response to OSTP Request for Information on Artificial Intelligence
Topic (3): Safety and control issues for Al
Dr. Christopher Rouff Dr. Aurora Schmidt, Dr. Christine Piatko and Mr. David Scheidt
Johns Hopkins University Applied Physics Laboratory

The great potential of artificial intelligence (AI) and autonomy is that they will enable the execution of complex and nuanced responses and adaptations to different environments without the need for human intervention. However, this advanced ability to operate autonomously necessitates an assurance of safety to the public during development and operation, which is challenging to provide due to the complexity and sophistication of AI.

The recent report of the Department of Defense Research and Engineering Autonomy Community of Interest Test and Evaluation, Verification, and Validation Working Group identified the need for formal assurance arguments for autonomous systems as one of the key research challenges to being able to justify that a system is acceptably safe and secure [TEVV 2015]. Formalizing assurance arguments will require new methods of incorporating safety parameters in requirements specifications and increasing use of formal methods for testing, verifying, and certifying the safety of large autonomous systems. Building systems that are directly constrained, during design, to satisfy critical requirements can revolutionize the cost and speed tradeoffs in building autonomous systems where humans depend on robust operations.

The first main challenge to developing assurance arguments to ensure safe and trustworthy operation is creating clear and precise requirements, agreed upon prior to the design of autonomous controls systems. The requirements for system behavior may be encoded as a contract between humans and the system. These requirements can then be used during testing, verification and certification of autonomous system components. Composable assurance arguments can then he built up for how an overall system is expected to perform, based on the assurance arguments for its component systems. Articulating such requirements for composable assurance arguments will require research in new methods for eliciting interdependencies of safety and performance thresholds between the individual system components and the overall system, as well as describing how each relates when performing under different conditions [Scheidt and Piatko 2016].

The second main challenge is then testing, validating, and certifying the intelligent system and its components given these operating requirements. This problem is much more difficult than traditional verification of software. Intelligent systems learn and adapt to

their environment and are constantly changing as they learn. Anticipating how intelligent systems will change is difficult because it is based on the environment the system is exposed to and the learning algorithms the system is employing. Because of this variation, traditional simulation testing and field testing cannot encompass all possible scenarios that would be needed to ensure the system will be safe to operate no matter which environment or stimulus it may encounter. Formal guarantees of safe and correct behavior are needed to provide a high level of assurance that all autonomous functions will continue to respect the safety and behavioral requirements.

A process called formal methods have been proven on a wide range of safety critical systems. Formal methods are mathematical-based methods and tools used to specify system behaviors and provide mathematical guarantees that a system's properties are correct. Though proven successful on a number of safety critical systems, formal methods have been viewed as limited in their ability to verify highly complex systems. However, through JHU-APL's approach of generating safety theorems independent from the particular control algorithms of the autonomous system being tested, we are making strides in providing a high level of assurance in increasing complex types of systems [Schmidt, et al. 2016; Garder, et al. 2016], ln addition, these methods allow for realistic dynamics models, nondeterministic ranges of uncertain environmental parameters, and worst-case or adversarial interactions with other systems and humans, allowing us to test for robustness to wide ranges of circumstances.

Rather than model a large system to be verified, we generate formally verified safety theorems on the range of system parameters and control decisions that ensure acceptable behavior. We use these theorems to search for cases in the full system where the autonomous decisions violate these safety theorems. This approach has been very effective means of verifying unprecedentedly complex and sophisticated intelligent systems, such as the da Vinci robotic surgery system [Kouskoulas, et al. 2013].

The Johns Hopkins University Applied Physics Laboratory (JHU/APL) has had success using this approach to verify the Federal Aviation Administration's next-generation advanced collision avoidance system (ACAS X) [Jeannin, et al. 2015]. The ACAS X system, while not engaging in online learning, employs tables for system behavior that exceed 1 million states. Traditional approaches to formal verification would not be able to test and provide strong guarantees for such a large system. Our approach of generating theorems of safe operation and comparing system behavior to these theorems uncovered stressing operating scenarios that would not be discoverable with traditional simulation testing approaches. In addition, we demonstrated that the safety theorems themselves can be operationalized as online system monitoring tools that can automatically constrain system behavior to safe operation in a way that only intervenes when absolutely necessary. This approach to guaranteeing safety can help to separate the task of learning and adaptation that intelligent systems should be doing from the continuing need to satisfy requirements for safe and reliable operation. Additionally, we can also use proven safety theorems to directly constrain

autonomous controllers during the design and optimization of their intelligent planning strategies, thereby ensuring strong safety guarantees during the design of the system and simplifying the testing and evaluation of the system in later phases of deployment. This new method for ensuring that an intelligent adaptive system under development automatically satisfies the critical requirements for safe and operationally reliable behavior would rapidly accelerate the availability of trustworthy intelligent systems that safely interact with each other and human teammates.

Additional research and expansion of these techniques for safe development, testing, and operation should be conducted as well as linking these methods to the emerging legal dialogue on accountability, public safety, and acceptable use of systems enabling or employing AI

References

TEVV 2015] Department of Defense Research & Engineering Autonomy Community of Interest (COI) Test and Evaluation, Verification and Validation (TEVV) Working Group Technology Investment Strategy www.defenseinnovationmarketplace.mil/resources/OSD_ATEVV_STRAT_DIST_A_SIGNED.PDF (2015).

Scheidt and Piatko 2016] "A Method for Specifying Autonomous System Requirements." AFRL Safe and Secure Systems and Software Symposium (S5) (2016).

Y. Kouskoulas, D. Renshaw, A. Platzer, P. Kazanzides. "Certifying the safe design of a virtual fixture control algorithms for a surgical robot". Hybrid Systems: Computation and Control, HSCC'.13, Philadelphia, PA, USA. ACM, Apr. 8-13, 2013.

A. Schmidt, C. Rouff, R . Gardner, D Genin, Y. Kouskoulas, G.Mullins. "Complementary Formal Techniques for Verification and Validation

Dr. Christopher Rouff (XXXXXXXXX)
Dr. Aurora Schmidt (XXXXXXXXX)
Dr. Christine Piatko (XXXXXXXXX)
Mr. David Scheidt (XXXXXXXXX)

Respondent 158

Duane Blackburn, MITRE Corporation

Context

The MITRE Corporation (www.mitre.org) is a not-for-profit company that runs seven Federally Funded Research and Development Centers (FFRDCs) for the U.S. government. MITRE's seven FFRDCs are:

National Security Engineering Center: sponsored by the Department of Defense, NSEC supports a broad and diverse set of sponsors within the Department of Defense and the Intelligence Community.

Center for Advanced Aviation System Development: Sponsored by the Federal Aviation Administration, CAASD works to advance the safety, effectiveness, and efficiency of global aviation.

Center for Enterprise Modernization: Sponsored by the Internal Revenue Service and co¬ sponsored by the Department of Veterans Affairs, CEM aims to support systems integration, engineer better technical solutions, deliver more efficient business processes, and implement new legislative requirements.

CMS Alliance to Modernize Healthcare: The Centers for Medicare & Medicaid Services works with CAMH toward an integrated health system with improved access and quality at sustainable cost.

Homeland Security Systems Engineering and Development Institute: Operated on behalf of the Department of Homeland Security, HSSEDI™ works to safeguard our nation against terrorist threats, aid the flow of legal commerce and immigration, and recover from natural disasters.

Judiciary Engineering and Modernization Center: Sponsored by the Administrative Office of the U.S. Courts on behalf of the federal judiciary, JEMC provides objective assessments of the technical challenges facing the judiciary including available and emerging technologies.

National Cybersecurity FFRDC: Sponsored by the National Institute of Standards and Technology, this FFRDC works to enhance cybersecurity and protect national information systems.

We are pleased to respond to your request for information regarding directions for research and determining challenges and opportunities in artificial intelligence (Al).

Our response will focus on topics related to Safety and Control for Al (topic # 3 in the RFI).

Safety and Control for Al

One general observation: there is a growing consensus that it is essential to develop Al systems that are safe and controllable, and to communicate the appropriate level of trust for the use of Al systems in critical roles.

A specific observation: there is not much collaboration evident yet between safety-critical systems communities and Al/machine learning communities. The Al community shows very little familiarity with basic techniques and understood limits to approaches that are familiar to the safety-critical community. MITRE has encountered common tools and techniques related to safety and assurance across a wide range of domains (e.g., weapons systems,

military and commercial aviation, medical devices and clinical decision support systems, automobiles, air traffic control, cyber security). A deliberate and sustained collaboration between the Al and safety-critical systems communities would be of significant value to the advancement of Al¬ based opportunities.

Two specific examples of opportunities for research to advance the application of safety-critical systems approaches to Al systems are structured assurance cases and advancing hazard analysis.

Assurance Cases for Critical At Systems

"Due to a lack of sufficient data to support or contradict any particular approach, a software system may not be declared dependable based on the method by which it was constructed. Rather, it should be regarded as dependable -certifiably dependable¬ only when adequate evidence has been marshaled in support of an argument for dependability that can be independently assessed. The goal of certifiably dependable software cannot therefore be achieved by mandating particular process and approaches, regardless of their effectiveness in certain situations. Instead, software developers should marshal evidence to justify an explicit dependability claim that makes clear which properties in the real world the system is intended to establish. Such evidence forms a dependability case, and creating a dependability case is the cornerstone of the committee's approach to developing certifiably dependable systems.
... Few, if any, existing certification regimes encompass the combination of characteristics recommended in this report-namely, explicit dependability claims, evidence for these claims, and a rigorous argument that demonstrates that the evidence is sufficient to establish the validity of the claims."

Software for Dependable Systems: Sufficient Evidence? National Research Council, 2007

We suggest that a research agenda for Al should include the development and assessment of assurance cases as a framework for verification, validation, and certification of complex Al systems. Assurance cases must become first-class engineering artifacts supporting rigorous analysis and scrutiny, not remain volumes of technical prose informally read and reviewed for certification. Focused research is required to develop notations, tools, and techniques useful for the analysis of assurance cases for Al systems.

There is no such thing as a fully autonomous system. Even humans must occasionally take orders and react to environmental changes. Most autonomous systems must sometimes cede control back to a human. This leads to the handoff problem, which can introduce safety risks. In fast and dangerous situations, like those that might occur in a self-driving car, that handoff must happen quickly. Often the human is not ready quickly enough or lacks the context to make a good decision. An example is Air France flight 447, which crashed into the ocean during a storm in 2009. In this case, a blocked air sensor caused the autopilot to turn

off. With the main pilot asleep, the two copilots had to quickly take control, but they lacked experience flying under such conditions and failed to understand the strange readings from the blocked air sensor. While this was clearly a case of a bad handoff, it also illustrates how important the design and "choreography" of human-machine interaction is in safety-critical systems. In fact, viewing these complex teams of human and machine agents as an all-or-nothing human is in charge vs. machine is in charge perspective is wholly inadequate. What is required is interdisciplinary research to enhance the design of a complex spectrum of shared roles among humans and machines, where aspects of control shift fluidly as best fits the situation. Ensuring that these shifting roles preserve safety and control is crucial, as the Air France example illustrates.

However, the ease with which humans have already integrated computational systems into decision making ranging from ordinary to critical, from simple to complex, belies a deeper truth: this area of inquiry is still in its infancy relative to where multi¬ disciplinary research could take it over the next generation. This state of affairs has generated an environment that is ripe for a rethinking of human-computer collaboration in the context of complex decision making. The vast amount of information that can be brought to bear does not guarantee better decisions or a more straightforward or reliable decision-making process.

Complex Operational Decision Making in Networked Systems of Humans and Machines, National Research Council, 2014

The verification and validation of complex assurance cases for Al will likely require human insight, oversight, and perhaps foresight. An assurance case focus is consistent with a variety of related activity in the United States and Europe for regulated and safety-critical software¬ intensive systems; focusing this research on Al systems will provide support for continued harmonization with European and other assurance and certification approaches for Al systems as they evolve.

Hazard Analysis for Al Systems
"Be careful how you fix what you don't understand." Fred Brooks, The Design of Design

Much of the considerable effort in Al systems (industry, academia, government) is focused on extending our reach-exploring new capabilities and applications. However, more rigorous work is also needed to understand the performance envelopes and hazards in increasingly critical Al systems. A range of powerful safety analysis tools and techniques exist that have not been widely applied to emerging Al systems, and that could produce a wealth of new insights into risks that must be mitigated and paths that may be taken to accomplish such mitigation. There is also fundamental new research to be done to develop new safety analysis tools and techniques to keep pace with the rapidly evolving technologies applied in Al systems.

Much work has been done in the system and software safety communities on both

traditional (e.g., Hazards and Operability Studies, Failure Modes and Effects Analysis, Fault Tree Analysis) and emerging (e.g., Systems Theoretic Accident Model and Processes, Systems-Theoretic Process Analysis, Architecture-Led Safety Analysis) analysis techniques, but very little of this work has been applied systematically to Al systems. There have been some analyses of human/automation interaction failure modes, and of novel failure modes in machine learning (deep learning, support vector machines, etc.). With very few exceptions, attention on mitigating risks in adopting Al for critical applications is focused on either strengthening the underlying trust model (to reduce the risks of over-trust and under-trust, and including the cognitive and human factor aspects of trust decisions), or mitigating already identified potential hazards (e.g., addressing the limitations to acceptance testing for learning adaptive systems by shifting to more runtime instrumentation and monitoring). Both of these areas of attention are of course critical. What is also needed is a sustained focus on uncovering novel failure modes and potential hazards introduced by reliance on novel emerging Al systems that in many ways the community do not yet adequately understand.

The community is increasingly putting Al systems in mission- or safety critical-roles, in many cases with an incomplete understanding of the potential failure modes and hazards introduced. Some of these failure modes have been recognized for a long time in human-machine interactions (HMls), but often have not been given adequate attention in the growing application of emerging Al systems to critical roles. HMI safety concerns include human cognitive biases, such as the combination of confirmation bias and automation bias, which can lead to false confidence, mixed communication modalities, and shared contexts between humans and cognitive systems that cause misunderstandings; and the speed and complexity of machine reasoning, which can pose communication challenges, especially when the HMI team includes multiple machines that can communicate directly with one another.

In other cases, underlying mechanisms such as machine learning introduce some novel failure modes. One missing area in research is learning causal models (vice statistical ones). This might help to advance explainable systems, and perhaps mitigate some aspects of the HMI safety concerns previously described. And the multidisciplinary challenges in fully identifying failure modes and hazards in critical Al systems have typically received less attention than demonstrations of the latest impressive new capability. Rigorous and systematic explorations of failure modes related to Al systems have been relatively rare, but when they are performed they have been widely cited and used to improve the collective understanding of the potential risks and mitigations. OSTP could promote research focused on the next steps in understanding the performance envelopes and potential hazards in Al systems, to accelerate the development of mitigations and to shape the informed adoption of this technology.

Respondent 159

Adam Theirer, Mercatus Center at George Mason University

PREPARING FOR THE FUTURE OF ARTIFICIAL INTELLIGENCE
ADAM THIERER
Senior Research Fellow, Mercatus Center at George Mason University
ANDREA CASTILLO
Program Manager, Mercatus Center at George Mason University
Request for Information on Artificial Intelligence Agency: Office of Science and Technology Policy Proposed: June 27, 2016
Comment period closes: July 22, 2016
Submitted: July 22, 2016
Document Number: 2016 -15082

The Office of Science and Technology Policy (OSTP) has requested comments pertaining to the governance of artificial intelligence (AI) technologies. 1

The Technology Policy Program of the Mercatus Center at George Mason University is dedicated to advancing knowledge of the impact of regulation on society. It conducts careful and independent analyses employing contemporary economic scholarship to assess policy issues from the perspective of the public interest.

We write here to comment on the appropriate policy framework for artificial intelligence (AI) technologies at this nascent stage of their development and to make the case for prudence, patience, and a continuing embrace of "permissionless innovation."Permissionless innovation refers to the idea that "experimentation with new technologies and business models should generally be permitted by default. Unless a compelling case can be made that a never invention will bring serious harm to society, innovation should be allowed to continue unabated and problems, if they develop at all, can be addressed later."2

Policymakers may be tempted to preemptively restrict AI technologies out of an abundance of caution for the perceived risks these new innovations might seem to pose. However, an examination of the history of US technology policy demonstrates that these concerns can be adequately addressed without quashing a potentially revolutionary new industry.

Specifically, as policymakers consider the governance of AI, they would be wise to consider the lessons that can be drawn from our recent experience with the Internet. The United States made permissionless innovation the basis of Internet policy beginning in the early 1990s, and it soon became the "secret sauce" that propelled the rise of the modern digital revolution. 3

If policymakers wish to replicate America's success with the Internet, they need to adopt a similar "light-touch" approach for the governance of AI technologies. To highlight the

benefits of permissionless innovation, the Mercatus Center at George Mason University has recently published a book,4 a series of law review articles, and several agency filings that explain what this policy vision entails for different technologies and sectors.5 A summary of the major insights from these studies can be found in a recent Mercatus Center paper called "Permissionless Innovation and Public Policy: A IO-Point Blueprint."6

If one's sole conception of a technology comes from Hollywood depictions of dystopian science fiction or killer robotic systems run amok, it is understandable that one might want to use the force of regulation to clamp down decisively on these "threats." But these fictional representations are just that: fictional. AI technologies are both much more benign and fantastic in reality.

The economic benefits of AI are projected to be enormous. One recent study used benchmarks derived from methodologically conservative studies of broadband Internet, mobile phones, and industrial robotics to estimate that the economic impact of AI could be between $1.49 trillion and $2.95 trillion over the next ten years.7 With less strict assumptions, the economic benefits could be greater still.

However, some skeptics are already making the case for a preemptive regulation of AI technologies. The rationales for control are varied, including concerns ranging from deindustrialization to dehumanization, 8 as well as worries about the "fairness" of the algorithms behind AI systems.9

Due to these anxieties associated with AI, some academics argue that policymakers should "legislate early and often" to "get ahead of" these hypothetical problems. 10 Specifics are often in short supply, with some critics simply hinting that "something must be done" to address amorphous concerns.

Other scholars have provided more concrete regulatory blueprints, however. They propose, among other things, the passage of broad-based legislation 11 such as an "Artificial Intelligence Development Act,"12 as well as the creation of a federal AI agency 13 or possibly a "Federal Robotics Commission"14 or "National Algorithmic Technology Safety Administration."15 These proposed laws and agencies would establish a certification process requiring innovators to subject their technologies to regulatory review to "ensure the safety and security of their A.I."16 Or, at a minimum, such agencies would advise other federal, state, and local officials and organizations on how to craft policy for AI and robotics.

Such proposals are based on "precautionary principle" reasoning. The precautionary principle refers to the belief that new innovations should be curtailed or disallowed until their developers can prove that they will not cause any harms to individuals, groups, specific entities, cultural norms, or various existing laws, norms, or traditions.

It is certainly true that AI technologies might give rise to some of the problems that critics

suggest. And we should continue to look for constructive solutions to the potentially thorny problems that some of these critics discuss. That does not mean that top-down, technocratic regulation is sensible, however.

Traditional administrative regulatory systems have a tendency to be overly rigid, bureaucratic, and slow to adapt to new realities. This is particularly problematic as it pertains to the governance of new, fast-moving technologies.

Prior restraints on innovative activities are a recipe for stagnation. By focusing on preemptive remedies that aim to predict hypothetical problems that may not ever come about, regulators run the risk of making bad bets based on overconfidence in their ability to predict the future.17 Worse yet, by preempting beneficial experiments that yield new and better ways of doing things, administrative regulation stifles the sort of creative, organic, bottom-u p solutions that will be needed to solve problems that may be unforeseeable today.18

This risk is perhaps more pronounced when dealing with AI technologies. How "artificial intelligence" is regulated makes little sense until policymakers define what it actually entails. The boundaries of AI are amorphous and ever changing. AI technologies are already all around us-examples include voice-recognition software, automated fraud detection systems, and medical diagnostic technologies-and new systems are constantly emerging and evolving rapidly.19 Policymakers should keep in mind the rich and distinct variety of opportunities presented by AI technologies, lest regulations more appropriate for one kind of application inadvertently stymie the development of another.20

Toward that end, we suggest that a different policy approach for AI is needed, one that is rooted in humility and a recognition that we possess limited knowledge about the future.21

This does not mean there is no role for government as it pertains to AI technologies. But it does mean that policymakers should first seek out less restrictive remedies to complex social and economic problems before resorting to top-down proposals that are preemptive and proscriptive.

Policymakers must carefully ensure they have a full understanding of the boundaries and promises of all of the technologies they address. Many AI technologies pose little or no risks to safety, fair market competition, or consumer welfare. These applications should not be stymied due to an inappropriate regulatory scheme that seeks to address an entirely separate technology. They should be distinguished and exempted from regulations as appropriate.

Other AI technologies may warrant more regulatory consideration if they generate substantial risks to public welfare. Still, regulators should proceed cautiously. To the extent that policymakers wish to spur the development of a wide array of new life¬ enriching

technologies, while also looking to devise sensible solutions to complex challenges, policymakers should consider a more flexible, bottom-up, permissionless innovation approach as the basis of America's policy regime for AI technologies.

1. Ed Felten, “How to Prepare for the Future of Artificial Intelligence,” White House blog, June 27, 2016
2. Adam Thierer, Permissionless Innovation: The Continuing Case for Comprehensive Technological Freedom (Arlington, VA: Mercatus Center at George Mason University, 2016).
3. Adam Thierer, "Embracing a Culture of Permissionless Innovation," Cato Online Forum, November 2014.
4. Adam Thierer, Permissionless Innovation.
5. Subjects include the Internet of Things, wearable devices, smart cars, commercial drones, cryptocurrency, 3D printing, robotics, the sharing economy, and advanced medical devices. Our research can be accessed at permissionlessinnovation.org.
6. Adam Thierer and Michael Wilt, "Permissionless Innovation: A 10-Point Checklist for Public Policymakers," Economic Perspectives, Mercatus Center at George Mason University, March 31, 2016.
7. Nicholas Chen et al., "Global Economic Impacts Associated with Artificial Intelligence" (Study, Analysis Group, Boston, MA, February 25, 2016), "Growth in Al producing sectors could lead to increased revenues, and employment within these existing firms, as well as the potential creation of entirely new economic activity. Productivity improve¬ments in existing sectors could be realized through faster and more efficient processes and decision making as well as increased knowledge and access to information."
8. Nicholas Carr, The Glass Cage: Automation and Us (New York: W. W. Norton & Company, 2014); Jerry Kaplan, Hu¬ mans Need Not Apply: A Guide to Wealth and Work in the Age of Artificial Intelligence (New Haven, CT: Yale University Press, 2015), 7. (Kaplan suggests that Al systems "can wreak havoc on an unimaginable scale in the blink of an eye.")
9. Frank Pasquale, The Black Box Society: The Secret Algorithms That Control Money and Information (Cambridge, MA: Harvard University Press, 2015).
10. John Frank Weaver. "We Need to Pass Legislation on Artificial Intelligence Early and Often," Slate, September 12, 2014.
11. Alex Rosenblat, Tamara Kneese and danah boyd, "Understanding Intelligent Systems" (Data & Society Working Paper. Data & Society Research Institute, October 8, 2014), 11.
12. Matthew U. Scherer, "Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies," Harvard Journal of Law and Technology 29, no. 2 (2016): 43-45. Also see Weaver, "We Need to Pass Legislation."
13. Matthew U. Scherer, "Regulating Artificial Intelligence Systems," 45-47.
14. Ryan Calo. "The Case for a Federal Robotics Commission" (Report, Brookings Institution, Washington, DC, September 2014).
15. Andrew Tutt, "An FDA for Algorithms" (working paper, 2016), available through SSRN at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2747994.
16. Scherer, "Regulating Artificial Intelligence Systems," 43
17. Thierer, Permissionless Innovation, 82. ("Trying to preemptively plan for every

hypothetical worst-case scenario means that many best-case scenarios will never come about.'')
18. Aaron Wildavsky, Searching for Safety (New Brunswick, CT: Transaction Books, 1988). 183. ("Regulation, because it deals with the general rather than with the particular, necessarily results in forbidding some actions that might be beneficial. Regulators cannot devise specifications sufficiently broad to serve as guidelines for every contingency without also limiting some actions that might increase safety. Because regulation is anticipatory, regulators frequently guess wrong about which things are dangerous; therefore, they compensate by blanket prohibitions.")
19. AJ Agrawal, "7 Ways Artificial Intelligence Is Improving Consumer Experiences," Customer Think, July 14, 2016.
20. Robert D. Atkinson, '"It's Going to Kill Us!' and Other Myths about the Future of Artificial Intelligence" (Report, Information Technology and Innovation Foundation, June 2016), 10. ("If we want progress-an increase in economic growth, improved health, a better environment, etc.-then it is time to regain our sense of optimism about the promise of technological innovation," argues Robert Atkinson of ITIF. "In particular, when it comes to Al, we should be enthusi¬astic and excited, not fearful and cautious.")
21. Maureen K. Ohlhausen, "The Internet of Things and the FTC: Does Innovation Require Intervention?" (Remarks before the US Chamber of Commerce, Federal Trade Commission, Washington, DC, October 18, 2013). ("It is . . . vital that government officials, like myself, approach new technologies with a dose of regulatory humility, by working hard to educate ourselves and others about the innovation, understand its effects on consumers and the marketplace, identify benefits and likely harms, and, if harms do arise, consider whether existing laws and regulations are sufficient to address them, before assuming that new rules are required.")

Respondent 160

Fusun Yaman , BBN Technologies

Preparing for the Future of Artificial Intelligence
Written by Fusun Yaman and Aaron Adler, XXXXXXXXX BBN Technologies, Cambridge, MA, USA

(5) the most pressing, fundamental questions in AI research, common to most or all scientific fields;

The fundamental question is how can we characterize an AI system in terms of its requirements, functionality, strengths, and weaknesses so that utilizing AI effectively in any domain does not require an AI expert.

There are textbooks full of AI techniques, and although the application of some may be formulaic, this simplicity is deceiving. As cross-disciplinary applications of AI grow, it is

critical that the proper techniques be applied to the proper problems, and further that the underlying biases of the input data or technique itself is understood. It is important to know which problems or questions the combination of a particular set of data and technique can answer, and equally if not more importantly which they cannot. For example, the danger of drawing conclusions about all mice (or all humans) based only on a study of male mice, selected because female mice have hormones that make experimentation more difficult.

In order to address this issue, we propose that a meta-AI system is needed, where data and technique can be annotated with their properties. This system would then reason over these inputs and be able to describe potential pitfalls. This would help ensure that applications of Al were consistent and sound. This is especially important for non-experts or the public that can easily download and apply machine learning libraries, but are less likely to be able to select appropriate input data or draw the correct inferences from the output. This is a challenging and fundamental problem, and could perhaps head off problems that can also arise in statistics where it is possible to intentionally or unintentionally influence the results based on the input samples and methodology used.

If instead this question is asking what is the fundamental goal of Al, i.e., similar to the goal of physics being to understand how the universe behaves, then AI is the study of teaching or programming a computer, most of which process data as ones and zeros, to learn and interact with the world, humans primarily, and be a productive and responsible part of the society. And by more deeply understanding this, what does it teach us about ourselves.

(6) the most important research gaps in AI that must be addressed to advance this field and benefit the public;

Ability to work with humans: For maximum benefit to public, AI based reasoners and decision support systems should be able to work with humans as opposed to work for humans. Working with humans requires gaining acceptance as trusted teammates. Such trust will foster from a combination of the following core capabilities: 1) naturally interacting with humans, 2) intuitively explaining rationale, 3) repeatedly demonstrating proficiency in a task. Depending on the goal and the team dynamics the importance of these three core capabilities might vary, as is the case with human teammates. As of yet, no AI systems have fully addressed any of the above areas.

(7) the scientific and technical training that will be needed to take advantage of harnessing the potential of AI technology;

Much as not everyone needs to be trained as a doctor to take advantage of centuries of medical knowledge, not everyone should have to be a trained AI researcher to take advantage of AI advances. Our previous discussion of a meta-AI system would help make advances in AI accessible to the general public and enable more cross¬ disciplinary research and advances. That said, clearly general computer literacy is important, in a similar way that

general biology, physics, chemistry, geology, etc., is important. The usage of the AI system will not require a lot of expertise if people know the core high-level concepts.

To extend the doctor analogy, just as you have genera list and specialist doctors, we envision that you will have generalist AI experts, but also specialists in neural nets, or Support Vector Machines (SVMs). The experts can focus on evaluating the existing methods, enabling the meta-AI-reasoner, and coming up with new solutions.

(8) the specific steps that could be taken by the federal government, research institutes, universities, and philanthropies to encourage multi-disciplinary AI research;

In order to encourage such interdisciplinary work, the academic world should not punish researchers for trying to bridge different fields. There is also a need to address the notion that the interdisciplinary researchers are not seen as first class citizens in either field.
From a funding perspective, additional funding for interdisciplinary work, including funding of long term basic research in AI. There are many difficult problems and expecting new, fundamentally different or advanced approaches is not compatible with expectations of results within 1-3 years. Finally, we would suggest more collaboration within funding agencies. If the researchers are expected to collaborate, then groups within the funding agencies should be encouraged to work together.

Clear cut research/funding areas leaves interdisciplinary researchers in a tough spot as all of the communities may feel like the research is out of their sweet spot or that they are not equipped to evaluate it.
In sum, the artificial lines between disciplines, for funding and academic evaluation, need to be blurred to better support interdisciplinary research.

Respondent 161

, TOYOTA

TOYOTA
Office of Science and Technology Policy Request for Information on Artificial Intelligence
RESPONSE
July 22, 2016

Background

As the largest automaker in the world, Toyota recognizes the potential that artificial intelligence (Al) has to save lives and improve the quality of life for millions of people globally. In 20 l 5, Toyota announced the formation of the Toyota Research Institute (TRI), a company focused on AI research and development. TRI is headquartered in Palo Alto, California, and has additional research facilities in Cambridge, Massachusetts, and Ann

Arbor, Michigan. TRI's CEO, Gill Pratt, is the former program manager of the Defense Advanced Research Projects Agency (DARPA) robotics challenge.

TRI is focusing its efforts on bridging the gap between fundamental research and product development with the goal of strategically leveraging AI technology to save and improve the quality of lives. TRl's goals are to: create a car that is incapable of causing a crash, increase access to cars for those who otherwise cannot drive (including the disabled and the elderly), help translate outdoor mobility technology into indoor mobility products, and accelerate scientific discovery by applying techniques from AI and machine learning. Toyota's approach to automated vehicles (AVs) is rooted in the belief that human-AI partnership will result in safer outcomes. For that reason, TRI is working diligently and carefully to develop Al technologies that can more consistently and reliably protect human drivers, even in highly uncertain situations.

The term AI is used in a variety of different contexts and with different meanings. TRl's Al system will make decisions based on algorithmic processes whose inputs include "live" sensor data, previously-recorded (or simulated) data, and a mathematical function that measures the quality of the AI system's decisions. Al algorithms attempt to maximize the quality of the outcome by making the "best" choice from a set of options. Both the set of options and the mathematical function used by the AI are designed by human engineers, which means that the AI cannot make decisions beyond the prescribed set of options, nor can it modify the mathematical function.

Al for Public Good

The National Highway Traffic Safety Administration (NHTSA) recently estimated that 35,200 people died in motor vehicle traffic crashes in 2015, representing a 7.7% increase from the previous year.1 Since the vast majority of crashes are caused by human error2, AV technology has the potential to significantly reduce the number of lives lost in vehicle crashes.
In addition, AVs are likely to result in a net decrease in long-term traffic congestion and vehicle emissions. A reduction in congestion caused by collisions, efficient driving that decreases idling and rapid acceleration, and optimized routes that minimize travel time are each likely to contribute to a decline in vehicle emissions.

Al also holds immense promise for increasing mobility and independence for the elderly and disabled. In addition to AV technology that can provide increased access to vehicles for these groups, TRI is working to develop indoor mobility products that will provide these individuals with a greater degree of ease and comfort when navigating indoor spaces. These advancements in personal mobility capabilities will provide typically-underserved groups with new options to function safely in their homes and communities, contributing to a greater sense of individual independence and dignity.

AI Safety and Control Issues

While the benefits of AV technology are great, realizing the technology's potential is not without challenges. Toyota sells about 10 million vehicles each year globally, and each of those vehicles is typically on the road for at least 10 years. This means that, at any given point in time, there are approximately 100 million Toyota vehicles 011 the road throughout the world. If one assumes that a vehicle drives about 10,000 miles each year, the global fleet of Toyota vehicles collectively drive approximately 1 trillion miles per year.

Although most driving is relatively easy and predictable, some driving is quite difficult. It is when driving is difficult that Al will likely play the most critical role. Even if only 1% of all driving is difficult, Toyota vehicles alone would be engaged in l0 billion miles of difficult driving out of 1 trillion total miles driven each year. The challenge of delivering autonomous vehicle technology that is capable of achieving "trillion mile" reliability and functioning appropriately in most - if not all - of those 10 billion miles of difficult driving cannot be understated.

As a result, TRI is pursuing two paths to autonomy. The first is series autonomy, which can be described as ''chauffeur mode." Under "chauffeur mode," the AV technology takes over the driving task from the human driver completely. Since a human driver is not in the loop, this type of AV technology needs to be able to perform at all times and in all circumstances. The ''chauffeur mode" is arguably the best model for providing mobility to the elderly and disabled, who would otherwise be limited by vehicle systems that require a human driver. The second is parallel autonomy, which can be described as "guardian angel mode." Under "guardian angel mode," the AV technology acts as a high-level driver assist system. The "guardian angel mode" is always monitoring the environment, but steps in only when a collision is imminent. If the objective is safety and reducing the number of traffic-related fatalities, the "guardian angel mode" may arguably be as effective as the "chauffeur mode." Moreover, with "guardian angel mode," these safety benefits can be realized in the nearer term.

Millions of test-driven miles is probably not sufficient to achieve the trillion-mile reliability that TRI seeks for its autonomous vehicle technology. As a result, TRI is dedicating a significant portion of its work to Al computer simulation to accelerate and expand the range of testing of these systems.

Legal and Governance Implications of AI

Since AI is an emerging technology, regulatory frameworks are premature. Preemptive legislation and regulation premised on a possible risk of potential harm would needlessly stifle the growth and development of a technology that has the potential to vastly improve safety and quality of life for millions globally. Instead, government should rely on industry to develop any consensus¬ driven standards and best practices that may be needed.

Since AI is being used across industries and sectors, it requires a cohesive and flexible interagency approach. The government's approach should also account for the fact that Al technology will be used by both highly-regulated and traditionally unregulated sectors. If AI is subjected to restrictive and cautionary regulatory frameworks in some sectors, and allowed to flourish unimpeded in traditionally unregulated sectors, beneficial technological advances may be arbitrarily stymied in some industries to the detriment of the public good. For that reason, a common government-wide approach to AI is appropriate.

The federal government should consider whether regulatory agencies have sufficient resources and expertise to handle the emergence of Al. There may also be a need to review existing laws to determine whether agencies have sufficient authority to accommodate, or even encourage the deployment of AI, and whether any existing statutory or regulatory provisions would present unnecessary obstacles for Al deployment.

AI technology will likely raise questions about data privacy and cybersecurity. However, these questions are fundamentally the same as questions that have historically been raised during the emergence of other information and communications technologies. In fact, many industries planning for near-term Al deployment have already designed and implemented robust data privacy and cybersecurity protections for similar technologies.3 The lessons learned in these other areas can be used to promote secure Al technologies that protect consumer privacy.

Widespread consumer acceptance requires AI to make choices that reflect society's collective understanding of "ethical" behavior, even in the most difficult driving situations. Al researchers spend a great deal of time designing these systems to be compatible with what humans would typically consider the "right" behavior. The AV uses an algorithm that maximizes the utility of the outcome and most often makes a decision that matches what would be expected of a typical human driver. Even then, if the vehicle encounters a "trolley" situation (e.g., where it could save the driver but harm a pedestrian, or harm the driver but save the pedestrian), the Al will - as a human driver would - necessarily have to make a choice. As with a human driver, either decision could subject the Al to legal and ethical scrutiny. It is unclear whether the AI system and its developers would be subject to the reasonable person standard (that is, the standard of care by which a human driver would be judged) or some higher standard (such as strict liability).

Despite the good performance of a human driver, an unlikely combination of factors may still contribute to a collision and make it difficult to assign moral blame. Similarly, it is conceptually possible for an AV to reach a level of performance such that, if a crash occurs, moral fault and culpability become equally difficult to assess. At the same time, just as human drivers sometimes make incorrect decisions, it is possible (although arguably less 'likely) that the Al will generate an incorrect 'human" decision. A decision that results in harm to one or more individuals should not necessarily undermine the aggregate lives

saved by the same technology.

Social and Economic Implications of Al

AI technology is likely to expand access to and enhance the quality of services provided to citizens. Improved vehicle safety technologies will likely shrink the number of driver, cyclist, and pedestrian fatalities and injuries that take place due to human error. AV technology will also likely contribute to resolution of the first and last-mile problem that many members of society face, including those from low-income and remote communities. Government has the opportw1ity to leverage data and Al to predict needs for housing, education, healthcare, public assistance, and public transportation and more proactively and precisely implement effective solutions.

There is concern that Al will result in job loss and worker displacement. While these concerns are reasonable, it is important to understand that the majority of emerging AI is "narrow AI," designed to perform a specific task. This type of Al means that a portion - but most likely not all - of an employee's job might be replaced or made easier by AI, freeing up time to focus on other responsibilities. While the impact of AI on driver-related jobs is not yet fully understood, it will likely differ depending on the specific mode of autonomy employed. For example, in the case of AVs, "guardian angel mode" parallel autonomy maintains the central role for the human driver and would presumably displace fewer driver-related jobs.

Fundamental Questions in AI Research

One concern specific to AVs is the driver-vehicle handoff problem. Series autonomy at various levels may require the driver to take over operation of the vehicle if the system encounters a situation that it cannot handle. Researchers are striving, but struggling, to delineate the boundaries for if, when, and how AV systems should disengage and command driver engagement. The dilemma is how to design a system that ensures driver awareness and attention precisely when it is needed, when the driver may have a latent expectation that he or she does not need to remain alert.

Cross-Sector Collaboration in Al Research and Training

The federal government can facilitate Al-related AV research by supporting partnerships focusing on AV research and development. For example, the Mcity test environment was designed and developed by the University of Michigan's interdisciplinary Mobility Transformation Center (MTC), a partnership among industry, government and academia. The federal government has the opportunity to create similar test environments that not only provide testing locations for companies, but also harness the power of cross-sector collaboration.

Finally, the federal government should continue to encourage partnerships among school

districts and universities, science agencies, businesses, and other community partners to prepare the next generation of leaders in Al. Investment in STEM programs and an emphasis on computer science proficiency will equip the workforce of the future with the skills to program, operate, audit, and advance Al in the decades to come.

1. Karen Aldana , "NHTSA Data Shows Traffic Deaths Up 7.7 Percent in 2015," USDOT NHTSA Press Release, July 1, 20I6, http://www.nhtsa.gov/About+NHTSA/Press+Releases/2015/2014-traffic-deaths-drop-but-2015-trending¬higher.
2. "Traffic Safety Facts: A Brief Statistical Summary," USDOT NHTSA Publication, February 2015, https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812115.
3. For example, in November of 2014, the auto industry unveiled the Privacy Principles for Vehicle Technologies and Services. The Principles were designed to address current vehicle technologies, as well as future vehicle technologies such as AVs, and include meaningful consumer protections on the collection and use of vehicle data.