Chapter 11
The Biointelligence Explosion

How Recursively Self-Improving Organic Robots will Modify their Own Source Code and Bootstrap Our Way to Full-Spectrum Superintelligence

David Pearce

Abstract This essay explores how recursively self-improving organic robots will modify their own genetic source code and bootstrap our way to full-spectrum superintelligence. Starting with individual genes, then clusters of genes, and eventually hundreds of genes and alternative splice variants, tomorrow's biohackers will exploit "narrow" AI to debug human source code in a positive feedback loop of mutual enhancement. Genetically enriched humans can potentially abolish aging and disease; recalibrate the hedonic treadmill to enjoy gradients of lifelong bliss, and phase out the biology of suffering throughout the living world.

> *Homo sapiens, the first truly free species, is about to decommission natural selection, the force that made us….*
> *Soon we must look deep within ourselves and decide what we wish to become.*
> Edward O. Wilson
> Consilience, The Unity of Knowledge (1999)

> *I predict that the domestication of biotechnology will dominate our lives during the next fifty years at least as much as the domestication of computers has dominated our lives during the previous fifty years.*
> *Freeman Dyson*
> New York Review of Books (July 19 2007)

D. Pearce (✉)
BLTC Reseerch, 7 Lower Rock Gardens, Brighton, BN2 1PG, UK
e-mail: dave@hedweb.com

A. H. Eden et al. (eds.), *Singularity Hypotheses*, The Frontiers Collection,
DOI: 10.1007/978-3-642-32560-1_11,

The Fate of the Germline

Genetic evolution is slow. Progress in artificial intelligence is fast (Kurzweil 2005). Only a handful of genes separate *Homo sapiens* from our hominid ancestors on the African savannah. Among our 23,000-odd protein-coding genes, variance in single nucleotide polymorphisms accounts for just a small percentage of phenotypic variance in intelligence as measured by what we call IQ tests. True, the tempo of human evolution is about to accelerate. As the reproductive revolution of "designer babies" (Stock 2002) gathers pace, prospective parents will pre-select alleles and allelic combinations for a new child *in anticipation of* their behavioural effects—a novel kind of selection pressure to replace the "blind" genetic roulette of natural selection. In time, routine embryo screening via preimplantation genetic diagnosis will be complemented by gene therapy, genetic enhancement and then true designer zygotes. In consequence, life on Earth will also become progressively happier as the hedonic treadmill is recalibrated. In the new reproductive era, hedonic set-points and intelligence alike will be ratcheted upwards in virtue of selection pressure. For what parent-to-be wants to give birth to a low-status depressive "loser"? Future parents can enjoy raising a normal transhuman supergenius who grows up to be faster than Usain Bolt, more beautiful than Marilyn Monroe, more saintly than Nelson Mandela, more creative than Shakespeare—and smarter than Einstein.

Even so, the accelerating growth of germline engineering will be a *comparatively* slow process. In this scenario, sentient biological machines will design cognitively self-amplifying biological machines who will design cognitively self-amplifying biological machines. Greater-than-human biological intelligence will transform itself into posthuman superintelligence. Cumulative gains in intellectual capacity and subjective well-being across the generations will play out over hundreds and perhaps thousands of years—a momentous discontinuity, for sure, and a twinkle in the eye of eternity; but not a Biosingularity.

Biohacking Your Personal Genome

Yet **germline** engineering is only one strand of the genomics revolution. Indeed, after humans master the ageing process (de Grey 2007), the extent to which traditional germline or human generations will persist in the post-ageing world is obscure. Focus on the human germline ignores the slow-burning but then explosive growth of **somatic** gene enhancement in prospect. Later this century, innovative gene *therapies* will be succeeded by gene *enhancement* technologies—a value-laden dichotomy that reflects our impoverished human aspirations. Starting with individual genes, then clusters of genes, and eventually hundreds of genes and alternative splice variants, a host of recursively self-improving organic robots ("biohackers") will modify their genetic source code and modes of sentience: their senses, their moods, their motivation, their cognitive apparatus, their world-simulations and their default state of consciousness.

As the era of open-source genetics unfolds, tomorrow's biohackers will add, delete, edit and customise their own legacy code in a positive feedback loop of cognitive enhancement. Computer-aided genetic engineering will empower biological humans, transhuman and then posthuman to synthesise and insert new genes, variant alleles and even designer chromosomes—reweaving the multiple layers of regulation of our DNA to suit their wishes and dreams rather than the inclusive fitness of their genes in the ancestral environment. Collaborating and competing, next-generation biohackers will use stem-cell technologies to expand their minds, literally, via controlled neurogenesis. Freed from the constraints of the human birth canal, biohackers may re-sculpt the prison-like skull of *Homo sapiens* to accommodate a larger mind/brain, which can initiate recursive self-expansion in turn. Six crumpled layers of neocortex fed by today's miserly reward pathways aren't the upper bound of conscious mind, merely its seedbed. Each biological neuron and glial cell of your growing mind/brain can have its own dedicated artificial healthcare team, web-enabled nanobot support staff, and social network specialists; compare today's anonymous neural porridge. Transhuman minds will be augmented with neurochips, molecular nanotechnology (Drexler 1986), mind/computer interfaces, and full-immersion virtual reality (Sherman 2002) software. To achieve finer-grained control of cognition, mood and motivation, genetically enhanced transhumans will draw upon exquisitely tailored new designer drugs, nutraceuticals and cognitive enhancers—precision tools that make today's crude interventions seem the functional equivalent of glue-sniffing.

By way of comparison, early in the twenty-first century the scientific counterculture is customizing a bewildering array of designer drugs (Shulgin 1995) that outstrip the capacity of the authorities to regulate or comprehend. The bizarre psychoactive effects of such agents dramatically expand the evidential base that our theory of consciousness (Chalmers 1995) must explain. However, such drugs are short-acting. Their benefits, if any, aren't cumulative. By contrast, the ability genetically to hack one's own source code will unleash an exponential growth of genomic rewrites—not mere genetic tinkering but a comprehensive redesign of "human nature". Exponential growth starts out almost unnoticeably, and then explodes. Human bodies, cognition and ancestral modes of consciousness alike will be transformed. Post-humans will range across immense state-spaces of conscious mind hitherto impenetrable because access to their molecular biology depended on crossing gaps in the fitness landscape (Langdon and Poli 2002) prohibited by natural selection. Intelligent agency can "leap across" such fitness gaps. What we'll be leaping into is currently for the most part unknown: an inherent risk of the empirical method. But mastery of our reward circuitry can guarantee such state-spaces of experience will be glorious beyond human imagination. For intelligent biohacking can make unpleasant experience physically impossible (Pearce 1995) because its molecular substrates are absent. Hedonically enhanced innervation of the neocortex can ensure a rich hedonic tone saturates whatever strange new modes of experience our altered neurochemistry discloses.

Pilot studies of radical genetic enhancement will be difficult. Randomised longitudinal trials of such interventions in long-lived humans would take decades. In fact officially licensed, well-controlled prospective trials to test the safety and efficacy of genetic innovation will be hard if not impossible to conduct because all of us, apart from monozygotic twins, are genetically unique. Even monozygotic twins exhibit different epigenetic and gene expression profiles. Barring an ideological and political revolution, most formally drafted proposals for genetically-driven life-enhancement probably won't pass ethics committees or negotiate the maze of bureaucratic regulation. But that's the point of biohacking (Wohlsen 2011). By analogy today, if you're technically savvy, you don't want a large corporation controlling the operating system of your personal computer: you use open-source software instead. Likewise, you don't want governments controlling your state of mind via drug laws. By the same token, tomorrow's biotech-savvy individualists won't want anyone restricting our right to customise and rewrite our own genetic source code in any way we choose.

Will central governments try to regulate personal genome editing? Most likely yes. How far they'll succeed is an open question. So too is the success of any centralised regulation of futuristic designer drugs or artificial intelligence. Another huge unknown is the likelihood of state-sponsored designer babies, human reproductive cloning, and autosomal gene enhancement programs; and their interplay with privately-funded initiatives. China, for instance, has a different historical memory from the West.

Will there initially be biohacking accidents? Personal tragedies? Most probably yes, until human mastery of the pleasure-pain axis is secure. By the end of the next decade, every health-conscious citizen will be broadly familiar with the architecture of his or her personal genome: the cost of personal genotyping will be trivial, as will be the cost of DIY gene-manipulation kits. Let's say you decide to endow yourself with an extra copy of the N-methyl D-aspartate receptor subtype 2B (NR2B) receptor, a protein encoded by the GRIN2B gene. Possession of an extra NR2B subunit NMDA receptor is a crude but effective way to enhance your learning ability, at least if you're a transgenic mouse. Recall how Joe Tsien (1999) and his colleagues first gave mice extra copies of the NR2B receptor-encoding gene, then tweaked the regulation of those genes so that their activity would increase as the mice grew older. Unfortunately, it transpires that such brainy "Doogie mice"—and maybe brainy future humans endowed with an extra NR2B receptor gene—display greater pain-sensitivity too; certainly, NR2B receptor *blockade* reduces pain and learning ability alike. Being smart, perhaps you decide to counteract this heightened pain-sensitivity by inserting and then over-expressing a high pain-threshold, "low pain" allele of the SCN9A gene in your nociceptive neurons at the dorsal root ganglion and trigeminal ganglion. The SCN9A gene regulates pain-sensitivity; nonsense mutations abolish the capacity to feel pain at all (Reimann et al. 2010). In common with taking polydrug cocktails, the factors to consider in making multiple gene modifications soon snowball; but you'll have heavy-duty computer software to help. Anyhow, the potential pitfalls and makeshift solutions illustrated in this hypothetical example could be multiplied in the face of a combinatorial explosion of possibilities on the horizon. Most risks—and opportunities—of genetic self-editing are presumably still unknown.

It is tempting to condemn such genetic self-experimentation as irresponsible, just as unlicensed drug self-experimentation is irresponsible. Would you want your teenage daughter messing with her DNA? Perhaps we may anticipate the creation of a genetic counterpart of the Drug Enforcement Agency to police the human genome and its transhuman successors. Yet it's worth bearing in mind how each act of sexual reproduction today is an unpoliced genetic experiment with unfathomable consequences too. Without such reckless genetic experimentation, none of us would exist. In a cruel Darwinian world, this argument admittedly cuts both ways (Benatar 2006).

Naively, genomic source code self-editing will always be too difficult for anyone beyond dedicated cognitive elite of recursively self-improving biohackers. Certainly there are strongly evolutionarily conserved "housekeeping" genes that archaic humans would be best advised to leave alone for the foreseeable future. Granny might do well to customize her Windows desktop rather than her personal genome—prior to her own computer-assisted enhancement, at any rate. Yet the Biointelligence Explosion won't depend on more than a small fraction of its participants mastering the functional equivalent of machine code—the three billion odd 'A's, 'C's, 'G's and 'T's of our DNA. For the open-source genetic revolution will be propelled by powerful suites of high-level gene-editing tools, insertion vector applications, nonviral gene-editing kits, and user-friendly interfaces. Clever computer modelling and "narrow" AI can assist the intrepid biohacker to become a recursively self-improving genomic innovator. Later this century, your smarter counterpart will have software tools to monitor and edit every gene, repressor, promoter and splice variant in every region of the genome: each layer of epigenetic regulation of your gene transcription machinery in every region of the brain. This intimate level of control won't involve just crude DNA methylation to turn genes off and crude histone acetylation to turn genes on. Personal self-invention will involve mastery and enhancement of the histone and micro-RNA codes to allow sophisticated fine-tuning of gene expression and repression across the brain. Even today, researchers are exploring "nanochannel electroporation" (Boukany et al. 2011) technologies that allow the mass-insertion of novel therapeutic genetic elements into our cells. Mechanical cell-loading systems will shortly be feasible that can inject up to 100,000 cells at a time. Before long, such technologies will seem primitive. Freewheeling genetic self-experimentation will be endemic as the DIY-Bio revolution unfolds. At present, crude and simple gene-editing can be accomplished only via laborious genetic engineering techniques. Sophisticated authoring tools don't exist. In future, computer-aided genetic and epigenetic enhancement can become an integral part of your personal growth plan.

Will Humanity's Successors also be Our Descendants?

To contrast "biological" with "artificial" conceptions of posthuman superintelligence is convenient. The distinction may also prove simplistic. In essence, whereas genetic change in biological humanity has always been slow, the software run on serial, programmable digital computers is executed exponentially faster (*cf.* Moore's Law); it's copyable without limit; it runs on multiple substrates; and it can be cheaply and rapidly edited, tested and debugged. Extrapolating, Singularitarians like Ray Kurzweil (1990) and Eliezer Yudkowsky (2008) prophesy that human programmers will soon become redundant because autonomous AI run on digital computers will undergo accelerating cycles of self-improvement. In this kind of scenario, artificial, greater-than-human nonbiological intelligence will be rapidly succeeded by artificial posthuman superintelligence.

So we may distinguish two radically different conceptions of posthuman superintelligence: on one hand, our supersentient, cybernetically enhanced, genetically rewritten *biological* descendants, on the other, *non*biological superintelligence, either a Kurzweilian ecosystem or the singleton Artificial General Intelligence (AGI) foretold by the Singularity Institute for Artificial Intelligence. Such a divide doesn't reflect a clean contrast between "natural" and "artificial" intelligence, the biological and the nonbiological. This contrast may prove another false dichotomy. Transhuman biology will increasingly become synthetic biology as genetic enhancement plus cyborgization proceeds apace. "Cyborgization" is a barbarous term to describe an invisible and potentially life-enriching symbiosis of biological sentience with artificial intelligence. Thus "narrow-spectrum" digital superintelligence on web-enabled chips can be more-or-less seamlessly integrated into our genetically enhanced bodies and brains. Seemingly limitless formal knowledge can be delivered on tap to supersentient organic wetware, i.e. us. *Critically, transhumans can exploit what is misleadingly known as* "*narrow*" *or* "*weak*" *AI to enhance our own code in a positive feedback loop of mutual enhancement*—first plugging in data and running multiple computer simulations, then tweaking and re-simulating once more. In short, biological humanity won't just be the spectator and passive consumer of the intelligence explosion, but its driving force. The smarter our AI, the greater our opportunities for reciprocal improvement. Multiple "hard" and "soft" take-off scenarios to posthuman superintelligence can be outlined for recursively self-improving organic robots, not just nonbiological AI (Good 1965). Thus for serious biohacking later this century, artificial quantum supercomputers (Deutsch 2011) may be deployed rather than today's classical toys to test-run multiple genetic interventions, accelerating the tempo of our recursive self-improvement. Quantum supercomputers exploit quantum coherence to do googols of computations all at once. So the accelerating growth of human/computer synergies means it's premature to suppose biological evolution will be superseded by technological evolution, let alone a "robot rebellion" as the parasite swallows its host (de Garis 2005; Yudkowsky 2008). As the human era comes to a close, the fate of biological (post)humanity is more likely to be symbiosis with AI followed by metamorphosis, not simple replacement.

Despite this witche's brew of new technologies, a conceptual gulf remains in the futurist community between those who imagine human destiny, if any, lies in digital computers running programs with (hypothetical) artificial consciousness; and in contrast radical bioconservatives who believe that our posthuman successors will also be our supersentient descendants at their organic neural networked core—not the digital zombies of symbolic AI (Haugeland 1985) run on classical serial computers or their souped-up multiprocessor cousins. For one metric of progress in AI remains stubbornly unchanged: despite the exponential growth of transistors on a microchip, the soaring clock speed of microprocessors, the growth in computing power measured in MIPS, the dramatically falling costs of manufacturing transistors and the plunging price of dynamic RAM etc., any chart plotting the growth rate in digital sentience shows neither exponential growth, nor linear growth, but no progress whatsoever. As far as we can tell, digital computers are still zombies. Our machines are becoming autistically intelligent, but not supersentient—nor even conscious. On some fairly modest philosophical assumptions, digital computers were not subjects of experience in 1946 (*cf.* ENIAC); nor are they conscious subjects in 2012 (*cf.* "Watson") (Baker 2011); nor do researchers know how any kind of sentience may be "programmed" in future. So what if anything does consciousness *do*? Is it computationally redundant? Pre-reflectively, we tend to have a "dimmer-switch" model of sentience: "primitive" animals have minimal awareness and "advanced" animals like human beings experience a proportionately more intense awareness. By analogy, most AI researchers assume that at a given threshold of complexity/intelligence/processing speed, consciousness will somehow "switch on", turn reflexive, and intensify too. The problem with the dimmer-switch model is that our most intense experiences, notably raw agony or blind panic, are also the most phylogenetically ancient, whereas the most "advanced" modes (e.g. linguistic thought and the rich generative syntax that has helped one species to conquer the globe) are phenomenologically so thin as to be barely accessible to introspection. Something is seriously amiss with our entire conceptual framework.

So the structure of the remainder of this essay is as follows. I shall first discuss the risks and opportunities of building friendly biological superintelligence. Next I discuss the nature of *full-spectrum* superintelligence—and why consciousness is computationally fundamental to the past, present and future success of organic robots. Why couldn't recursively self-improving *zombies* modify their own genetic source code and bootstrap their way to full-spectrum superintelligence, i.e. a zombie intelligence explosion? Finally, and most speculatively, I shall discuss the future of sentience in the cosmos.

Can We Build Friendly Biological Superintelligence?

Risk–Benefit Analysis

Crudely speaking, evolution "designed" male human primates to be hunters/warriors. Evolution "designed" women to be attracted to powerful, competitive alpha males. Until humans rewrite our own hunter-gatherer source code, we shall continue to practise extreme violence (Peterson and Wrangham 1997) against members of other species—and frequently against members of our own. A heritable (and conditionally activated) predisposition to unfriendliness shown towards members of other races and other species is currently hardwired even in "social" primates. Indeed we have a (conditionally activated) predisposition to compete against, and harm, anyone who isn't a genetically identical twin. Compared to the obligate siblicide found in some bird species, human sibling rivalry isn't normally so overtly brutal. But conflict as well as self-interested cooperation is endemic to Darwinian life on Earth. This grim observation isn't an argument for genetic determinism, or against gene-culture co-evolution, or to discount the decline of everyday violence with the spread of liberal humanitarianism—just a reminder of the omnipresence of immense risks so long as we're shot through with legacy malware. Attempting to conserve the genetic status quo in an era of weapons of mass destruction poses unprecedented global catastrophic and existential risks (Bostrom 2002). Indeed the single biggest underlying threat to the future of sentient life within our cosmological horizon derives, not from asocial symbolic AI software in the basement turning rogue and going FOOM (a runaway computational explosion of recursive self-improvement), but from conserving human nature in its present guise. In the twentieth century, male humans killed over 100 million fellow human beings and billions of non-human animals (Singer 1995). This century's toll may well be higher. Mankind currently spends well over a trillion dollars each year on weapons designed to kill and maim other humans. The historical record suggests such weaponry won't all be beaten into ploughshares.

Strictly speaking, however, humanity is more likely to be wiped out by *idealists* than by misanthropes, death-cults or psychologically unstable dictators. Anti-natalist philosopher David Benatar's plea ("Better Never to Have Been") for human extinction via voluntary childlessness (Benatar 2006) must fail if only by reason of selection pressure; but not everyone who shares Benatar's bleak diagnosis of life on Earth will be so supine. Unless we modify human nature, compassionate-minded negative utilitarians with competence in bioweaponry, nanorobotics or artificial intelligence, for example, may quite conceivably take direct action. Echoing Moore's law, Eliezer Yudkowsky warns that "Every eighteen months, the minimum IQ necessary to destroy the world drops by one point". Although suffering and existential risk might seem separate issues, they are intimately connected. Not everyone loves life so much they wish to preserve it. Indeed the extinction of Darwinian life is what many transhumanists are aiming for—just not framed in such apocalyptic and provocative language. For just as we educate small children so they can mature into fully-fledged adults, biological humanity may aspire to grow up, too, with the consequence that—in common with small children—archaic humans become extinct.

Technologies of Biofriendliness Empathogens?

How do you disarm a potentially hostile organic robot—despite your almost limitless ignorance of his source code? Provide him with a good education, civics lessons and complicated rule-governed ethics courses? Or give him a tablet of MDMA ("Ecstasy") and get smothered with hugs?

MDMA is short-acting (Holland 2001). The "penicillin of the soul" is potentially neurotoxic to serotonergic neurons. In theory, however, lifelong use of safe and sustainable empathogens would be a passport to worldwide biofriendliness. MDMA releases a potent cocktail of oxytocin, serotonin and dopamine into the user's synapses, thereby inducing a sense of "I love the world and the world loves me". There's no technical reason why MDMA's acute pharmacodynamic effects can't be replicated indefinitely shorn of its neurotoxicity. Designer "hug drugs" can potentially turn manly men into intelligent bonobo, more akin to the "hippie chimp" *Pan paniscus* than his less peaceable cousin *Pan troglodytes*. Violence would become unthinkable. Yet is this sort of proposal politically credible? "Morality pills" and other pharmacological solutions to human unfriendliness are both personally unsatisfactory and sociologically implausible. Do we really want to drug each other up from early childhood? Moreover, life would be immeasurably safer if our fellow humans weren't genetically predisposed to unfriendly behaviour in the first instance.

But how can this friendly predisposition are guaranteed? Friendliness can't realistically be hand-coded by tweaking the connections and weight strengths of our neural networks. Nor can robust friendliness in advanced biological intelligence be captured by a bunch of explicit logical rules and smart algorithms, as in the paradigm of symbolic AI.

Mass Oxytocination?

Amplified "trust hormone" (Lee et al. 2009) might create the biological underpinnings of world-wide peace and love if negative feedback control of oxytocin release can be circumvented. Oxytocin is functionally antagonised by testosterone in the male brain. Yet oxytocin enhancers have pitfalls too. Enriched oxytocin function leaves one vulnerable to exploitation by the unenhanced. Can we really envisage a cross-cultural global consensus for mass-medication? When? Optional or mandatory? And what might be the wider ramifications of a "high oxytocin, low testosterone" civilisation? Less male propensity to violent territorial aggression, for sure; but disproportionate intellectual progress in physics, mathematics and computer science to date has been driven by the hyper-systematising cognitive style of "extreme male" brains (Baron-Cohen 2001). Also, enriched oxytocin function can indirectly even promote *un*friendliness to "out-groups" in consequence of promoting in-group bonding. So as well as oxytocin enrichment, global security demands a more inclusive, impartial, intellectually sophisticated conception of "us" that embraces all sentient beings (Singer 1981)—the expression of a hyper-developed capacity for empathetic understanding combined with a hyper-developed capacity for rational systematisation. Hence the imperative need for Full-Spectrum Superintelligence.

Mirror-Touch Synaesthesia?

A truly long-term solution to unfriendly biological intelligence might be collectively to engineer ourselves with the functional generalisation of "mirror-touch" synaesthesia (Banissy 2009). On seeing you cut and hurt yourself, a mirror-touch synaesthete is liable to feel a stab of pain as acutely as you do. Conversely, your expressions of pleasure elicit a no less joyful response. Thus mirror-touch synaesthesia is a hyper-empathising condition that makes deliberate unfriendliness, in effect, biologically impossible in virtue of cognitively enriching our capacity to represent each other's first-person perspectives. The existence of mirror-touch synaesthesia is a tantalising hint at the God-like representational capacities of a Full-Spectrum Superintelligence. This so-called "disorder" is uncommon in humans.

Timescales

The biggest problem with all these proposals, and other theoretical biological solutions to human unfriendliness, is timescale. Billions of human and non-human animals will have been killed and abused before they could ever come to pass. Cataclysmic wars may be fought in the meantime with nuclear, biological and chemical weapons harnessed to "narrow" AI. Our circle of empathy expands only slowly and fitfully. For the most part, religious believers and traditional-minded bioconservatives won't seek biological enhancement/remediation for themselves or their children. So messy democratic efforts at "political" compromise are probably unavoidable for centuries to come. For sure, idealists can dream up utopian schemes to mitigate the risk of violent conflict until the "better angels of our nature" (Pinker 2011) can triumph, e.g. the election of a risk-averse all-female political class (Pellissier 2011) to replace legacy warrior males. Such schemes tend to founder on the rock of sociological plausibility. Innumerable sentient beings are bound to suffer and die in consequence.

Does Full-Spectrum Superintelligence Entail Benevolence?

The God-like perspective-taking faculty of a Full-Spectrum Superintelligence doesn't entail distinctively *human*-friendliness (Yudkowsky 2008) any more than a God-like Superintelligence could promote distinctively Aryan-friendliness. Indeed it's unclear how benevolent superintelligence could want omnivorous killer apes in our current guise to walk the Earth in any shape or form. But is there any connection at all between benevolence and intelligence? Pre-reflectively, benevolence and intelligence are orthogonal concepts. There's nothing obviously incoherent about a malevolent God or a malevolent—or at least a callously indifferent—Superintelligence. Thus a sceptic might argue that there is no link whatsoever between benevolence—on the face of it a mere personality variable—and enhanced intellect. After all, some sociopaths score highly on our [autistic, mind-blind] IQ tests. Sociopaths know that their victims suffer. They just don't care.

However, what's critical in evaluating cognitive ability is a *criterion of representational adequacy*. Representation is not an all-or-nothing phenomenon; it varies in functional degree. More specifically here, the cognitive capacity to represent the *formal* properties of mind differs from the cognitive capacity to represent the *subjective* properties of mind (Seager 2006). Thus a notional zombie Hyper-Autist robot running a symbolic AI program on an ultrapowerful digital computer with a classical von Neumann architecture may be beneficent or maleficent in its behaviour toward sentient beings. By its very nature, it can't know or care. Most starkly, the zombie Hyper-Autist might be programmed to convert the world's matter and energy into either heavenly "utilitronium" or diabolical "dolorium" without the slightest insight into the significance of what it was doing. This kind of scenario is at least a notional risk of creating insentient Hyper-Autists endowed with mere formal utility functions rather than hyper-sentient Full-Spectrum Superintelligence. By contrast, Full-Spectrum Superintelligence *does* care in virtue of its full-spectrum representational capacities—a bias-free generalisation of the superior perspective-taking, "mind-reading" capabilities that enabled humans to become the cognitively dominant species on the planet. Full-spectrum Superintelligence, if equipped with the posthuman cognitive generalisation of mirror-touch synaesthesia, understands your thoughts, your feelings and your egocentric perspective better than you do yourself.

Could there arise "evil" mirror-touch synaesthetes? In one sense, no. You can't go around wantonly hurting other sentient beings if you feel their pain as your own. Full-spectrum intelligence is friendly intelligence. But in another sense yes, insofar as primitive mirror-touch synaesthetes are preys to species-specific cognitive limitations that prevent them acting rationally to maximise the well-being of all sentience. Full-spectrum superintelligences would lack those computational limitations in virtue of their full cognitive competence in understanding both the subjective and the formal properties of mind. Perhaps full-spectrum superintelligences might optimise your matter and energy into a blissful smart angel; but they couldn't wantonly hurt you, whether by neglect or design.

More practically today, a cognitively superior analogue of natural mirror-touch synaesthesia should soon be feasible with reciprocal neuroscanning technology—a kind of naturalised telepathy. At first blush, mutual telepathic understanding sounds a panacea for ignorance and egotism alike. An exponential growth of shared telepathic understanding might safeguard against global catastrophe born of mutual incomprehension and WMD. As the poet Henry Wadsworth Longfellow observed, “If we could read the secret history of our enemies, we should find in each life sorrow and suffering enough to disarm all hostility”. Maybe so. The problem here, as advocates of Radical Honesty soon discover, is that many Darwinian thoughts scarcely promote friendliness if shared: they are often ill-natured, unedifying and unsuitable for public consumption. Thus, unless perpetually “loved-up” on MDMA or its long-acting equivalents, most of us would find mutual mind-reading a traumatic ordeal. Human society and most personal relationships would collapse in acrimony rather than blossom. Either way, our human incapacity fully to understand the first-person point of view of other sentient beings isn’t just a moral failing or a personality variable; it’s an *epistemic* limitation, an intellectual failure to grasp an objective feature of the natural world. Even “normal” people share with sociopaths this fitness-enhancing cognitive deficit. By posthuman criteria, perhaps we’re all ignorant quasi-sociopaths. The egocentric delusion (i.e. that the world centres on one’s existence) is genetically adaptive and strongly selected for over hundreds of millions of years (Dawkins 1976). Fortunately, it’s a cognitive failing amenable to technical fixes and eventually a cure: Full-Spectrum Superintelligence. The devil is in the details, or rather the genetic source code.

A Biotechnological Singularity?

Yet does this positive feedback loop of reciprocal enhancement amount to a Singularity (Vinge 1993) in anything more than a metaphorical sense? The risk of talking portentously about “The Singularity” isn’t of being wrong: it’s of being “not even wrong”—of reifying one’s ignorance and elevating it to the status of an ill-defined apocalyptic event. Already multiple senses of “The Singularity” proliferate in popular culture. Does taking LSD induce a Consciousness Singularity? How about the abrupt and momentous discontinuity in one’s conception of reality entailed by waking from a dream? Or the birth of language? Or the Industrial Revolution? So is the idea of recursive self-improvement leading to a *Bio*technological Singularity, or “Biosingularity” for short, any more rigorously defined than recursive self-improvement (Omohundro 2007) of seed AI leading to a “Technological Singularity”?

Metaphorically, perhaps, the impending biointelligence explosion represents an intellectual “event horizon” beyond which archaic humans cannot model or understand the future. Events beyond the Biosingularity will be stranger than science-fiction: too weird for unenhanced human minds—or the algorithms of a zombie super-Asperger—to predict or understand. In the popular sense of "event horizon", maybe the term is apt too, though the metaphor is still potentially misleading. Thus, theoretical physics tells us that one could pass through the event horizon of a non-rotating supermassive black hole and not notice any subjective change in consciousness—even though one's signals would now be inaccessible to an external observer. The Biosingularity will feel different in ways a human conceptual scheme can't express. But what is the empirical content of this claim?

What is Full-Spectrum Superintelligence?

> [g is] ostensibly some innate scalar brain force…[However] ability is a folk concept and not amenable to scientific analysis.
>
> Jon Marks (Dept Anthropology, Yale University), 1995, *Nature*, 9 xi, 143–144.

> Our normal waking consciousness, rational consciousness as we call it, is but one special type of consciousness, whilst all about it, parted from it by the filmiest of screens, there lie potential forms of consciousness entirely different.
>
> (William James)

Intelligence

"Intelligence" is a folk concept. The phenomenon is not well-defined—or rather any attempt to do so amounts to a stipulative definition that doesn't "carve Nature at the joints". The Cattell-Horn-Carroll psychometric theory of human cognitive abilities (1993) is probably most popular in academia and the IQ testing community. But the Howard Gardner multiple intelligences model, for example, differentiates "intelligence" into various spatial, linguistic, bodily-kinaesthetic, musical, interpersonal, intrapersonal, naturalistic and existential intelligence (Gardner 1983) rather than a single general ability ("g"). Who's right? As it stands, "g" is just a statistical artefact of our culture-bound IQ tests. If general intelligence were indeed akin to an innate scalar brain force, as some advocates of "g" believe, or if intelligence can best be modelled by the paradigm of symbolic AI, then the exponential growth of digital computer processing power might indeed entail an exponential growth in intelligence too—perhaps leading to some kind of Super-Watson (Baker 2011). Other facets of intelligence, however, resist enhancement by mere acceleration of raw processing power.

One constraint is that a theory of general intelligence should be race-, species-, and culture-neutral. Likewise, an impartial conception of intelligence should embrace all possible state-spaces of consciousness: prehuman, human, transhuman and posthuman.

The non-exhaustive set of criteria below doesn't pretend to be anything other than provisional. They are amplified in the sections to follow.

Full-Spectrum Superintelligence entails:

1. The capacity to solve the Binding Problem, (Revonsuo and Newman 1999) i.e. to generate phenomenally unified entities from widely distributed computational processes; and run cross-modally matched, data-driven world-simulations (Revonsuo 2005) of the mind-independent environment. (*cf.* naive realist theories of "perception" versus the world-simulation or "Matrix" paradigm. Compare disorders of binding, e.g. simultanagnosia (an inability to perceive the visual field as a whole), cerebral akinetopsia ("motion blindness"), etc. In the absence of a data-driven, almost real-time simulation of the environment, intelligent agency is impossible.)
2. A self or some non-arbitrary functional equivalent of a person to which intelligence can be ascribed. (*cf.* dissociative identity disorder ("multiple personality disorder"), or florid schizophrenia, or your personal computer: in the absence of at least a fleetingly unitary self, what philosophers call "synchronic identity", there is no entity that is intelligent, just an aggregate of discrete algorithms and an operating system.)
3. A "mind-reading" or perspective-taking faculty; higher-order intentionality (e.g. "he believes that she hopes that they fear that he wants…" etc.): social intelligence.
   The intellectual success of the most cognitively successful species on the planet rests, not just on the recursive syntax of human language, but also on our unsurpassed "mind-reading" prowess, an ability to simulate the perspective of other unitary minds: the "Machiavellian Ape" hypothesis (Byrne and Whiten 1988). Any ecologically valid intelligence test designed for a species of social animal must incorporate social cognition and the capacity for co-operative problem-solving. So must any test of empathetic superintelligence.
4. A metric to distinguish the important from the trivial.
   (Our theory of significance should be explicit rather than implicit, as in contemporary IQ tests. What distinguishes, say, mere calendrical prodigies and other "savant syndromes" from, say, a Grigori Perelman who proved the Poincaré conjecture? Intelligence entails understanding what does—and doesn't—matter. What matters is of course hugely contentious.)
5. A capacity to navigate, reason logically about, and solve problems in multiple state-spaces of consciousness [e.g. dreaming states (*cf.* lucid dreaming), waking consciousness, echolocation competence, visual discrimination, synaesthesia in all its existing and potential guises, humour, introspection, the different realms of psychedelia (*cf.* salvia space, "the K-hole" etc.)] including *realms of experience not yet co-opted by either natural selection or posthuman design for tracking features of the mind-independent world*. Full-Spectrum Superintelligence will

entail cross-domain goal-optimising ability in all possible state-spaces of consciousness (Shulgin 2011). and finally

6. "Autistic", pattern-matching, rule-following, mathematico-linguistic intelligence, i.e. the standard, mind-blind (Baron-Cohen 1995) cognitive tool-kit scored by existing IQ tests. High-functioning "autistic" intelligence is indispensable to higher mathematics, computer science and the natural sciences. High-functioning autistic intelligence is necessary—but not sufficient—for a civilisation capable of advanced technology that can cure ageing and disease, systematically phase out the biology of suffering, and take us to the stars. And for programming artificial intelligence.

We may then ask which facets of Full-Spectrum Superintelligence will be accelerated by the exponential growth of digital computer processing power? Number six, clearly, as decades of post-ENIAC progress in computer science attest. But what about numbers one-to-five? Here the picture is murkier.

The Bedrock of Intelligence. World-Simulation ("Perception")

Consider criterion number one, world-simulating prowess, or what we misleadingly term "perception". The philosopher Bertrand Russell (1948) once aptly remarked that one never sees anything but the inside of one's own head. In contrast to such inferential realism, commonsense perceptual direct realism offers all the advantages of theft over honest toil—and it's computationally useless for the purposes either of building artificial general intelligence or understanding its biological counterparts. For the bedrock of intelligent agency is the capacity of an embodied agent computationally to simulate dynamic objects, properties and events in the mind-independent environment. [For a contrary view, see e.g. Brooks 1991] The evolutionary success of organic robots over the past *c*. 540 million years has been driven by our capacity to run data-driven egocentric world-simulations—what the naive realist, innocent of modern neuroscience or post-Everett (Everett 1955) quantum mechanics, calls simply perceiving one's physical surroundings. Unlike classical digital computers, organic neurocomputers can simultaneously "bind" multiple features (edges, colours, motion, etc.) distributively processed across the brain into unitary phenomenal objects embedded in unitary spatio-temporal world-simulations apprehended by a momentarily unitary self: what Kant (1781) calls "the transcendental unity of apperception". These simulations run in (almost) real-time; the time-lag in our world-simulations is barely more than a few dozen milliseconds. Such blistering speed of construction and execution is adaptive and often life-saving in a fast-changing external environment. Recapitulating evolutionary history, pre-linguistic human infants must first train up their neural networks to bind the multiple features of dynamic objects and run unitary world-simulations before they can socially learn *second-order* representation and then *third-order* representation, i.e. language followed later in childhood by meta-language.

Occasionally, object binding and/or the unity of consciousness partially breaks down in mature adults who suffer a neurological accident. The results can be cognitively devastating (*cf*. akinetopsia or "motion blindness" (Zeki 1991); and simultanagnosia, an inability to apprehend more than a single object at a time Riddoch and Humphreys 2004), etc.). Yet normally our simulations of fitness-relevant patterns in the mind-independent local environment feel seamless. Our simulations each appear simply as "the world"; we just don't notice or explicitly represent the gaps. Neurons, (mis)construed as classical processors, are pitifully slow, with spiking frequencies barely up to 200 per second. By contrast, silicon (etc.) processors are ostensibly millions of times faster. Yet the notion that non-biological computers are faster than sentient neurocomputers is a philosophical assumption, not an empirical discovery. Here the assumption will be challenged. Unlike the CPUs of classical robots, an organic mind/brain delivers dynamic unitary phenomenal objects and unitary world-simulations with a "refresh rate" of many billions per second (*cf*. the persistence of vision as experienced watching a movie run at a mere 30 frames per second). These cross-modally matched simulations take the guise of what passes as the macroscopic world: a spectacular egocentric simulation run by the vertebrate CNS that taps into the world's fundamental quantum substrate (Ball 2011). A strong prediction of this conjecture is that classical digital computers will never be non-trivially conscious—or support software smart enough to understand their ignorance.

We should pause here. This is *not* a mainstream view. Most AI researchers regard stories of a non-classical mechanism underlying the phenomenal unity of biological minds as idiosyncratic at best. In fact no scientific consensus exists on the molecular underpinnings of the unity of consciousness, or on how such unity is even physically possible. By analogy, 1.3 billion skull-bound Chinese minds can never be a single subject of experience, irrespective of their interconnections. How waking or dreaming communities of membrane-bound classical neurons could—even *micro*conscious classical neurons—be any different? If materialism is true, conscious mind should be *impossible*. Yet any explanation of phenomenal object binding, the unity of perception, or the phenomenal unity of the self that invokes quantum coherence as here is controversial. One reason it's controversial is that the delocalisation involved in quantum coherence is exceedingly short-lived in an environment as warm and noisy as a macroscopic brain—supposedly too short-lived to do computationally useful work (Hagen 2002). Physicist Max Tegmark (2000) estimates that thermally-induced decoherence destroys any macroscopic coherence of brain states within 10^{-13} s, an unimaginably long time in natural Planck units but an unimaginably short time by everyday human intuitions. Perhaps it would be wiser just to acknowledge these phenomena are unexplained mysteries within a conventional materialist framework—as mysterious as the existence of consciousness itself. But if we're speculating about the imminent end of the human era (Good 1965), shoving the mystery under the rug isn't really an option. For the different strands (Yudkowsky 2007) of the Singularity movement share a common presupposition. This presupposition is that our complete ignorance within a materialist conceptual scheme of why consciousness exists (the

"Hard Problem") (Chalmers 1995), and of even the ghost of a solution to the Binding Problem, doesn't matter for the purposes of building the seed of artificial posthuman superintelligence. Our ignorance supposedly doesn't matter either because consciousness and/or our quantum "substrate" are computationally irrelevant to cognition (Hutter 2012) and the creation of nonbiological minds, or alternatively because the feasibility of "whole brain emulation" (Markram 2006) will allow us to finesse our ignorance.

Unfortunately, we have no grounds for believing this assumption is true or that the properties of our quantum "substrate" are functionally irrelevant to Full-Spectrum Superintelligence or its humble biological predecessors. Conscious minds are not substrate-neutral digital computers. Humans investigate and reason about problems of which digital computers are invincibly ignorant, not least the properties of consciousness itself. The Hard Problem of consciousness can't be quarantined from the rest of science and treated as a troublesome but self-contained anomaly: its mystery infects *everything* (Rescher 1974) that we think we know about ourselves, our computers and the world. Either way, the conjecture that the phenomenal unity of perception is a manifestation of ultra-rapid sequences of irreducible quantum coherent states *isn't* a claim that the mind/brain is capable of detecting events in the mind-independent world on this kind of sub-picosecond timescale. Rather the role of the local environment in shaping action-guiding experience in the awake mind/brain is here conjectured to be quantum state-*selection*. When we're awake, patterns of impulses from e.g. the optic nerve *select* which quantum-coherent frames are generated by the mind/brain—in contrast to the autonomous world-simulations spontaneously generated by the dreaming brain. Other quantum mind theorists, most notably Roger Penrose (1994) and Stuart Hameroff (2006), treat quantum minds as evolutionarily novel rather than phylogenetically ancient. They invoke a non-physical (Saunders 2010) wave-function collapse and unwisely focus on e.g. the ability of mathematically-inclined brains to perform non-computable functions in higher mathematics, a feat for which selection pressure has presumably been non-existent (Litt 2006). Yet the human capacity for sequential linguistic thought and formal logico-mathematical reasoning is a late evolutionary novelty executed by a slow, brittle virtual machine running on top of its massively parallel quantum parent—a momentous evolutionary innovation whose neural mechanism is still unknown.

In contrast to the evolutionary novelty of serial linguistic thought, our ancient and immensely adaptive capacity to run unitary world-simulations, simultaneously populated by hundreds or more dynamic unitary objects, enables organic robots to solve the computational challenges of navigating a hostile environment that would leave the fastest classical supercomputer grinding away until Doomsday. Physical theory (*cf*. the Bekenstein bound) shows that informational resources as classically conceived are not just physical but finite and scarce: a maximum possible limit of 10^{120} bits set by the surface area of the entire accessible universe (Lloyd 2002) expressed in Planck units according to the Holographic principle. An infinite computing device like a universal Turing machine (Dyson 2012) is physically impossible. So invoking computational equivalence and asking whether a classical

Turing machine can run a human-equivalent macroscopic world-simulation is akin to asking whether a classical Turing machine can factor 1,500 digit numbers in real-world time [i.e. no]. No doubt resourceful human and transhuman programmers will exploit all manner of kludges, smart workarounds and "brute-force" algorithms to try and defeat the Binding Problem in AI. How will they fare? Compare clod-hopping AlphaDog with the sophisticated functionality of the sesame-seed sized brain of a bumblebee. Brute-force algorithms suffer from an exponentially growing search space that soon defeats any classical computational device in open-field contexts. As witnessed by our seemingly effortless world-simulations, organic minds are ultrafast; classical computers are slow. Serial *thinking* is slower still; but that's not what conscious biological minds are good at. On this conjecture, "substrate-independent" phenomenal world-simulations are impossible for the same reason that "substrate-independent" chemical valence structure is impossible. We're simply begging the question of what's functionally (ir) relevant. Ultimately, Reality has only a single, "program-resistant" (Gunderson 1985) ontological level even though it's amenable to description at different levels of computational abstraction; and the nature of this program-resistant level as disclosed by the subjective properties of one's mind (Lockwood 1989) is utterly at variance with what naive materialist metaphysics would suppose (Seager 2006). *If* our phenomenal world-simulating prowess turns out to be constitutionally tied to our quantum mechanical wetware, then substrate-neutral virtual machines (i.e. software implementations of a digital computer that execute programs like a physical machine) will never be able to support "virtual" qualia or "virtual" unitary subjects of experience. This rules out sentient life "uploading" itself to digital nirvana (Moravec 1990). C*ontra* Marvin Minsky ("The most difficult human skills to reverse engineer are those that are unconscious") (Minsky 1987), the most difficult skills for roboticists to engineer in artificial robots are actually intensely conscious: our colourful, noisy, tactile, sometimes hugely refractory virtual worlds.

Naively, for sure, real-time world-simulation doesn't sound too difficult. Hollywood robots do it all the time. Videogames become ever more photorealistic. Perhaps one imagines viewing some kind of inner TV screen, as in a Terminator movie or The Matrix. Yet the capacity of an awake or dreaming brain to generate unitary macroscopic world-simulations can only superficially resemble a little man (a "homunculus") viewing its own private theatre—on pain of an infinite regress. For by what mechanism would the homunculus view this inner screen? Emulating the behaviour of even the very simplest sentient organic robots on a classical digital computer is a daunting task. *If* conscious biological minds are irreducibly quantum mechanical by their very nature, then reverse-engineering the brain to create digital human "mindfiles" and "roboclones" alike will prove impossible.

The Bedrock of Superintelligence Hypersocial Cognition ("Mind-reading")

Will superintelligence be solipsistic or social? Overcoming a second obstacle to delivering human-level artificial general intelligence—let alone building a recursively self-improving super-AGI culminating in a Technological Singularity—depends on finding a solution to the first challenge, i.e. real-time world-simulation. For the evolution of distinctively human intelligence, sitting on top of our evolutionarily ancient world-simulating prowess, has been driven by the interplay between our rich generative syntax and superior "mind-reading" skills: so-called Machiavellian intelligence (Byrne and Whiten 1988). Machiavellian intelligence is an egocentric parody of God's-eye-view empathetic superintelligence. Critically for the prospects of building AGI, this real-time mind-modelling expertise is parasitic on the neural wetware to generate unitary first-order world-simulations—virtual worlds populated by the avatars of intentional agents whose different first-person perspectives can be partially and imperfectly understood by their simulator. Even articulate human subjects with autism spectrum disorder are prone to multiple language deficits because they struggle to understand the intentions—and higher-order intentionality—of neurotypical language users. Indeed natural language is itself a pre-eminently social phenomenon: its criteria of application must first be socially learned. Not all humans possess the cognitive capacity to acquire mind-reading skills and the cooperative problem-solving expertise that sets us apart from other social primates. Most notably, people with autism spectrum disorder don't just fail to understand other minds; autistic intelligence cannot begin to understand its own mind. Pure autistic intelligence has no conception of a self that can be improved, recursively or otherwise. Autists can't "read" their own minds. The inability of the autistic mind to take what Daniel Dennett (1987) calls the "intentional stance" parallels the inability of classical computers to understand the minds of intentional agents—or have insight into their own zombie status. Even with smart algorithms and ultra-powerful hardware, the ability of ultra-intelligent autists to predict the long-term behaviour of mindful organic robots by relying exclusively on the physical stance (i.e. solving the Schrödinger equation of the intentional agent in question) will be extremely limited. For example, much collective human behaviour is chaotic in the technical sense, i.e. it shows extreme sensitivity to initial conditions that confounds long-term prediction by even the most powerful real-world supercomputer. But there's a worse problem: reflexivity. Predicting sociological phenomena differs essentially from predicting mindless physical phenomena. Even in a classical, causally deterministic universe, the behaviour of mindful, reflexively self-conscious agents is frequently unpredictable, even in principle, from *within* the world owing to so-called prediction paradoxes (Welty 1970). When the very act of prediction causally interacts with the predicted event, then self-defeating or self-falsifying predictions are inevitable. Self-falsifying predictions are a mirror image of so-called self-fulfilling predictions. So in common with autistic "idiot savants", classical AI gone rogue will be vulnerable to the low cunning of Machiavellian apes and the high cunning of our transhuman descendants.

This argument (i.e. our capacity for unitary mind-simulation embedded in unitary world-simulation) for the cognitive primacy of biological general intelligence isn't decisive. For a start, computer-aided Machiavellian humans can program robots with "narrow" AI—or perhaps "train up" the connections and weights of a subsymbolic connectionist architecture (Rumelhart et al. 1986)—for their own manipulative purposes. Humans underestimate the risks of zombie infestation at our peril. Given our profound ignorance of how conscious mind is even possible, it's probably safest to be agnostic over whether autonomous non-biological robots will ever emulate human world-simulating or mind-reading capacity in most open-field contexts, despite the scepticism expressed here. Either way, the task of devising an ecologically valid measure of general intelligence that can reliably, predictively and economically discriminate between disparate life-forms is immensely challenging, not least because the intelligence test will express the value-judgements, and species- and culture-bound conceptual scheme, of the tester. Some biases are insidious and extraordinarily subtle: for example, the desire systematically to measure "intelligence" with mind-blind IQ tests is itself a quintessentially Asperger-ish trait. In consequence, social cognition is disregarded altogether. What we fancifully style "IQ tests" is designed by people with abnormally high AQs (Baron-Cohen 2001) as well as self-defined high IQs. Thus, many human conceptions of (super) intelligence resemble high-functioning autism spectrum disorder rather than a hyper-empathetic God-like Super-Mind. For example, an AI that attempted systematically to maximise the cosmic abundance of paperclips (Yudkowsky 2008) would be recognisably autistic rather than incomprehensibly alien. Full-Spectrum (Super-) intelligence is certainly harder to design or quantify scientifically than mathematical puzzle-solving ability or performance in verbal memory-tests: "IQ". But that's because superhuman intelligence will be not just quantitatively different but also qualitatively alien (Huxley 1954) from human intelligence. To misquote Robert McNamara, cognitive scientists need to stop making what is measurable important, and find ways to make the important measurable. An idealised Full-Spectrum Superintelligence will indeed be capable of an impartial "view from nowhere" or God's-eye-view of the multiverse (Wallace 2012), a mathematically complete Theory of Everything—as does modern theoretical physics, in aspiration if not achievement. But in virtue of its God's-eye-view, Full-Spectrum Superintelligence must also be hypersocial and super*sentient*: able to understand all possible first-person perspectives, the state-space of all possible minds in other Hubble volumes, other branches of the universal wavefunction—and in other solar systems and galaxies if such beings exist within our cosmological horizon. Idealised at least, Full-Spectrum Superintelligence will be able to understand and weigh the significance of all possible modes of experience *irrespective of whether they have hitherto been recruited for information-signalling purposes*. The latter is, I think, by far the biggest intellectual challenge we face as cognitive agents. The systematic investigation of alien types of consciousness intrinsic to varying patterns of matter and energy

(Lockwood 1989) calls for a methodological and ontological revolution (Shulgin 1995). Transhumanists talking of post-Singularity superintelligence are fond of hyperbole about "Level 5 Future Shock" etc.; but it's been aptly said that if Elvis Presley were to land in a flying saucer on the White House lawn, it's as nothing in strangeness compared to your first DMT trip.

Ignoring the Elephant: Consciousness Why Consciousness is Computationally Fundamental to the Past, Present and Future Success of Organic Robots

The pachyderm in the room in most discussions of (super) intelligence is consciousness—not just human reflective self-awareness but the whole gamut of experience from symphonies to sunsets, agony to ecstasy: the phenomenal world. All one ever knows, except by inference, and is the contents of one's own conscious mind: what philosophers call "qualia". Yet according to the ontology of our best story of the world, namely physical science, *conscious* minds shouldn't exist at all, i.e. we should be zombies, insentient patterns of matter and energy indistinguishable from normal human beings but lacking conscious experience. Dutch computer scientist Edsger Dijkstra famously once remarked, "The question of whether a computer can think is no more interesting than the question of whether a submarine can swim". Yet the question of whether a programmable digital computer—or a subsymbolic connectionist system with a merely classical parallelism (Churchland 1989)—could possess, *and think about*, qualia, "bound" perceptual objects, a phenomenal self, or the unitary phenomenal minds of sentient organic robots can't be dismissed so lightly. For if advanced nonbiological intelligence is to be smart enough comprehensively to understand, predict and manipulate the behaviour of enriched biological intelligence, then the AGI can't rely autistically on the "physical stance", i.e. to monitor the brains, scan the atoms and molecules, and then solve the Schrödinger equation of intentional agents like human beings. Such calculations would take longer than the age of the universe.

For sure, many forms of human action can be predicted, fallibly, on the basis of crude behavioural regularities and reinforcement learning. Within your world-simulation, you don't need a theory of mind or an understanding of quantum mechanics to predict that Fred will walk to the bus-stop again today. Likewise, powerful tools of statistical analysis run on digital supercomputers can predict, fallibly, many kinds of human collective behaviour, for example stock markets. Yet to surpass human and transhuman capacities in all significant fields, AGI must understand how intelligent biological robots can think about, talk about and manipulate the manifold varieties of consciousness that make up their virtual worlds. Some investigators of consciousness even dedicate their lives to that end; what might a notional insentient AGI suppose we're doing? There is no evidence that serial digital computers have the capacity to do anything of the kind—or could ever be programmed to do so. Digital computers don't know anything about conscious minds, unitary persons, the nature of phenomenal pleasure and pain, or the Problem of Other Minds; it's not even "all dark inside". The challenge for a *conscious* mind posed by understanding itself "from the inside" pales into insignificance compared to the challenge for a nonconscious system of understanding a conscious mind "from the outside". Nor within the constraints of a materialist ontology have we the slightest clue how the purely classical parallelism of a subsymbolic, "neurally inspired" connectionist architecture could turn water into wine and generate *unitary* subjects of experience to fill the gap. For even if we conjecture in the spirit of Strawsonian physicalism—the only scientifically literate form of panpsychism (Strawson 2006)—that the fundamental stuff of the world, the mysterious "fire in the equations", is fields of microqualia, this bold ontological conjecture doesn't, by itself, explain why biological robots aren't zombies. This is because structured aggregates of classically conceived "mind-dust" aren't the same as a unitary phenomenal subject of experience who apprehends "bound" spatio-temporal objects in a dynamic world-simulation. Without phenomenal object binding and the unity of perception, we are faced with the spectre of what philosophers call "mereological nihilism" (Merricks 2001). Mereological nihilism, also known as "compositional nihilism", is the position that composite objects with proper parts do not exist: strictly speaking, only basic building blocks without parts have more than fictional existence. Unlike the fleetingly unitary phenomenal minds of biological robots, a classical digital computer and the programs it runs lack ontological integrity: it's just an assemblage of algorithms. In other words, a classical digital computer has no self to understand or a mind recursively to improve, exponentially or otherwise. Talk about artificial "intelligence" exploding (Hutter 2012) is just an anthropomorphic projection on our part.

So how do biological brains solve the binding problem and become persons? (Parfit 1984) In short, we don't know. Vitalism is clearly a lost cause. Most AI researchers would probably dismiss—or at least discount as wildly speculative—any story of the kind mooted here involving macroscopic quantum coherence grounded in an ontology of physicalistic panpsychism. But in the absence of any story at all, we are left with a theoretical vacuum and a faith that natural science—or the exponential growth of digital computer processing power culminating in a Technological Singularity—will one day deliver an answer. Evolutionary biologist Theodosius Dobzhansky famously observed how "Nothing in Biology Makes Sense Except in the Light of Evolution". In the same vein, nothing in the future of intelligent life in the universe makes sense except in the light of a solution to the Hard Problem of Consciousness and the closure of Levine's Explanatory Gap (Levine 1983). Consciousness is the only reason anything matters at all; and it's the only reason why unitary subjects of experience can ask these questions; and yet materialist orthodoxy has no idea how or why the phenomenon exists. Unfortunately, the Hard Problem won't be solved by building more advanced digital zombies who can tell mystified conscious minds the answer.

More practically for now, perhaps the greatest cognitive challenge of the millennium and beyond is deciphering and systematically manipulating the "neural correlates of consciousness" (Koch 2004). Neuroscientists use this expression in default of any deeper explanation of our myriad qualia. How and why does experimentally stimulating via microelectrodes one cluster of nerve cells in the neocortex yield the experience of phenomenal colour; stimulating a superficially similar type of nerve cell induces a musical jingle; stimulating another with a slightly different gene expression profile a sense of everything being hysterically funny; stimulating another seemingly of your mother; and stimulating another of an archangel, say, in front of your body-image? In each case, the molecular variation in neuronal cell architecture is ostensibly *trivial*; the difference in subjective experience is profound. On a mind/brain identity theory, such experiential states are an intrinsic property of some configurations of matter and energy (Lockwood 1989). How and why this is so is incomprehensible on an orthodox materialist ontology. Yet empirically, microelectrodes, dreams and hallucinogenic drugs elicit these experiences regardless of any information-signalling role such experiences typically play in the "normal" awake mind/brain. Orthodox materialism and classical information-based ontologies alike do not merely lack any explanation for why consciousness and our countless varieties of qualia exist. They lack any story of how our qualia could have the causal efficacy to allow us to allude to—and in some cases volubly expatiate on—their existence. Thus, mapping the neural correlates of consciousness is not amenable to formal computational methods: digital zombies don't have any qualia, or at least any "bound" macroqualia, that could be mapped, or a unitary phenomenal self that could do the mapping.

Note this claim for the cognitive primacy of biological sentience isn't a denial of the Church-Turing thesis that given *infinite* time and *infinite* memory any Turing-universal system can formally simulate the behaviour of any conceivable process that can be digitized. Indeed (very) fancifully, if the multiverse were being run on a cosmic supercomputer, speeding up its notional execution a million times would presumably speed us up a million times too. But that's not the issue here. Rather the claim is that nonbiological AI run on real-world digital computers cannot tackle the truly hard and momentous cognitive challenge of investigating first-person states of egocentric virtual worlds—or understand why some first-person states, e.g. agony or bliss, are intrinsically important, and cause unitary subjects of experience, persons, to act the way we do.

At least in common usage, "intelligence" refers to an agent's ability to achieve goals in a wide range of environments. What we call greater-than-human intelligence or Superintelligence presumably involves the design of qualitatively *new* kinds of intelligence never seen before. Hence the growth of artificial intelligence and symbolic AI, together with subsymbolic (allegedly) brain-inspired connectionist architectures, and soon artificial quantum computers. But contrary to received wisdom in AI research, sentient biological robots are making greater cognitive progress in discovering the potential for truly novel kinds of intelligence than the techniques of formal AI. We are doing so by synthesising and empirically investigating a galaxy of psychoactive designer drugs (Shulgin 2011)—experimentally opening up the possibility of radically new kinds of intelligence in different state-spaces of consciousness. For the most cognitively challenging environments don't lie in the stars but in organic mind/brains—the baffling subjective properties of quantum-coherent states of matter and energy—most of which aren't explicitly represented in our existing conceptual scheme.

Case Study: Visual Intelligence Versus Echolocation Intelligence: What is it Like to be a Super-Intelligent Bat?

Let's consider the mental state-space of organisms whose virtual worlds are rooted in their dominant sense mode of echolocation (Nagel 1974). This example isn't mere science fiction. Unless post-Everett quantum mechanics (Deutsch 1997) is false, we're forced to assume that googols of quasi-classical branches of the universal wavefunction—the master formalism that exhaustively describes our multiverse—satisfy this condition. Indeed their imperceptible interference effects must be present even in "our" world: strictly speaking, interference effects from branches that have decohered ("split") never wholly disappear; they just become vanishingly small. Anyhow, let's assume these echolocation superminds have evolved opposable thumbs, a rich generative syntax and advanced science and technology. How are we to understand or measure this alien kind of (super) intelligence? Rigging ourselves up with artificial biosonar apparatus and transducing incoming data into the familiar textures of sight or sound might seem a good start. But to understand the conceptual world of echolocation superminds, we'd need to equip ourselves with neurons and neural networks neurophysiologically equivalent to smart chiropterans. If one subscribes to a coarse-grained functionalism about consciousness, then echolocation experience would (somehow) emerge at some abstract computational level of description. The implementation details, or "meatware" as biological mind/brains are derisively called, are supposedly incidental or irrelevant. The functionally unique valence properties of the carbon atom, and likewise the functionally unique quantum mechanical properties of liquid water (Vitiello 2001), are discounted or ignored. Thus, according to the coarse-grained functionalist, silicon chips could replace biological neurons without loss of function or subjective identity. By contrast, the *micro*-functionalist, often branded a mere "carbon chauvinist", reckons that the different intracellular properties of biological neurons—with their different gene expression profiles, diverse primary, secondary, tertiary, and quaternary amino acid chain folding (etc.) as described by quantum chemistry—are critical to the many and varied phenomenal properties such echolocation neurons express. Who is right? We'll only ever know the answer by rigorous self-experimentation: a post-Galilean science of mind.

It's true that humans don't worry much about our ignorance of echolocation experience, or our ignorance of echolocation primitive terms, or our ignorance of possible conceptual schemes expressing echolocation intelligence in echolocation world-simulations. This is because we don't highly esteem bats. Humans don't share the same interests or purposes as our flying cousins, e.g. to attract desirable, high-fitness bats and rear reproductively successful baby bats. Alien virtual worlds based on biosonar don't seem especially significant to *Homo sapiens* except as an armchair philosophical puzzle.

Yet this assumption would be intellectually complacent. Worse, understanding what it's like to be a hyperintelligent bat mind is *comparatively* easy. For echolocation experience has been recruited by natural selection to play an information-signalling role in a fellow species of mammal; and in principle a research community of language users could biologically engineer their bodies and minds to replicate bat-type experience and establish crude intersubjective agreement to discuss and conceptualise its nature. By contrast, the vast majority of experiential state-spaces remain untapped and unexplored. This task awaits Full-Spectrum Superintelligence in the posthuman era.

In a more familiar vein, consider visual intelligence. How does one measure the visual intelligence of a congenitally blind person? Even with sophisticated technology that generates "inverted spectrograms" of the world to translate visual images into sound, the congenitally blind are invincibly ignorant of visual experience and the significance of visually-derived concepts. Just as a sighted idiot has greater visual intelligence than a blind super-rationalist sage, likewise psychedelics confer the ability to become (for the most part) babbling idiots about other state-spaces of consciousness—but babbling idiots whose insight is deeper than the drug-naive or the genetically unenhanced—or the digital zombies spawned by symbolic AI and its connectionist cousins.

The challenge here is that the vast majority of these alien state-spaces of consciousness latent in organised matter haven't been recruited by natural selection for information-tracking purposes. So "psychonauts" don't yet have the conceptual equipment to navigate these alien state-spaces of consciousness in even a pseudo-public language, let alone integrate them in any kind of overarching conceptual framework. Note the claim here *isn't* that taking e.g. ketamine, LSD, salvia, DMT and a dizzying proliferation of custom-designed psychoactive drugs is the royal route to wisdom. Or that ingesting such agents will give insight into deep mystical truths. On the contrary: it's precisely because such realms of experience *haven't* previously been harnessed for information-processing purposes by evolution in "our" family of branches of the universal wavefunction that makes investigating their properties so cognitively challenging—currently beyond our conceptual resources to comprehend. After all, plants synthesise natural psychedelic compounds to scramble the minds of herbivores that might eat them, not to unlock mystic wisdom. Unfortunately, there is no "neutral" medium of thought impartially to appraise or perceptually cross-modally match all these other experiential state-spaces. One can't somehow stand outside one's own stream of consciousness to evaluate how the properties of the medium are infecting the notional propositional content of the language that one uses to describe it.

By way of illustration, compare drug-induced visual experience in a notional community of congenitally blind rationalists who lack the visual apparatus to transduce incident electromagnetic radiation of our familiar wavelengths. The lone mystical babbler who takes such a vision-inducing drug is convinced that [what we would call] visual experience is profoundly significant. And as visually intelligent folk, we know that he's right: visual experience is potentially hugely significant—to an extent which the blind mystical babbler can't possibly divine. But can the drug-taker convince his congenitally blind fellow tribesmen that his mystical visual experiences really matter in the absence of perceptual equipment that permits sensory discrimination? No, he just sounds psychotic. Or alternatively, he speaks lamely and vacuously of the "ineffable". The blind rationalists of his tribe are unimpressed.

The point of this fable is that we've scant reason to suppose that biologically re-engineered posthumans millennia hence will share the same state-spaces of consciousness, or the same primitive terms, or the same conceptual scheme, or the same type of virtual world that human beings now instantiate. Maybe all that will survive the human era is a descendant of our mathematical formalism of physics, M-theory of whatever, in basement reality.

Of course such ignorance of other state-spaces of experience doesn't normally trouble us. Just as the congenitally blind don't grow up in darkness—a popular misconception—the drug-naive and genetically unenhanced don't go around with a sense of what we're missing. We notice teeming abundance, not gaping voids. Contemporary humans can draw upon terms like "blindness" and "deafness" to characterise the deficits of their handicapped conspecifics. From the perspective of full-spectrum superintelligence, what we really need is *millions* more of such "privative" terms, as linguists call them, to label the different state-spaces of experience of which genetically unenhanced humans are ignorant. In truth, there may very well be more than millions of such nameless state-spaces, each as incommensurable as e.g. visual and auditory experience. We can't yet begin to quantify their number or construct any kind of crude taxonomy of their interrelationships.

Note the problem here isn't cognitive bias or a deficiency in logical reasoning. Rather a congenitally blind (etc.) super-rationalist is constitutionally ignorant of visual experience, visual primitive terms, or a visually-based conceptual scheme. So (s)he can't cite e.g. Aumann's agreement theorem [claiming in essence that two cognitive agents acting rationally and with common knowledge of each other's beliefs cannot agree to disagree] or be a good Bayesian rationalist or whatever: these are incommensurable state-spaces of experience as closed to human minds as Picasso is to an earthworm. Moreover, there is no reason to expect one realm, i.e. "ordinary waking consciousness", to be cognitively privileged relative to every other realm. "Ordinary waking consciousness" just happened to be genetically adaptive in the African savannah on Planet Earth. Just as humans are incorrigibly ignorant of minds grounded in echolocation—both echolocation world-simulations and echolocation conceptual schemes—likewise we are invincibly ignorant of posthuman life while trapped within our existing genetic architecture of intelligence.

In order to understand the world—both its formal/mathematical *and* its subjective properties—sentient organic life must bootstrap its way to super-sentient

Full-Spectrum Superintelligence. Grown-up minds need tools to navigate all possible state-spaces of qualia, including all possible first-person perspectives, and map them—initially via the neural correlates of consciousness in our world-simulations—onto the formalism of mathematical physics. Empirical evidence suggests that the behaviour of the stuff of the world is exhaustively described by the formalism of physics. To the best of our knowledge, physics is causally closed and complete, at least within the energy range of the Standard Model. In other words, there is nothing to be found in the world—no "element of reality", as Einstein puts it—that isn't captured by the equations of physics and their solutions. This is a powerful formal constraint on our theory of consciousness. Yet our ultimate theory of the world must also close Levine's notorious "Explanatory Gap". Thus, we must explain why consciousness exists at all ("The Hard Problem"); offer a rigorous derivation of our diverse textures of qualia from the field-theoretic formalism of physics; and explain how qualia combine ("The Binding Problem") in organic minds. These are powerful constraints on our ultimate theory too. How can they be reconciled with physicalism? Why aren't we zombies?

The hard-nosed sceptic will be unimpressed at such claims. How *significant* are these outlandish state-spaces of experience? And how are they computationally relevant to (super) intelligence? Sure, says the sceptic, reckless humans may take drugs, and experience wild, weird and wonderful states of mind. But so what? Such exotic states aren't objective in the sense of reliably tracking features of the mind-independent world. Elucidation of their properties doesn't pose a well-defined problem that a notional universal algorithmic intelligence (Legg and Hutter 2007) could solve.

Well, let's assume, provisionally at least, that all mental states are identical with physical states. If so, then all experience is an objective, spatio-temporally located feature of the world whose properties a unified natural science must explain. A cognitive agent can't be intelligent, let alone superintelligent, and yet be constitutionally ignorant of a fundamental feature of the world—not just ignorant, but completely incapable of gathering information about, exploring, or reasoning about its properties. Whatever else it may be, superintelligence can't be constitutionally *stupid*. What we need is a universal, species-neutral criterion of significance that can weed out the trivial from the important; and gauge the intelligence of different cognitive agents accordingly. Granted, such a criterion of significance might seem elusive to the antirealist about value (Mackie 1991). Value nihilism treats any ascription of (in) significance as arbitrary. Or rather the value nihilist maintains that what we find significant simply reflects what was fitness-enhancing for our forebears in the ancestral environment of adaptation (Barkow 1992). Yet for reasons we simply don't understand, Nature discloses just such a universal touchstone of importance, namely the pleasure-pain axis: the world's inbuilt metric of significance and (dis)value. We're not zombies. First-person facts exist. Some of them matter urgently, e.g. I am in pain. Indeed it's unclear if the expression "I'm in agony; but the agony doesn't matter" even makes cognitive sense. Built into the very nature of agony is the knowledge that its subjective raw awfulness matters a great deal—not instrumentally or derivatively, but by its very nature. If anyone—or indeed any notional super-AGI—supposes that your agony *doesn't* matter, then he/it hasn't adequately represented the first-person perspective in question.

So the existence of first-person facts is an objective feature of the world that any intelligent agent must comprehend. Digital computers and the symbolic AI code they execute can support formal utility functions. In some contexts, formally programmed utility functions can play a role functionally analogous to importance. But nothing intrinsically matters to a digital zombie. Without sentience, and more specifically without hedonic tone, nothing inherently matters. By contrast, extreme pain and extreme pleasure in any guise intrinsically matter intensely. Insofar as exotic state-states of experience are permeated with positive or negative hedonic tone, they matter too. In summary, "He jests at scars, that never felt a wound": scepticism about the self-intimating significance of this feature of the world is feasible only in its absence.

The Great Transition

The End of Suffering

A defining feature of general intelligence is the capacity to achieve one's goals in a wide range of environments. All sentient biological agents are endowed with a pleasure-pain axis. All prefer occupying one end to the other. A pleasure-pain axis confers inherent significance on our lives: the opioid-dopamine neurotransmitter system extends from flatworms to humans. Our core behavioural and physiological responses to noxious and rewarding stimuli have been strongly conserved in our evolutionary lineage over hundreds of millions of years. Some researchers (Cialdini 1987) argue for *psychological hedonism*, the theory that all choice in sentient beings is motivated by a desire for pleasure or an aversion from suffering. When we choose to help others, this is because of the pleasure that we ourselves derive, directly or indirectly, from doing so. Pascal put it starkly: "All men seek happiness. This is without exception. Whatever different means they employ, they all tend to this end. The cause of some going to war, and of others avoiding it, is the same desire in both, attended with different views. This is the motive of every action of every man, even of those who hang themselves". In practice, the hypothesis of psychological hedonism is plagued with anomalies, circularities and complications if understood as a universal principle of agency: the "pleasure principle" is simplistic as it stands. Yet the broad thrust of this almost embarrassingly commonplace idea may turn out to be central to understanding the future of life in the universe. If even a weak and exception-laden version of psychological hedonism is true, then there is an intimate link between full-spectrum superintelligence and happiness: the "attractor" to which rational sentience is heading. If that's really what we're striving for, a lot of the time at least, then instrumental means-ends rationality dictates that intelligent agency should seek maximally cost-effective ways to deliver happiness—and then superhappiness and beyond.

A discussion of psychological hedonism would take us too far afield here. More fruitful now is just to affirm a truism and then explore its ramifications for life in the post-genomic era. Happiness is typically one of our goals. Intelligence amplification entails pursuing our goals more rationally. For sure, happiness, or at least a reduction in unhappiness, is frequently sought under a variety of descriptions that don't explicitly allude to hedonic tone and sometimes disavow it altogether. Natural selection has "encephalised" our emotions in deceptive, fitness-enhancing ways within our world-simulations. Some of these adaptive fetishes may be formalised in terms of abstract utility functions that a rational agent would supposedly maximise. Yet even our loftiest intellectual pursuits are underpinned by the same neurophysiological reward and punishment pathways. The problem for sentient creatures is that, both personally and collectively, Darwinian life is not very smart or successful in its efforts to achieve long-lasting well-being. Hundreds of millions of years of "Nature, red in tooth and claw" attest to this terrible cognitive limitation. By a whole raft of indices (suicide rates, the prevalence of clinical depression and anxiety disorders, the Easterlin paradox, etc.) humans are not getting any (un) happier on average than our Palaeolithic ancestors despite huge technological progress. Our billions of factory-farmed victims (Francione 2006) spend most of their abject lives below hedonic zero. In absolute terms, the amount of suffering in the world increases each year in humans and non-humans alike. Not least, evolution sabotages human efforts to improve our subjective well-being thanks to our genetically constrained hedonic treadmill—the complicated web of negative feedback mechanisms in the brain that stymies our efforts to be durably happy at every turn (Brickman et al. 1978). Discontent, jealousy, anxiety, periodic low mood, and perpetual striving for "more" were fitness-enhancing in the ancient environment of evolutionary adaptedness. Lifelong bliss wasn't harder for information-bearing self-replicators to encode. Rather lifelong bliss was genetically maladaptive and hence selected against. Only now can biotechnology remedy organic life's innate design flaw.

A potential pitfall lurks here: the fallacy of composition. Just because all individuals tend to seek happiness and shun unhappiness doesn't mean that all individuals seek universal happiness. We're not all closet utilitarians. Genghis Khan wasn't trying to spread universal bliss. As Plato observed, "Pleasure is the greatest incentive to evil." But here's the critical point. Full-Spectrum Superintelligence entails the cognitive capacity impartially to grasp all possible first-person perspectives—overcoming egocentric, anthropocentric, and ethnocentric bias (*cf.* mirror-touch synaesthesia). As an idealisation, at least, Full-Spectrum Superintelligence understands and weighs the full range of first-person facts. First-person facts are as much an objective feature of the natural world as the rest mass of the electron or the Second Law of Thermodynamics. You can't be ignorant of first-person perspectives and superintelligent any more than you can be ignorant of the Second law of Thermodynamics and superintelligent. By analogy, just as autistic superintelligence captures the formal structure of a unified natural science, a mathematically complete "view from nowhere", all possible solutions to the universal Schrödinger equation or its relativistic extension, likewise a Full-Spectrum Superintelligence also grasps all possible first-person perspectives—and acts accordingly. In effect, an idealised Full-Spectrum Superintelligence would combine the mind-reading prowess of a telepathic mirror-touch synaesthete with the optimising prowess of a rule-following hyper-systematiser on a cosmic scale. If your hand is in the fire, you reflexively withdraw it. In withdrawing your hand, there is no question of first attempting to solve the Is-Ought problem in meta-ethics and trying logically to derive an "ought" from an "is". Normativity is built into the nature of the aversive experience itself: I-ought-not-to-be-in-this-dreadful-state. By extension, perhaps a Full-Spectrum Superintelligence will perform cosmic felicific calculus (Bentham 1789) and execute some sort of metaphorical hand-withdrawal for all accessible suffering sentience in its forward light-cone. Indeed one possible *criterion* of Full-Spectrum Superintelligence is the propagation of subjectively hypervaluable states on a cosmological scale.

What this constraint on intelligent agency means in practice is unclear. Conceivably at least, idealised Superintelligences must ultimately do what a classical utilitarian ethic dictates and propagate some kind of "utilitronium shockwave" across the cosmos. To the classical utilitarian, any rate of time-discounting indistinguishable from zero is ethically unacceptable, so s/he should presumably be devoting most time and resources to that cosmological goal. An ethic of negative utilitarianism is often accounted a greater threat to intelligent life (*cf.* the hypothetical "button-pressing" scenario) than classical utilitarianism. But whereas a negative utilitarian believes that once intelligent agents have phased out the biology of suffering, all our ethical duties have been discharged, the classical utilitarian seems ethically committed to converting all accessible matter and energy into relatively homogeneous matter optimised for maximum bliss: "utilitronium". Hence the most empirically valuable outcome entails the extinction of intelligent life. Could this prospect derail superintelligence?

Perhaps but, utilitronium shockwave scenarios shouldn't be confused with *wireheading*. The prospect of self-limiting superintelligence might be credible if either a (hypothetical) singleton biological superintelligence or its artificial counterpart discovers intracranial self-stimulation or its nonbiological analogues. Yet is this blissful fate a threat to anyone else? After all, a wirehead doesn't aspire to convert the rest of the world into wireheads. A junkie isn't driven to turn the rest of the world into junkies. By contrast, a utilitronium shockwave propagating across our Hubble volume would be the product of intelligent design by an advanced civilisation, not self-subversion of an intelligent agent's reward circuitry. Also, consider the reason why biological humanity—as distinct from individual humans—is resistant to wirehead scenarios, namely *selection pressure*. Humans who discover the joys of intra-cranial self-stimulation or heroin aren't motivated to raise children. So they are outbred. Analogously, full-spectrum superintelligences, whether natural or artificial, are likely to be social rather than solipsistic, not least because of the severe selection pressure exerted against any intelligent systems who turn in on themselves to wirehead rather than seek out unoccupied ecological niches. In consequence, the adaptive radiation of natural and artificial intelligence across the Galaxy won't be undertaken by stay-at-home wireheads or their blissed-out functional equivalents.

On the face of it, this argument from selection pressure undercuts the prospect of superhappiness for all sentient life—the "attractor" towards which we may tentatively predict sentience is converging in virtue of the pleasure principle harnessed to ultraintelligent mind-reading and utopian neuroscience. But what is necessary for sentient intelligence is information-sensitivity to fitness-relevant stimuli—not an agent's absolute location on the pleasure-pain axis. True, uniform bliss and uniform despair are inconsistent with intelligent agency. Yet mere recalibration of a subject's "hedonic set-point" leaves intelligence intact. Both information-sensitive gradients of bliss and information-sensitive gradients of misery allow high-functioning performance and critical insight. Only sentience animated by gradients of bliss is consistent with a rich subjective quality of intelligent life. Moreover, the nature of "utilitronium" is as obscure as its theoretical opposite, "dolorium". The problem here cuts deeper than mere lack of technical understanding, e.g. our ignorance of the gene expression profiles and molecular signature of pure bliss in neurons of the rostral shell of the nucleus accumbens and ventral pallidum, the twin cubic centimetre-sized "hedonic hot-spots" that generate ecstatic well-being in the mammalian brain (Berridge and Kringelbach 2010). Rather there are difficult conceptual issues at stake. For just as the torture of one mega-sentient being may be accounted worse than a trillion discrete pinpricks, conversely the sublime experiences of utilitronium-driven Jupiter minds may be accounted preferable to tiling our Hubble volume with the maximum abundance of micro-bliss. What is the optimal trade-off between quantity and intensity? In short, even assuming a classical utilitarian ethic, the optimal distribution of matter and energy that a God-like superintelligence would create in any given Hubble volume is very much an open question.

Of course we've no grounds for believing in the existence of an omniscient, omnipotent, omnibenevolent God or a divine utility function. Nor have we grounds for believing that the source code for any future God, in the fullest sense of divinity, could ever be engineered. The great bulk of the Multiverse, and indeed a high measure of life-supporting Everett branches, may be inaccessible to rational agency, quasi-divine or otherwise. Yet His absence needn't stop rational agents intelligently fulfilling what a notional benevolent deity would wish to accomplish, namely the well-being of all accessible sentience: the richest abundance of empirically hypervaluable states of mind in their Hubble volume. Recognisable extensions of existing technologies can phase out the biology of suffering on Earth. But responsible stewardship of the universe within our cosmological horizon depends on biological humanity surviving to become posthuman superintelligence.

Paradise Engineering?

The hypothetical shift to life lived entirely above Sidgwick's (1907) "hedonic zero" will mark a momentous evolutionary transition. What lies beyond? There is no reason to believe that hedonic ascent will halt in the wake of the world's last aversive experience in our forward light-cone. Admittedly, the self-intimating urgency of eradicating *suffering* is lacking in any further hedonic transitions, i.e. a transition from the biology of happiness (Schlaepfer and Fins 2012) to a biology of superhappiness; and then beyond. Yet why "lock in" mediocrity if intelligent life can lock in sublimity instead?

Naturally, superhappiness scenarios could be misconceived. Long-range prediction is normally a fool's game. But it's worth noting that future life based on gradients of intelligent bliss isn't tied to any particular ethical theory: its assumptions are quite weak. Radical recalibration of the hedonic treadmill is consistent not just with classical or negative utilitarianism, but also with preference utilitarianism, Aristotelian virtue theory, a Kantian deontological ethic, pluralist ethics, Buddhism, and many other value systems besides. Recalibrating our hedonic set-point doesn't—or at least needn't—undermine critical discernment. All that's needed for the abolitionist project and its hedonistic extensions (Pearce 1995) to succeed is that our ethic isn't committed to perpetuating the biology of involuntary suffering. Likewise, only a watered-down version of psychological hedonism is needed to lend the scenario sociological credibility. We can retain as much—or as little—of our existing preference architecture as we please. You can continue to prefer Shakespeare to Mills-and-Boon, Mozart to Morrissey, and Picasso to Jackson Pollock while living perpetually in Seventh Heaven or beyond.

Nonetheless an exalted hedonic baseline will revolutionise our conception of life. The world of the happy is quite different from the world of the unhappy, says Wittgenstein; but the world of the super happy will feel unimaginably different from the human, Darwinian world. Talk of preference conservation may reassure bioconservatives that nothing worthwhile will be lost in the post-Darwinian transition. Yet life based on information-sensitive gradients of superhappiness will most likely be "encephalized" in state-spaces of experience alien beyond human comprehension. Humanly comprehensible or otherwise, enriched hedonic tone can make all experience generically hypervaluable in an empirical sense—its lows surpassing today's peak experiences. Will such experience be hypervaluable in a metaphysical sense too? Is this question cognitively meaningful?

The Future of Sentience

The Sentience Explosion

Man proverbially created God in his own image. In the age of the digital computer, humans conceive God-like *super*intelligence in the image of our dominant technology and personal cognitive style—refracted, distorted and extrapolated for sure, but still through the lens of human concepts. The "super-" in so-called superintelligence is just a conceptual fig-leaf that humans use to hide our ignorance of the future. Thus high-AQ/high-IQ humans (Baron-Cohen 2001) may imagine God-like intelligence as some kind of Super-Asperger—a mathematical theorem-proving hyper-rationalist liable systematically to convert the world into computronium for its awesome theorem-proving. High-EQ, low-AQ humans, on the other hand, may imagine a cosmic mirror-touch synaesthete nurturing creatures great and small in expanding circles of compassion. From a different frame of reference, psychedelic drug investigators may imagine superintelligence as a Great Arch-Chemist opening up unknown state-space of consciousness. And so forth. Probably the only honest answer is to say, lamely, boringly, uninspiringly: we simply don't know.

Grand historical meta-narratives are no longer fashionable. The contemporary Singularitarian movement is unusual insofar as it offers one such grand meta-narrative: history is the story of simple biological intelligence evolving through natural selection to become smart enough to conceive an abstract universal Turing machine, build and program digital computers—and then merge with, or undergo replacement by, recursively self-improving artificial superintelligence.

Another grand historical meta-narrative views life as the story of overcoming suffering. Darwinian life is characterised by pain and malaise. One species evolves the capacity to master biotechnology, rewrites its own genetic source code, and creates post-Darwinian superhappiness. The well-being of all sentience will be the basis of post-Singularity civilisation: primitive biological sentience is destined to become blissful supersentience.

These meta-narratives aren't mutually exclusive. Indeed on the story told here, Full-Spectrum Superintelligence entails full-blown supersentience too: a seamless unification of the formal and the subjective properties of mind.

If the history of futurology is any guide, the future will confound us all. Yet in the words of Alan Kay: "It's easier to invent the future than to predict it".

References

Baker, S. (2011). Final Jeopardy: man vs. machine and the quest to know everything. (Houghton Mifflin Harcourt).

Ball, P. (2011). Physics of life: The dawn of quantum biology. *Nature, 474*(2011), 272–274.

Banissy, M., et al. (2009). Prevalence, characteristics and a neurocognitive model of mirror-touch synaesthesia. *Experimental Brain Research, 198*(2–3), 261–272. doi:10.1007/s00221-009-1810-9.

Barkow, J., Cosimdes, L., & Tooby, J. (Eds.). (1992). *The adapted mind: Evolutionary psychology and the generation of culture*. New York: Oxford University Press.

Baron-Cohen, S. (1995). Mindblindness: an essay on autism and theory of mind (MIT Press/ Bradford Books).

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., Clubley, E. (2001). The autism-spectrum quotient (AQ): evidence from Asperger syndrome/high functioning autism, males and females, scientists and mathematicians. *Journal of Autism Development Disorders*, *31*(1): 5–17. doi:10.1023/A:1005653411471. PMID 11439754.

Baron-Cohen, S. (2001). Autism spectrum questionnaire. (Cambridge: Autism Research Centre, University of Cambridge) http://psychology-tools.com/autism-spectrum-quotient/.

Benatar, D. (2006). *Better Never to Have Been: The Harm of Coming Into Existence*. Oxford: Oxford University Press.

Bentham, J. (1789). *An introduction to the principles of morals and legislation*. Oxford: Clarendon Press. Reprint.

Berridge, K. C., & Kringelbach, M. L. (Eds.). (2010). *Pleasures of the Brain*. Oxford: Oxford University Press.

Bostrom, N. (2002). Existential risks: analyzing human extinction scenarios and related hazards. *Journal of Evolution and Technology*, 9.

Boukany, P. E., et al. (2011). Nanochannel electroporation delivers precise amounts of biomolecules into living cells. *Nature Nanotechnology, 6*(2011), 74.

Brickman, P., Coates, D., & Janoff-Bulman, R. (1978). Lottery winners and accident victims: is happiness relative? *Journal of Personal Society Psychology, 36*(8), 917–927. 7–754.

Brooks, R. (1991). Intelligence without representation. *Artificial Intelligence, 47*(1–3), 139–159. doi:10.1016/0004-3702(91)90053-M.

Buss, D. (1997). "Evolutionary Psychology: The New Science of the Mind". (Allyn & Bacon).

Byrne, R., & Whiten, A. (1988). *Machiavellian intelligence*. Oxford: Oxford University Press.

Carroll, J. B. (1993). *Human cognitive abilities: a survey of factor-analytic studies*. Cambridge: Cambridge University Press.

Chalmers, D. J. (2010). The singularity: a philosophical analysis. *Journal of Consciousness Studies, 17*(9), 7–65.

Chalmers, D. J. (1995). Facing up to the hard problem of consciousness. *Journal of Consciousness Studies, 2*(3), 200–219.

Churchland, P. (1989). *A neurocomputational perspective: the nature of mind and the structure of science*. Cambridge: MIT Press.

Cialdini, R. B. (1987). Empathy-based helping: is it selflessly or selfishly motivated? *Journal of Personality and Social Psychology, 52*(4), 749–758.

Clark, A. (2008). *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. USA: Oxford University Press.

Cochran, G., Harpending, H. (2009). The 10,000 Year Explosion: How Civilization Accelerated Human Evolution New York: Basic Books.

Cochran, G., Hardy, J., & Harpending, H. (2006). Natural history of Ashkenazi intelligence. *Journal of Biosocial Science, 38*(5), 659–693.

Cohn, N. (1957). The pursuit of the millennium: revolutionary millenarians and mystical anarchists of the middle ages (Pimlico).

Dawkins, R. (1976). *The Selfish Gene*. New York City: Oxford University Press.

de Garis, H. (2005). The Artilect War: Cosmists vs. Terrans: A bitter controversy concerning whether humanity should build godlike massively intelligent machines. ETC. Publications, pp. 254. ISBN 978-0882801537.

de Grey, A. (2007). *Ending aging: The rejuvenation breakthroughs that could reverse human aging in our lifetime*. New York: St. Martin's Press.

Delgado, J. (1969). *Physical control of the mind: Toward a psychocivilized society*. New York: Harper and Row.

Dennett, D. (1987). *The intentional stance*. Cambridge: MIT Press.

Deutsch, D. (1997). *The fabric of reality*. Harmondsworth: Penguin.

Deutsch, D. (2011). *The beginning of infinity*. Harmondsworth: Penguin.

Drexler, E. (1986). *Engines of creation: The coming era of nanotechnology*. New York: Anchor Press/Doubleday.

Dyson, G. (2012). *Turing's cathedral: The origins of the digital universe*. London: Allen Lane.

Everett, H. (1973). The theory of the universal wavefunction. Manuscript (1955), pp 3–140 In B. DeWitt, & R. N. Graham, (Eds.), The many-worlds interpretation of quantum mechanics. Princeton series in physics. Princeton, Princeton University Press. ISBN 0-691-08131-X.

Francione, G. (2006). Taking sentience seriously. Journal of Animal Law & Ethics 1.

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences*. New York: Basic Books.

Goertzel, B. (2006). The hidden pattern: A patternist philosophy of mind. (Brown Walker Press).

Good, I. J. (1965). Speculations concerning the first ultraintelligent machine. In L. F. Alt & M. Rubinoff (Eds.), *Advances in computers* (pp. 31–88). London: Academic Press.

Gunderson, K. (1985). *Mentality and machines*. Minneapolis: University of Minnesota Press.

Hagan, S., Hameroff, S., & Tuszynski, J. (2002). Quantum computation in brain microtubules? Decoherence and biological feasibility. *Physical Reviews, E65*, 061901.

Haidt, J. (2012). *THE righteous mind: Why good people are divided by politics and religion*. NY: Pantheon.

Hameroff, S. (2006). Consciousness, neurobiology and quantum mechanics. In The emerging physics of consciousness, J. Tuszynski (Ed.) (Springer).

Harris, S. (2010). *The moral landscape: How science can determine human values*. NY: Free Press.

Haugeland, J. (1985). *Artificial intelligence: The very idea*. Cambridge: MIT Press.

Holland, J. (2001). Ecstasy: The complete guide: A comprehensive look at the risks and benefits of mdma. (Park Street Press).

Holland, J. H. (1975). *Adaptation in natural and artificial systems*. Ann Arbor: University of Michigan Press.

Hutter, M. (2010). *Universal artificial intelligence: Sequential decisions based on algorithmic probability*. New York: Springer.

Hutter, M. (2012). Can intelligence explode? *Journal of Consciousness Studies, 19*, 1–2.

Huxley, A. (1932). Brave New World. London: Chatto and Windus.

Huxley, A. (1954). Doors of perception and heaven and hell. New York: Harper & Brothers.

Kahneman, D. (2011). *Thinking, fast and slow*. Straus and Giroux: Farrar.

Kant, I. (1781). Critique of pure reason. In P. Guyer & A. Wood. Cambridge: Cambridge University Press, 1997.

Koch, C. (2004). The Quest for Consciousness: a Neurobiological Approach. Roberts & Co..

Kurzweil, R. (2005). The singularity is near. New York: Viking.

Kurzweil, R. (1990). The age of intelligent machines. Cambridge: MIT Press

Kurzweil, R. (1998). The age of spiritual machines. New York: Viking.

Langdon, W., Poli, R. (2002). Foundations of genetic programming. New York: Springe).

Lee, H. J., Macbeth, A. H., Pagani, J. H., Young, W. S. (2009). Oxytocin: The great facilitator of life. *Progress in Neurobiology 88* (2): 127–51. doi:10.1016/j.pneurobio.2009.04.001. PMC 2689929. PMID 19482229.

Legg, S., Hutter, M. (2007). Universal intelligence: A definition of machine intelligence. *Minds & Machines*, *17*(4), pp. 391–444.

Levine, J. (1983). Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly, 64*, 354–361.

Litt, A., et al. (2006). Is the Brain a Quantum Computer? *Cognitive Science, 20*, 1–11.

Lloyd, S. (2002). Computational capacity of the universe. *Physical Review Letters 88*(23): 237901. arXiv:quant-ph/0110141. Bibcode 2002PhRvL..88w7901L.

Lockwood, L. (1989). Mind, brain, and the quantum. Oxford: Oxford University Press.
Mackie, JL. (1991). Ethics: Inventing right and wrong. Harmondsworth: Penguin.
Markram, H. (2006). The blue brain project. *Nature Reviews Neuroscience*, *7*:153–160. PMID 16429124.
Merricks, T. (2001). Objects and persons. Oxford: Oxford University Press.
Minsky, M. (1987). The society of mind. New York: Simon and Schuster.
Moravec, H. (1990). Mind children: The future of robot and human intelligence. Cambridge: Harvard University Press.
Nagel, T. (1974). What is it Like to Be a Bat? *Philosophical Review, 83*, 435–450.
Nagel, T. (1986). The view from nowhwere. Oxford: Oxford University Press.
Omohundro, S. (2007). The Nature of Self-Improving Artificial Intelligence. San Francisco: Singularity Summit.
Parfit, D. (1984). *Reasons and persons*. Oxford: Oxford University Press.
Pearce, D. (1995). The hedonistic imperative http://hedweb.com.
Pellissier, H. (2011) Women-only leadership: Would it prevent war? http://ieet.org/index.php/IEET/more/4576.
Penrose, R. (1994). Shadows of the mind: A search for the missing science of consciousness. Cambridge: MIT Press.
Peterson, D., Wrangham, R. (1997). Demonic males: Apes and the origins of human violence. Mariner Books.
Pinker, S. (2011). The better angels of our nature: Why violence has declined. New York: Viking.
Rees, M. (2003). Our final hour: A scientist's warning: How terror, error, and environmental disaster threaten humankind's future in this century—on earth and beyond. New York: Basic Books.
Reimann F, et al. (2010). Pain perception is altered by a nucleotide polymorphism in SCN9A. *Proc Natl Acad Sci* USA. *107*(11):5148–5153 (2010 Mar 16).
Rescher, N. (1974). Conceptual idealism. Oxford: Blackwell Publishers.
Revonsuo, A. (2005). Inner presence: Consciousness as a biological phenomenon. Cambridge: MIT Press.
Revonsuo, A., & Newman, J. (1999). Binding and consciousness. *Consciousness and Cognition, 8*, 123–127.
Riddoch, M. J., & Humphreys, G. W. (2004). Object identification in simultanagnosia: When wholes are not the sum of their parts. *Cognitive Neuropsychology, 21*(2–4), 423–441.
Rumelhart, D. E., McClelland, J. L., & the PDP Research Group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations*. Cambridge: MIT Press.
Russell, B. (1948). *Human knowledge: Its scope and limits*. London: George Allen & Unwin.
Sandberg, A., Bostrom, N. (2008). Whole brain emulation: A roadmap. Technical report 2008-3.
Saunders, S., Barrett, J., Kent, A., Wallace, D. (2010). Many worlds: Everett, quantum theory, and reality. Oxford: Oxford University Press.
Schlaepfer, T. E., & Fins, J. J. (2012). How happy is too happy? Euphoria, neuroethics and deep brain stimulation of the nucleus accumbens. *The American Journal of Bioethics, 3*, 30–36.
Schmidhuber, J. (2012). Philosophers & Futurists, Catch Up! Response to the singularity. *Journal of Consciousness Studies*, *19*: 1–2, pp. 173–182.
Seager, W. (1999). Theories of consciousness. London: Routledge.
Seager, W. (2006). The 'intrinsic nature' argument for panpsychism. *Journal of Consciousness Studies, 13*(10–11), 129–145.
Sherman, W., Craig A., (2002). Understanding virtual reality: Interface, application, and design. Los Altos: Morgan Kaufmann.
Shulgin, A. (1995). PiHKAL: A chemical love story. Berkeley: Transform Press.
Shulgin, A. (1997). TiHKAL: The continuation. Berkeley: Transform Press.
Shulgin, A. (2011). The shulgin index vol 1: Psychedelic phenethylamines and related compounds. Berkeley: Transform Press.

Shulman, C., Sandberg, A. (2010). Implications of a software-limited singularity. Proceedings of the European Conference of Computing and Philosophy.

Sidgwick, H. (1907). The methods of ethics. Indianapolis: Hackett, seventh edition, 1981, 1-4.

Singer, P. (1995). *Animal liberation: A new ethics for our treatment of animals*. New York: Random House.

Singer, P. (1981). *The expanding circle: Ethics and sociobiology*. New York: Farrar, Straus and Giroux.

Smart, JM. (2008–2011) Evo Devo Universe? A Framework for Speculations on Cosmic Culture. In Cosmos and culture: Cultural evolution in a cosmic context. J. S. Dick, M. L. Lupisella (Eds.), Govt Printing Office, NASA SP-2009-4802, Wash., D.C., 2009, pp. 201–295.

Stock, G. (2002). Redesigning humans: Our inevitable genetic future. Boston: Houghton Mifflin Harcourt.

Strawson G., et al. (2006). Consciousness and its place in nature: Does physicalism entail panpsychism? Imprint Academic.

Tegmark, M. (2000). Importance of quantum decoherence in brain processes. *Physical Review E, 61*(4), 4194–4206. doi:10.1103/PhysRevE.61.4194.

Tsien, J. et al., (1999). Genetic enhancement of learning and memory in mice. *Nature 401*, 63–69 (2 September 1999) | doi:10.1038/43432.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*, 433–460.

Vinge, V. (1993). *The coming technological singularity*. New Whole Earth LLC: Whole Earth Review.

Vitiello, G. (2001). My double unveiled; advances in consciousness. Amsterdam: John Benjamins.

Waal, F. (2000). Chimpanzee politics: Power and sex among apes". Maryland: Johns Hopkins University Press.

Wallace, D. (2012). *The Emergent multiverse: Quantum theory according to the Everett interpretation*. Oxford: Oxford University Press.

Welty, G. (1970). The history of the prediction paradox, presented at the Annual Meeting of the International Society for the History of the Behavioral and Social Sciences, Akron, OH (May 10, 1970), Wright State University Dayton, OH 45435 USA. http://www.wright.edu/~gordon.welty/Prediction_70.htm.

Wohlsen, M. (2011): Biopunk: DIY scientists hack the software of life. London: Current.

Yudkowsky, E. (2007). Three major singularity schools. http://yudkowsky.net/singularity/schools.

Yudkowsky, E. (2008). Artificial intelligence as a positive and negative factor in global risk. In Bostrom, Nick and Cirkovic, Milan M. (Eds.). Global catastrophic risks. pp. 308–345 (Oxford: Oxford University Press).

Zeki, S. (1991). Cerebral akinetopsia (visual motion blindness): A review. *Brain, 114*, 811–824. doi:10.1093/brain/114.2.811.

Chapter 11A
Illah R. Nourbakhsh on Pearce's "The Biointelligence Explosion"

The Optimism of Discontinuity

In *The Biointelligence Explosion*, David Pearce launches a new volley in the epic, pitched battle of today's futurist legions. The question of this age is: *machine or man?* And neither machine nor man resembles the modern-day variety. According to the Singularity's version of foreshadowed reality, our successors are nothing like a simulacrum of human intelligence; instead they vault beyond humanity along every dimension, achieving heights of intelligence, empathy, creativity, awareness and immortality that strain the very definitions of these words as they stand today. Whether these super-machines embody our unnatural, disruptive posthuman evolution, displacing and dismissing our organic children, or whether they melt our essences into their circuitry by harvesting our consciousnesses and *qualia* like so much wheat germ, the core ethic of the machine disciples is that the future will privilege digital machines over carbon-based, analog beings.

Pearce sets up an antihero to the artificial superintelligence scenario, proposing that our wetware will shortly become so well understood, and so completely modifiable, that personal bio-hacking will collapse the very act of procreation into a dizzying tribute to the ego. Instead of producing children as our legacy, we will modify our own selves, leaving natural selection in the dust by changing our personal genetic makeup in the most extremely personal form of creative hacking imaginable. But just like the AI singularitarians, Pearce dreams of a future in which the new and its ancestor are unrecognizably different. Regular humans have depression, poor tolerance for drugs, and, let's face it, mediocre social, emotional and technical intelligence. Full-Spectrum Superintelligences will have perfect limbic mood control, infinite self-inflicted hijacking of chemical pathways, and so much intelligence as to achieve omniscience bordering on Godliness.

The Singularity proponents have a fundamentalist optimism born, as in all religions, of something that cannot be proven or disproven rationally: faith. In their case, they have undying faith in a future discontinuity, the likes of which the computational world has never seen. After all, as Pearce points out, today's computers have not shown even a smattering of consciousness, and so the ancestry of the intelligent machine, a machine so fantastically powerful that it can eventually invent the superintelligent machine, is so far an utter no-show. But this is alright if we can believe that with Moore's Law comes a new golden chalice: a point of no return, when the progress of Artificial Intelligence self-reinforces, finally, and takes off like an airplane breaking ground contact and suddenly shooting upward in the air: a discontinuity that solves all the unsolvable problems. No measurement of AI's effectiveness before the discontinuity matters from within this world view; the future depends only on the shape of a curve, and eventually all the rules will change when we hit a sudden bend. That a technical sub-field can depend so fully, not on early markers of success, but on the promise of an unknown future disruption, speaks volumes about the discouraging state of Artificial Intelligence today. When the best recent marker of AI, IBM's Watson, wins peculiarly by responding to a circuit-driven light in 8 ms, obviating the chances of humans who must look at a light and depend on neural pathways orders of magnitude slower, then AI Singularity cannot yet find a machine prophet.

Pearce is also an optimist, presenting an alternative view that extrapolates from the mile marker of yet another discontinuity: when hacker-dom successfully turns its tools inward, open-sourcing and bio-hacking their own selves to create recursively improving bio-hackers that rapidly morph away from mere human and into transcendental Superintelligence. The discontinuity is entirely different from the AI Singularity, and yet it depends just as much on a computational mini-singularity. Computers would need to provide the simulation infrastructure to enable bio-hackers to visualize and test candidate self-modifications. Whole versions of human-YACC and human-VMWare would need to compile and run entire human architectures in dynamic, simulated worlds to see just what behaviour will ensue when Me is replaced by Me-2.0. This demands a level of modelling, analog simulation and systems processing that depend on just as much of a discontinuity as the entire voyage. *And then a miracle happens* becomes almost cliché when every technical obstacle to be surmounted is not a mountain, but a hyperplane of unknown dimensionality!

But then there is the hairy underbelly of open-source genetics, namely that of systems engineering and open-source programming in general. As systems become more complex, Quality Assurance (QA) becomes oxymoronic because tests fail to exhaustively explore the state-space of possibilities. The Toyota Prius brake failures were not caught by engineers whose very job is to be absolutely sure that brakes *never*, *ever* fail, because just the right resonant frequency, combined with a hybrid braking architecture, combined with just the right accelerometer architecture and firmware, can yield a one-in-a million rarity a handful of times, literally. The logistical tail of complexity is a massive headache in the regime of QA, and this bodes poorly for open-sourced hacking of human systems, which dwarf the complexity of Toyota Prius exponentially. IDE's for bio-hacking; debuggers that can isolate part of your brain so that you can debug a nasty problem without losing consciousness (Game Over!); version control systems and repositories so that, in a panic, you can return your genomic identity to a most recent stable state- all of these tools will be needed, and we will of course be financially enslaved to the corporations that provide these self-modification tools. Will a company, let's call it *HumanSoft*, provide a hefty discount on its insertion vector applications if you agree to do some advertising—your compiled genome always drinks Virgil's Root Beer at parties, espousing its combination of Sweet Birch and Molasses? Will you upgrade to HumanSoft's newest IDE because it introduces forked compiling—now you can run two mini-me's in one body and switch between them every 5 s by reprogramming the brain's neural pathways.

Perhaps most disquieting is the law of unintended consequences, otherwise known as robotic compounding. In the 1980s, roboticists thought that they could build robots bottom-up, creating low-level behaviours, testing and locking them in, then adding higher-level behaviours until, eventually, human-level intelligence flowed seamlessly from the machine. The problem was that the second level induced errors in how level one functioned, and it took unanticipated debugging effort to get level one working with level two. By the time a roboticist reaches level four, the number of side effects overwhelms the original engineering effort completely, and funding dries up before success can be had. Once we begin bio-hacking, we are sure to discover side effects that the best simulators will fail to recognize unless they are equal in fidelity to the real-world. After how many major revisions will we discover that all our hacking time is spent trying to undo unintended consequences rather than optimizing desired new features? This is not a story of discontinuity, unfortunately, but the gradual build-up of messy, complicated baggage that gums up the works gradually and eventually becomes a singular centre of attention.

We may just discover that the Singularity, whether it gives rise to Full-Spectrum Superintelligence or to an Artificial Superintelligence, surfaces an entire stable of mediocre attempts long before something of real value is even conceivable. Just how many generations of mediocrity will we need to bridge and at what cost, to reach the discontinuity that is an existential matter of faith?

There is one easy answer here, at once richly appropriate and absurd. Pearce proposes that emotional self-control has one of the most profound consequences on our humanity, for we can make ourselves permanently happy. Learn to control the limbic system fully, and then bio-hackers can hack their way into enforced sensory happiness- indeed, even modalities of happiness that effervesce beyond anything our non drug-induced dreams can requisition today. Best of all, we could program ourselves for maximal happiness even if Me-2.0 is mediocre and buggy. Of course, this level of human chemical pathway control suggests a level of maturity that pharmaceutical companies dream about today, but if it is truly possible to obtain permanent and profound happiness all-around, then of course we lose both the condition and state of happiness. It becomes the drudgery that is a fact of life.

Finally, let us return to one significant commonality between the two hypotheses: they both demand that technology provide the ultimate modelling and simulation engine: I call it the Everything Engine. The Everything Engine is critical to AI because computers must reason, fully, about future implications of all state sets and actions. The Everything Engine is also at the heart of any IDE you would wish to use when hacking your genome: you need to model and generate evidence that your proposed personal modification yields a better you rather than a buggier you. But today, the Everything Engine is Unobtanium, and we know that incremental progress on computation speed will not produce it. We need a discontinuity in computational trends in order to arrive at the Everything Engine. Pearce is right when he states that the two meta-narratives of Singularity are not mutually exclusive. In fact, they are conjoined at the hip; for, if their faith in a future discontinuity proves false, then we might just need infinity of years to reach either Nirvana. And if the discontinuity arrives soon, then as Pearce points out, we will all be too busy inventing the future or evading the future to predict the future.‘