An Attempt to Generalize AI Part 14: Mind Control Speculation

By Paul Almond

24 June 2010

This is the fourteenth in a series of articles attempting an overview of how minds may work and how similar systems could be implemented in computers. Previous articles described a probabilistic hierarchy based on *patterns*. A pattern has a specification describing a set, or population, of *pattern instances*, distributed throughout a hierarchy containing the pattern instances of all the patterns. Each pattern's set of pattern instances is used to obtain statistical information for probabilistic predictions. Each pattern's population of pattern instances is to be described in a very general way, to provide a very general ontology. The hierarchy's structure is based on the history of previous inputs/outputs, so "faking" outputs at the point of entry into the hierarchy as bottom-level pattern instances could be used to control the representation of the world in the hierarchy. The behavior of a human brain may be altered by such "faking" of the outputs that are stored in bottom-level pattern instances. The "self" has no qualitatively special status, but is merely an object in the hierarchy's representation of the world, and would therefore be altered in such a process. The extent to which such a process would affect a human would depend on the extent to which the human brain is like the hierarchical system described in this series of articles. A non-invasive version of the process, involving manipulation of the environment, is also discussed, as is combining this with the process of manipulating outputs. The more extreme idea of copying an entire mind, or some other computer system, into an artificial hierarchy or a human brain is discussed.

Table of Contents

List of Abbreviations

AI   artificial intelligence

BERP   basic, exploratory relevance process

EFS   evaluation function score

ERP   exploratory relevance process

1 Introduction

This article is the fourteenth in a series about artificial intelligence (AI) and how our own minds might work. The first article, *An Attempt to Generalize AI - Part 1: The Modeling System*, is available at http://www.paul-almond.com/AI01.pdf.[1] The second article, *An Attempt to Generalize AI - Part 2: Planning and Actions*, is at http://www/paul-almond.com/AI02.pdf.[2] The third article, *An Attempt to Generalize AI - Part 3: Forgetting*, is at http://www.paul-almond.com/AI03.pdf.[3]

These three articles described a hierarchy based on *patterns*, which are sets of *pattern instances*, and were intended to give an idea of how humans may model the world, plan actions and discard information from the model when it is no longer useful. The fourth article, *An Attempt to Generalize AI - Part 4: Modeling Efficiency*, which is at http://www.paul-almond.com/AI04.pdf, suggested that pattern instances should be allowed to have *incompletely specified pattern inputs*, so that it would be practical for the hierarchy to be "pruned" by some process seeking to maximize its relevance.[4] This required a *completely* probabilistic hierarchy, an issue dealt with in the fifth article of this series, *An Attempt to Generalize AI - Part 5: A Completely Probabilistic Hierarchy*, which is at http://www.paul-almond.com/AI05.pdf.[5]

That made a process to provide relevance in the hierarchy feasible. The sixth article, *An Attempt to Generalize AI – Part 6: Measuring Relevance*, which is at http://www.paul-almond.com/AI06.pdf, described a back-propagation process for measuring relevance in the hierarchy.[6] The problem is made tractable by the way in which the *action selection process*, described in the second article, *An Attempt to Generalize AI – Part 2: Planning and Actions*, works. The seventh article, *An Attempt to Generalize AI – Part 7: A Basic, Exploratory Relevance Process*, which is at http://www.paul-almond.com/AI07.pdf, described the *basic exploratory relevance process* (BERP), which uses this measuring process to direct the growth and pruning of the hierarchy.[7]

---

[1] Almond, P., 2010. *An Attempt to Generalize AI - Part 1: The Modeling System*. [Online] paul-almond.com. http://www.paul-almond.com/AI01.pdf or http://www.paul-almond.com/AI01.doc.

[2] Almond, P., 2010. *An Attempt to Generalize AI - Part 2: Planning and Actions*. [Online] paul-almond.com. http://www.paul-almond.com/AI02.pdf or http://www.paul-almond.com/AI02.doc.

[3] Almond, P., 2010. *An Attempt to Generalize AI - Part 3: Forgetting*. [Online] paul-almond.com. http://www.paul-almond.com/AI03.pdf or http://www.paul-almond.com/AI03.doc.

[4] Almond, P., 2010. *An Attempt to Generalize AI - Part 4: Modeling Efficiency*. [Online] paul-almond.com. http://www.paul-almond.com/AI04.pdf or http://www.paul-almond.com/AI04.doc.

[5] Almond, P., 2010. *An Attempt to Generalize AI - Part 5: A Completely Probabilistic Hierarchy*. [Online] paul-almond.com. http://www.paul-almond.com/AI05.pdf or http://www.paul-almond.com/AI05.doc.

[6] Almond, P., 2010. *An Attempt to Generalize AI - Part 6: Measuring Relevance*. [Online] paul-almond.com. http://www.paul-almond.com/AI06.pdf or http://www.paul-almond.com/AI06.doc.

[7] Almond, P., 2010. *An Attempt to Generalize AI - Part 7: A Basic, Exploratory Relevance Process.*

The eighth article, *An Attempt to Generalize AI – Part 8: Forgetting as Part of the Exploratory Relevance Process*, which is at http://www.paul-almond.com/AI08.pdf, removed the need for the forgetting process in the third article, instead incorporating forgetting into the BERP, or any other exploratory relevance process (ERP).[8] This was done by modifying the relevance measurement process (RMP) to take account of obsolescence. This article also introduced *ghost pattern instances*. A ghost pattern instance is one that persists temporarily, after "removal" by the ERP, as a simple probability value, while it is still needed as a pattern input by other pattern instances. The incompletely specified pattern inputs introduced earlier are now ghost pattern instances. The ninth article, *An Attempt to Generalize AI – Part 9: Improving the Exploratory Relevance Process*, which is at http://www.paul-almond.com/AI09.pdf, discussed ways in which the sophistication of the BERP might be increased, giving an improved ERP.[9]

Functioning of the system requires pattern instances to be placed in the hierarchy on an ongoing basis, and pattern instances need to belong to patterns. A way in which this could work had been described in the first article, *An Attempt to Generalize AI - Part 1: The Modeling System*.[10] Other approaches, conforming to the same general idea, were discussed in *An Attempt to Generalize AI – Part 10: Alternatives for Pattern Instance Construction*, which is at http://www.paul-almond.com/AI10.pdf.[11]

The view of cognition in this series was related to dreaming in humans, with a suggestion for how it occurs in *An Attempt to Generalize AI – Part 11: Explaining Dreaming*, which is at http://www.paul-almond.com/AI11.pdf.[12]

The issue of ensuring relevance for patterns, rather than just for pattern *instances*, as had been considered in previous discussion of the BERP and other ERPs, was discussed

---

[Online] paul-almond.com. http://www.paul-almond.com/AI07.pdf or http://www.paul-almond.com/AI07.doc.

[8] Almond, P., 2010. *An Attempt to Generalize AI – Part 8: Forgetting as Part of the Exploratory Relevance Process*. [Online] paul-almond.com. http://www.paul-almond.com/AI08.pdf or http://www.paul-almond.com/AI08.doc.

[9] Almond, P., 2010. *An Attempt to Generalize AI - Part 9: Improving the Exploratory Relevance Process*. [Online] paul-almond.com. http://www.paul-almond.com/AI09.pdf or http://www.paul-almond.com/AI09.doc.

[10] Almond, P., 2010. *An Attempt to Generalize AI - Part 1: The Modeling System*. [Online] paul-almond.com. http://www.paul-almond.com/AI01.pdf or http://www.paul-almond.com/AI01.doc.

[11] Almond, P., 2010. *An Attempt to Generalize AI - Part 10: Alternatives for Pattern Instance Construction*. [Online] paul-almond.com. http://www.paul-almond.com/AI0.pdf or http://www.paul-almond.com/AI10.doc.

[12] Almond, P., 2010. *An Attempt to Generalize AI - Part 11: Explaining Dreaming*. [Online] paul-almond.com. http://www.paul-almond.com/AI11.pdf or http://www.paul-almond.com/AI11.doc.

in *An Attempt to Generalize AI – Part 12: Pattern Relevance*, which is at http://www.paul-almond.com/AI12.pdf.[13]

Relevance was also discussed in *An Attempt to Generalize AI – Part 13: Reflexive Outputs*, which is at http://www.paul-almond.com/AI13.pdf.[14] This article discussed the possibility of using special outputs which are directed inwards, acting on the actual hierarchy itself.

This article, rather than being about developing the approach to AI and cognition, will be about one of its implications. The view of cognition taken in this series, if correct, suggests the possibility of a range of technologies relating to mind control: changing the personalities and behavior of humans. This and similar issues will be discussed in this article. So that there is no doubt what this article is about, it is about *making people do things*.

I have had to consider whether this is a suitable subject for an article. If technology like this is possible, there are clearly unethical uses of it. There are also ethical ones, however: Some people would *want* their minds modified. Other arguments could be made for and against publishing this, and I will not give them here: They are already widely known in other contexts. Ultimately, I have decided that nothing would be achieved by not writing about this matter.

---

[13] Almond, P., 2010. *An Attempt to Generalize AI - Part 12: Pattern Relevance*.
[Online] paul-almond.com. http://www.paul-almond.com/AI12.pdf or
http://www.paul-almond.com/AI12.doc.
[14] Almond, P., 2010. *An Attempt to Generalize AI - Part 13: Reflexive Outputs*.
[Online] paul-almond.com. http://www.paul-almond.com/AI13.pdf or
http://www.paul-almond.com/AI13.doc.

2 How Mind Control Might Work

2.1 The Nature of the "Self"

If the view of cognition described in this series of articles is correct, mind control technologies could be developed that rely on the nature of the "self" and the way that behavior is produced by the hierarchy.

The view of the "self" that has been described is that it has no qualitatively special status in the system, but is just part of the model generated from the history of inputs/outputs. The "self" happens to be a part of the model that particularly strongly relates past inputs/outputs with future outputs, and therefore describes intentionality. Metzinger has previously described a view in which the "self" is just part of a model.[15]

Like everything else in the hierarchy, the "self" is generated from the history of inputs/outputs, and therefore the history of inputs/outputs determines the self, as well as all the behavior of the system.

2.2 Modifying the "Self"

If we have an AI system like the one discussed in this series, we could change the future behavior and the "self" in the model by *changing the history of outputs*. We would need to access the part of the AI system where the outputs that have occurred are fed back into the bottom-level pattern instances of the hierarchy, and we would need to change the data being fed back into the hierarchy, replacing some of the real outputs with fake outputs, as they occur, so that the system is tricked into "thinking" it has made the fake outputs, instead of the ones that it has actually made. All of the hierarchy is produced from the history of inputs/outputs, and this fake behavioral history would form the basis of the "self" in the hierarchy, and the system's future behavior. The behavior of the system would be modified.

This hierarchical view of cognition described in this series of articles, as well as being an AI proposal, is a proposed model of cognition in *humans*. This seems to suggest that human minds could be similarly modified by altering the history of outputs. Although this would require some direct manipulation of the brain, it would presumably be a quite shallow one, in terms of intervening in a (relatively) simple process.

---

[15] Metzinger, T., 2003. *Being No One: The Self-Model Theory of Subjectivity*. Cambridge (MA): MIT Press. Metzinger, T., 2009. *The EGO Tunnel: The Science of the Mind and the Myth of the Self*. New York: Basic Books.

2.3 Practical Considerations

A process like this will have limitations. In reality, things may not be as simple as the cognitive model in this series of articles. Even if the cognitive model is generally correct, the mind may use this and lots of special-case processing, or it may even need more than one cognitive model to describe it. If the human mind is not completely as discussed in this series, this will limit any attempt to exploit features of the system described here for mind control. This does not mean that mind control would fail completely: If the cognitive model captures *some* general functioning of the mind, it is likely that mind control attempts performed on this basis would have *some* success – maybe even a significant degree of success.

The "self" generated by the hierarchy will be based on the past, real outputs, before any interference with the outputs started, as well as on the outputs after interference started. It will be an explanation of this sequence of outputs. If the change in behavior is very implausible, it will be more plausible to explain the behavior in terms of the new outputs not reflecting the genuine intentions of the "self". For example, if the subject's favorite food is pizza, and you suddenly start faking outputs that suggest that the subject hates pizza, it may be implausible that the subject has suddenly gone from liking pizza to hating it. Instead, a "self" might emerge which likes pizza but has been pretending to hate it for some reason, if this more plausibly explains the sudden change in behavior. For this reason, an attempt to modify the subject's behavior may need to proceed gradually, making only small changes to the outputs which give a significant behavioral change over a period of time. Furthermore, if the outputs are altered so much that the subject's situation should be very different from the one described by the inputs, then unless these are interfered with as well (and they may be completely faked in a very advanced version of this method), this would create a noticeable inconsistency.

If the subject knows or suspects that his/her behavior is being modified, then a "self" might result which explains the faked outputs in terms of some non-genuine behavior by the subject or remembered by the subject.

As the new "self" will result from the previous, unaltered output history and the more recent, modified output history, the previous, unaltered output history effectively needs to be overcome in the modification process. If the subject has had a particular attitude for a long time, for example, it may be less plausible for it to be changing. This suggests that the age of the subject could be an issue: An older subject would mean that there is more previous history that needs explaining away.

Even with various limitations in place, it may still be possible to effect significant modification to a subject's personality. Altering of outputs might be combined with other approaches to make up for the limitations of the process. Positive and negative reinforcement may be used after the behavioral modification to cause the subject to

exhibit the required behavior. Some memory erasure method may be used to reduce the problem of having to overcome a pre-existing history.

2.4 Non-Invasive Methods?

The method that has been described requires direct manipulation of the subject's brain, but could we use less invasive techniques to trick the brain into thinking that it has made outputs that it has not made?

This might be done by altering information in the external environment, to misrepresent evidence of what the person has done. For example, if we wanted to modify the subject's political views, we might ask the subject to answer a long series of questions. For each question, the subject could be asked to place a mark on a line to indicate his/her level of agreement or disagreement with some statement: Putting the mark on the far left could indicate complete agreement, while putting it on the far right could indicate complete disagreement, and putting it somewhere between could indicate some level of agreement or disagreement. Afterwards, we might discuss the answers with the subject, but we might slightly alter the position of the marks for some of answers without the subject knowing, moving the mark slightly to the left or right in each case. The idea would be to trick the subject's brain into thinking that his/her answers were slightly different to what they were, so that the "self" would be modified to be consistent with the slightly altered answers. Over time, the subject would be gradually "walked" over to a different political view.

This is far from guaranteed to work. The main problem is that the information available from the environment is being altered, but the outputs being fed into the bottom-level pattern instances are not being altered. Experimentation might be worthwhile, however.

2.5 Manipulation of Inputs and Outputs

An approach like the one just described, in 2.4, may work better if combined with the direct manipulation of the brain – the faking of the outputs fed into the hierarchy – that has already been described in 2.2, with the external environment being rearranged to be consistent with the "faked" outputs.

The most extreme example of this would be complete immersion in virtual reality, with the outputs made by the system being altered before being fed into the bottom level of the hierarchy, and the virtual reality environment being modified to be consistent with the faked outputs.

3 Copying a Mind

A process more extreme than the one just described would be one in which the *complete* input/output history of a hierarchy needed to set up a "self" was provided to it from outside.

Such a process would start with a hierarchy with no input/output history. All the inputs/outputs would then be provided to the bottom-level pattern instances, running the hierarchy through a "pre-recorded" history. The resulting structure of pattern instances would be derived from this history, and so would be consistent with it. If the inputs/outputs appeared to describe the behavior of a hierarchy containing a "self" then a "self" should result in the hierarchy, except that it would be one with memories, personality and motivations determined by the forced input/output history. The pre-recorded input/output history might be obtained from the running of another AI system, so that some approximation of it could be set up elsewhere.

A hierarchy of the kind being discussed in this series is actually set up for having behavioral patterns and a "self" transcribed into it. Because of the way the hierarchy constructs itself based on inputs/outputs, we can present a fake or pre-recorded input/output history to the system and it will actually construct an appropriate hierarchy for us: If we give the system a pre-recorded history which implies the existence of a "self", the system will actually make the "self" for us.

One use for this could be to make a replica of a human mind in an AI system, just based on records of inputs and outputs: something which I have discussed in a previous article.[16] If we record the input/output history of a human brain, and then force an AI system, of the kind being discussed here, with this, the AI system's hierarchy should set itself up with a model that matches that history, including a "self" with a personality and memories approximating those of the original person. For this to work with a human mind, the hierarchy would need to be capable of modeling one, and the current system may need to be extended. Such a copy would not be identical to the original, but it would be consistent with the original's history of inputs/outputs, which should make it very similar. Whether this could be considered a continuation of the original person to any extent, I will not explore in this article.

We are not restricted to just copying minds into artificial computers, however. If a human brain is a "pure" system of the type being discussed, then this means *it is already set up to have a mind downloaded into it*. This suggests various scenarios as follows.

Copying a human mind into someone else's brain – This would involve capturing the history of inputs/outputs for one human brain, and providing them as fake data to

---

[16] Almond, P., 2003. *Indirect Mind Uploading: Using AI to Avoid Staying Dead*. [Online] paul-almond.com. Available at: http://www.paul-almond.com/IndirectMindUploading.htm [Accessed 24 June 2010].

another human brain, which has previously had minimal or no experiences, or to which some kind of memory erasure process has been applied, so that a hierarchy results that is consistent with the history of the first brain and with a similar "self".

Copying software into a human mind – This would involve capturing the history of inputs/outputs for the software in some artificial computer and providing them as fake data to a human brain, so that a hierarchy results that is consistent with the history of the system. Effectively, something resembling the computer system's software would have been transcribed into a human brain. The system being copied would not necessarily have to behave like a human brain. For example, it could be a program that performs some calculus, or controls a vehicle. Alternatively, the system being copied could be an AI system, so that something resembling an AI system is transcribed into a human brain.

The extent to which you could actually do any of this is uncertain: It depends on how much the human brain is like the hierarchical system described in this series of articles.

Some of the above processes would be potentially ethically problematic.[17] People would disagree about what they imply, but some people would view them as equivalent to murder.

---

[17] Some readers will say that that is an understatement.

4 Conclusion

In the cognitive model developed in this series of articles, the entire structure of the hierarchy is put together to be consistent with the history of inputs/outputs, and this includes the "self" generated in the hierarchy, which is only a part of the model from which predictions and behavior are generated. Therefore, if the information about the system's outputs that is fed into the bottom-level pattern instances could be manipulated at the point where it is inserted into the hierarchy's bottom level, it may be possible to change someone's personality and behavior. This would include altering the "self", as that is merely a part of the model in the hierarchy.

This might be viewed in terms of the hierarchy producing a narrative from the history of inputs/outputs, and that narrative including a description of what the "self" is thinking, feeling and intending. Manipulation of the system's records of its own outputs can alter the narrative, and that includes the narrative of what is happening with the "self".

A process like this would have some limitations. It is based on the idea that the human brain works exactly as described in this series of articles. In reality, the human brain's functioning may be described by a general theory like this, but there may also be a lot of special-case processing. The hierarchy would tend to produce the simplest model that was consistent with the data, and it is possible that the model produced would be one in which the fake behavior would be recognized by the brain as not being its "true" behavior. The risk of this increases if the subject knows that his/her mind is being modified. The rate at which such a process is performed may be an issue: A sudden, apparent change in behavior might be more easily explained in terms of the "new" behavior being fake rather than resulting from a real personality change: For this reason, a process like this may work better if the apparent behavior fed into the hierarchy is changed slowly. Furthermore, if the behavior is changed too rapidly, the inputs would be noticeably inconsistent with the outputs, unless they are faked too. The new model has to be consistent with the existing history of inputs/outputs and the new data that is being manipulated. If the existing history is significant, it may be harder to explain a change in behavior. For this reason, the modification process may be more likely to work better on the mind of a younger subject than an older one. Despite these limitations, we might reasonably expect some change in personality to be caused by the process. The process could be more effective if drugs or memory erasure were used, or if positive of negative reinforcement were used to sustain the required behavior after the subject's mind has been modified.

A less invasive version of this process might be used, in which the outputs being fed into the hierarchy are not tampered with, but in which an attempt is instead made to trick the brain into thinking it has made certain outputs by manipulating the environment to create false evidence of those outputs being made. It is far from certain that this would work. This less invasive process could also be combined with the process of

manipulating outputs that are fed into the hierarchy, so that the outputs being fed into the bottom-level of the hierarchy are manipulated and inputs are manipulated as well.

More extreme processes involving copying entire minds, or copying other systems into brains, have also been discussed. If the human brain works exactly as described in this series, it should be possible to record the inputs/outputs for a human brain and use them to set up another brain with a similar mind. Another possibility is to transcribe the behavioral history of some computing system into a human brain, effectively programming it with the behavior of this system. Actions like these would present ethical issues. We should also be cautious about the extent to which they would work: They are based, again, on the idea that the human brain has no processing outside the scope of this series.

5 Bibliography

Almond, P., 2003. *Indirect Mind Uploading: Using AI to Avoid Staying Dead*. [Online] paul-almond.com. Available at: http://www.paul-almond.com/IndirectMindUploading.htm [Accessed 24 June 2010].

Almond, P., 2010. *An Attempt to Generalize AI - Part 1: The Modeling System*. [Online] paul-almond.com. Available at: http://www.paul-almond.com/AI01.pdf or http://www.paul-almond.com/AI01.doc [Accessed 24 June 2010].

Almond, P., 2010. *An Attempt to Generalize AI - Part 2: Planning and Actions*. [Online] paul-almond.com. Available at: http://www.paul-almond.com/AI02.pdf or http://www.paul-almond.com/AI02.doc [Accessed 24 June 2010].

Almond, P., 2010. *An Attempt to Generalize AI - Part 3: Forgetting*. [Online] paul-almond.com. Available at: http://www.paul-almond.com/AI03.pdf or http://www.paul-almond.com/AI03.doc [Accessed 24 June 2010].

Almond, P., 2010. An Attempt to Generalize AI - Part 4: Modeling Efficiency. [Online] paul-almond.com. Available at: http://www.paul-almond.com/AI04.pdf or http://www.paul-almond.com/AI04.doc [Accessed 24 June 2010].

Almond, P., 2010. *An Attempt to Generalize AI - Part 5: A Completely Probabilistic Hierarchy*. [Online] paul-almond.com. Available at: http://www.paul-almond.com/AI05.pdf or http://www.paul-almond.com/AI05.doc [Accessed 24 June 2010].

Almond, P., 2010. *An Attempt to Generalize AI - Part 6: Measuring Relevance*. [Online] paul-almond.com. Available at: http://www.paul-almond.com/AI06.pdf or http://www.paul-almond.com/AI06.doc [Accessed 24 June 2010].

Almond, P., 2010. *An Attempt to Generalize AI - Part 7: A Basic, Exploratory Relevance Process*. [Online] paul-almond.com. Available at: http://www.paul-almond.com/AI07.pdf or http://www.paul-almond.com/AI07.doc [Accessed 24 June 2010].

Almond, P., 2010. *An Attempt to Generalize AI - Part 8: Forgetting as Part of the Exploratory Relevance Process*. [Online] paul-almond.com. Available at: http://www.paul-almond.com/AI08.pdf or http://www.paul-almond.com/AI08.doc [Accessed 24 June 2010].

Almond, P., 2010. *An Attempt to Generalize AI - Part 9: Improving the Exploratory Relevance Process*. [Online] paul-almond.com. Available at: http://www.paul-almond.com/AI09.pdf or http://www.paul-almond.com/AI09.doc [Accessed 24 June 2010].

Almond, P., 2010. *An Attempt to Generalize AI - Part 10: Alternatives for Pattern Instance Construction*. [Online] paul-almond.com. Available at: http://www.paul-almond.com/AI10.pdf or http://www.paul-almond.com/AI10.doc [Accessed 24 June 2010].

Almond, P., 2010. *An Attempt to Generalize AI - Part 11: Explaining Dreaming*. [Online] paul-almond.com. Available at: http://www.paul-almond.com/AI11.pdf or http://www.paul-almond.com/AI11.doc [Accessed 24 June 2010].

Almond, P., 2010. *An Attempt to Generalize AI - Part 12: Pattern Relevance*. [Online] paul-almond.com. Available at: http://www.paul-almond.com/AI12.pdf or http://www.paul-almond.com/AI12.doc [Accessed 24 June 2010].

Almond, P., 2010. *An Attempt to Generalize AI - Part 13: Reflexive Outputs*. [Online] paul-almond.com. Available at: http://www.paul-almond.com/AI13.pdf or http://www.paul-almond.com/AI13.doc [Accessed 24 June 2010].

Metzinger, T., 2003. *Being No One: The Self-Model Theory of Subjectivity*. Cambridge (MA): MIT Press.

Metzinger, T., 2009. *The EGO Tunnel: The Science of the Mind and the Myth of the Self*. New York: Basic Books.