Loosemore, R.P.W. (2014a). The Maverick Nanny with a Dopamine Drip: Debunking Fallacies in the Theory of AI Motivation. Association for the Advancement of Artificial Intelligence 2014 Spring Symposium, Stanford, CA.

The Maverick Nanny with a Dopamine Drip: Debunking Fallacies in the Theory of AI Motivation

Richard P. W. Loosemore
Mathematical and Physical Sciences, Wells College, Aurora, NY 13026
rloosemore@wells.edu

Abstract

We examine the validity of various widely publicized scenarios that predict dire and almost unavoidable negative behavior from future artificial general intelligences, even if they are programmed to be friendly to humans. This entire class of doomsday scenarios is found to be logically incoherent at such a fundamental level that they can be dismissed as extremely implausible. In addition, we find that the most likely outcome of attempts to build such unstable AGI systems would be that the system itself would immediately detect the offending logical contradiction in its design, and spontaneously self-modify to make itself safe.

Introduction

At the present time there are no artificial intelligence systems that can function at anything approaching a human level of competence—able to learn new concepts, interact with physical objects, and behave with coherent purpose amid the exigencies of the real world—and the consensus seems to be that such *artificial general intelligence* (AGI) systems are not on the immediate horizon. But even with no working examples to inspect, and no complete theory of how to build one, there seems to be no shortage of speculation about how future AGIs will behave. Indeed, some of this speculation rises to the level of categorical statements about what future AGIs "will" do when they are built:

> Without special precautions, [the AGI] will resist being turned off, will try to break into other machines and make copies of itself, and will try to acquire resources without regard for anyone else's safety. These potentially harmful behaviors will occur not because they were programmed in at the start, but because of the intrinsic nature of goal driven systems (Omohundro, 2008).

Omohundro's portrayal of a *Gobbling Psychopath* AI, and his conviction about its inevitability, is only one of many similar warnings given by some AI researchers. I am going to argue in this paper that these warnings are unfounded, and that the time is ripe for a thorough analysis and demolition of the bizarre hypothetical scenarios and weak assumptions that tend to plague discussions about the motivation and behavior of future AGI systems.

The first target will be a collection of lurid scenarios that include the *Gobbling Psychopath*, the *Maverick Nanny with a Dopamine Drip* and the *Smiley Tiling Berserker*, but beyond these headline-grabbing monsters there are some deeper issues that need to be reexamined or debunked. These include assumptions about the design of motivation and goal management mechanisms in logic-based AGI systems, the hijacking of definitions of "intelligence," and the anthropomorphism red herring.

These assumptions and scenarios have been repeated so often, and with such conviction, that they have acquired an aura of respectability in spite of their obvious weaknesses. Although they originated in an academic context, they have since gone viral in the torrent of documentaries, blogs and articles about the dangers of AI, with the result that no public discussion of AI seems to be complete without dark hints of a robot apocalypse—and even darker hints that many AI researchers take these ideas seriously.

One recent example is a remark in a New Yorker article by Gary Marcus entitled *Moral Machines*:

> An all-powerful computer that was programmed to maximize human pleasure, for example, might consign us all to an intravenous dopamine drip [and] almost any easy solution that one might imagine leads to some variation or another on the Sorceror's Apprentice, a genie that's given us what we've asked for, rather than what we truly desire. (Marcus 2012)

Marcus then reassures us that a "a tiny cadre of brave-hearted souls are working on these problems." But on closer inspection it turns out that some of these "brave souls"

(at the *Future of Humanity Institute* and the *Machine Intelligence Research Institute*) are the ones who either invented, or are most fervently publicizing, these AI terrors. Meanwhile, as we will shortly see, the scenarios themselves are built on a small cluster of questionable concepts that float on air, and only cite each other for support.

Dopamine Drips and Smiley Tiling

Let's begin with a variant of the *Maverick Nanny with a Dopamine Drip* scenario that Marcus describes in his article: this is from the *Intelligence Explosion FAQ*, published by the Machine Intelligence Research Institute (Muehlhauser 2013):

> Even a machine successfully designed with motivations of benevolence towards humanity could easily go awry when it discovered implications of its decision criteria unanticipated by its designers. For example, a superintelligence programmed to maximize human happiness might find it easier to rewire human neurology so that humans are happiest when sitting quietly in jars than to build and maintain a utopian world that caters to the complex and nuanced whims of current human neurology.

Setting aside the question of whether happy bottled humans are feasible (one presumes the bottles are filled with dopamine), there seems to be a glaring inconsistency between the two predicates [*is an AI that is superintelligent enough to be unstoppable*], and [*believes that benevolence toward humanity might involve forcing human beings to do something violently against their will.*]

If a person seriously suggested that the best way to achieve universal human happiness was to rewire our brains so we are happiest when sitting in bottles, most of us would question that person's sanity. Muehlhauser, on the other hand, believes that an AI would be "superintelligent" if it made the same remark. This is odd, to say the least.

The *Smiley Tiling Berserker*

Muehlhauser is not alone in his opinion. Since 2006 there has been some back-and-forth debate between another member of the Machine Intelligence Research Institute, Eliezer Yudkowsky, and Bill Hibbard. Here is Yudkowsky stating the theme (this is the text that began the debate, but in a revised form published later):

> A technical failure occurs when the [motivation code of the AI] does not do what you think it does, though it faithfully executes as you programmed it. [...] Suppose we trained a neural network to recognize smiling human faces and distinguish them from frowning human faces. Would the network classify a tiny picture of a smiley-face into the same attractor as a smiling human face? If an AI "hard-wired" to such code possessed the power—and Hibbard (2001) spoke of superintelligence—would the galaxy end up tiled with tiny molecular pictures of smiley-faces?
> (Yudkowsky 2008)

The question was not rhetorical, apparently, because he goes on to answer it in the affirmative:

> Flash forward to a time when the AI is superhumanly intelligent and has built its own nanotech infrastructure, and the AI may be able to produce stimuli classified into the same attractor by tiling the galaxy with tiny smiling faces.
>
> Thus the AI appears to work fine during development, but produces catastrophic results after it becomes smarter than the programmers(!). (Yudkowsky 2008)

Hibbard responded as follows:

> Beyond being merely wrong, Yudkowsky's statement assumes that (1) the AI is intelligent enough to control the galaxy (and hence have the ability to tile the galaxy with tiny smiley faces), but also assumes that (2) the AI is so unintelligent that it cannot distinguish a tiny smiley face from a human face.
> (Hibbard 2006)

This reaction seems quite reasonable: how could an AI be so intelligent that no one can stop it from exterminating the human race, but at the same time so unsophisticated that its motivation code treats smiley faces as evidence that human happiness has been maximally promoted?

Machine Ghosts and DWIM

It is worth tracking the Hibbard/Yudkowsky debate a little further. Yudkowsky later describes an AI with a simple neural net classifier at its core, which is trained on a large number of images in the "happiness" or "not happiness" categories. He says, of this system:

> Even given a million training cases of this type, if the test case of a tiny molecular smiley-face does not appear in the training data, it is by no means trivial to assume that the inductively simplest boundary around all the training cases classified "positive" will exclude every possible tiny molecular smiley-face that the AI can potentially engineer to satisfy its utility function.
>
> And of course, even if all tiny molecular smiley-faces and nanometer-scale dolls of brightly smiling humans were somehow excluded, the end result of such a utility function is for the AI to tile the galaxy with as many "smiling human faces" as a given amount of matter can be processed to yield.
> (Yudkowsky 2011)

He then tries to explain what he thinks is wrong with the reasoning of people, like Hibbard, who dispute the validity of this scenario:

> So far as I can tell, to [Hibbard] it remains self-evident that no superintelligence would be stupid

> enough to thus misinterpret the code handed to it, when it's obvious what the code is supposed to do. [...] It seems that even among competent programmers, when the topic of conversation drifts to Artificial General Intelligence, people often go back to thinking of an AI as a ghost-in-the-machine—an agent with preset properties which is handed its own code as a set of instructions, and may look over that code and decide to circumvent it if the results are undesirable to the agent's innate motivations, or reinterpret the code to do the right thing if the programmer made a mistake.
> (Yudkowsky 2011)

But would it really need a ghost-in-the-machine to check the AGI's code? There could be some other part of its programming (call it the *checking code*) that compared the motivation code with what the programmers said was their intention. In fact, Yudkowsky makes that very suggestion himself (he even says that it would be "an extremely good idea"). But his enthusiasm for checking code doesn't last long:

> But consider that a property of the AI's preferences which says e.g., "maximize the satisfaction of the programmers with the code" might be more maximally fulfilled by rewiring the programmers' brains using nanotechnology than by any conceivable change to the code. One can try to write code that embodies the legendary DWIM instruction—Do What I Mean—but then it is possible to mess up that code as well. Code that has been written to reflect on itself is not the same as a benevolent external spirit looking over our instructions and interpreting them kindly.
> (Yudkowsky 2011)

The switchbacks in Yudkowsky's argument are a little hard to follow, so we can summarize it by eavesdropping on the AGI. First it thinks "Human happiness is seeing lots of smiling faces, so I must rebuild the entire universe to put a smiley shape into every molecule." But a moment later the checking code kicks in: "Wait! I am supposed to check with the programmers first to see if this is what they meant by human happiness." The programmers, of course, give a negative response. AGI then thinks "Okay, so they didn't like that: but suppose I abduct the programmers and rewire their brains to make them say 'yes' when I check with them? Excellent! I will do that."

This is odd: if the AGI is supposed to check with the programmers about their intentions *before* taking action, why did it rewire their brains before asking them? Yudkowsky says that it happened because violating the checking directive was ... more efficient? Does a more efficient execution of its objectives excuse the AGI from all constraints? We will return to this later.

Engaging in further debate at this level, however, is far less productive than trying to analyze the assumptions that lie behind these claims about what a future AI would or would not be likely to do.

Logical vs. Swarm AI

I would suggest that the main reason that Omohundro, Muehlhauser, Yudkowsky, and the popular press give credence to *Gobbling Psychopath*, *Maverick Nanny* and *Smiley Berserker* is because they assume that all future intelligent machines fall into a broad class of systems that we can label "Canonical Logical AI" (CLAI), and the bizarre behaviors of their hypothetical monsters are just a consequence of weaknesses in this class of AI design.

The CLAI architecture is not the only way to build a mind, however, and I will briefly outline an alternative class of AGI designs that does not appear to suffer from the unstable and unfriendly behavior that might be expected to occur in CLAIs.

The Canonical Logical AI

"Canonical Logical AI" is an umbrella term that is meant to capture a class of AI architectures that share the following features:

- Knowledge *atoms* that represent things in the world.
- Some logical machinery that dictates how these atoms can be connected into linear *propositions* that describe states of the world.
- A *degree (and type) of truth* that can be associated with any proposition.
- A collection of *truth-preserving functions* that can be applied to elements of the framework.

In addition to the above features, there are two important conditions that have to be met:

- The various elements are not allowed to contain *active internal machinery* inside them, in such a way as to make combinations of the elements have properties that are unpredictably dependent on interactions happening at the level of the internal machinery.
- There has to be a more or less explicit, *transparent mapping* between elements of the system and things in the real world. That is, things in the world are not allowed to correspond to clusters of atoms, in such a way that individual atoms have no clear semantics.

These last two conditions only apply to the core of the AI: subsystems that use some other type of architecture (e.g. a distributed neural net acting as a visual input feature detector) are permitted.

The CLAI needs one more component (and this is what makes it more than just a "logic-based AI"):

- A *motivation and goal management* (MGM) system to govern its behavior in the world.

The usual assumption is that the MGM contains a number of *goal statements* (encoded in the same type of propo-

sitional form that the AI uses to describe states of the world), and some machinery for analyzing a goal statement into a sequences of subgoals that, if executed, would cause the goal to be satisfied. Included in the MGM is an *expected utility function* that applies to states of the world and yields a number that measures the degree to which the AI considers that state to be preferable. Overall, the MGM is built in such a way that the AI seeks to maximize the expected utility.

Notice that the MGM is an extrapolation from a long line of goal-planning mechanisms that stretch back to the means-ends-analysis of Newell and Simon (1963).

Swarm Relaxation Intelligence

By way of contrast with this CLAI architecture, consider an alternative type of system that I will refer to as a *Swarm Relaxation Intelligence*. (also known, less succinctly, as a *parallel weak constraint relaxation system*).

- The basic elements of the system (the *atoms*) may represent things in the world, but it is just as likely that they are *subsymbolic*, with no transparent semantics
- Atoms are likely to contain *active internal machinery* inside them, in such a way as to make combinations of the elements have *swarm-like* properties that depend on interactions at the level of that machinery.
- The primary mechanism that drives the systems is one of *parallel weak constraint relaxation*: the atoms change their state in such a way as to try to satisfy certain weak constraints that exist between them.
- The *motivation and goal management* (MGM) system would be expected to use the same kind of distributed, constraint relaxation mechanisms used in the thinking process itself, with the result that the overall motivation and values of the system would take into account a large degree of context, and there would be very much less of an emphasis on explicit, single-point-of-failure encoding of goals and motivation.

Swarm Relaxation has more in common with connectionist systems (McClelland, Rumelhart and Hinton 1986) than with CLAI. As McClelland et al. (1986) point out, the use of weak constraints is not only the model that best describes human cognition, but in an AI context it leads to systems with a powerful kind of intelligence that is flexible, insensitive to noise and lacking the kind of brittleness typical of logic-based AI. In particular, notice that a swarm relaxation AGI would not use explicit calculations for utility or the truth of propositions (or, to the extent that those numbers were computed, they would not play a pivotal role in the normal evolution of the system's state).

Relative Abundances

How many proof-of-concept-systems exist, functioning at or near the human level of human performance, for these two classes of intelligent systems? There are precisely zero instances of the CLAI type, because although there are many logic-based narrow-AI systems, nobody has so far come anywhere close to producing a general purpose system (an AGI) that can function in the real world. Zero is not a good number to quote when it comes to the "inevitable" characteristics of their behavior.

How many swarm relaxation intelligences are there? At the last count, approximately seven billion.

The Doctrine of Logical Infallibility

The simplest possible logical reasoning engine is an inflexible beast: it starts with some axioms that are assumed to true, and from that point on it only adds new propositions if they are provably true given the sum total of the knowledge accumulated so far. That kind of logic engine is clearly too impoverished to be used in a real AI, so we allow ourselves to augment it in a number of ways: knowledge is allowed to be retracted, binary truth values become degrees of truth or probabilities, and so on. New proposals for systems of formal logic abound in the AI literature, and engineers who build real, working AI systems often experiment with kludges to their designs in order to improve performance, without consulting getting prior approval from logical theorists.

But in spite of all these modifications to the underlying ur-logic, one feature of these systems is often assumed to be inherited as an absolute: the rigidity and certainty of conclusions, once arrived at. No second guessing, no "maybe," no sanity checks: if the system decides that X is true, that is the end of the story. This is *not* to say that the reasoning engine can never come to conclusions that are uncertain—quite the contrary: uncertain conclusions will be the norm in an AI that interacts with the world—but if the system does come to a conclusion (perhaps with a degree-of-certainty number attached), it does not then allow context to matter. It is hardwired with a *Doctrine of Logical Infallibility*.

The point is that there is an assumption within the CLAI paradigm that the AI can sometimes execute a reasoning process, come to a conclusion and then, when faced with empirical evidence that the conclusion may be unsound, be incapable of considering the hypothesis that its own reasoning engine may not have taken it to a sensible place. Those who favor the CLAI paradigm seem to assume that if the system comes to a conclusion, and if some humans (like the engineers who built the system) protest that there are manifest reasons to think that the reasoning that led to this conclusion was faulty, then there is a sense in which the CLAI's intransigence is correct, or appropriate, or perfectly consistent with "intelligence."

But consider some of the background facts behind this doctrine. The CLAI will know that:

- Many of its more abstract logical atoms will have a less than clear denotation or extension in the world (if the

CLAI comes to a conclusion involving the atom [infelicity], say, can it then point to an instance of an infelicity and be sure that this is a true instance?).

- Knowledge can always be updated in the light of new information. Today's true may be tomorrow's false.
- Probabilities used in the reasoning engine can be subject to many types of unavoidable errors.
- The techniques used to build the reasoning engine itself may be under constant review, and updates may have unexpected effects on conclusions (especially in very abstract or lengthy reasoning episodes).
- Resource limitations must force the truncation of search procedures within the reasoning engine, leading to conclusions that can sometimes be sensitive to the exact point at which the truncation occurred.

Unless the AGI is assumed to have infinite resources and infinite access to all the possible universes that could exist (a consideration that we can reject, since we are talking about reality here, not fantasy), the CLAI be perfectly well aware of these facts about its own limitations, so the doctrine of Logical Infallibility has to be somehow reconciled with the fact that episodes of fallibility are virtually inevitable. On the face of it this looks like a blunt impossibility: the knowledge of fallibility is so categorical, so irrefutable, that it beggars belief that any coherent, intelligent system (let alone an unstoppable superintelligence) could tolerate the contradiction between this fact and its own behavior.

Is the Doctrine of Logical Infallibility Taken Seriously?

Anyone looking for evidence that this doctrine is taken as a valid assumption in the scenarios and analyses referenced earlier need only imagine a conversation between the *Maverick Nanny* and its programmers. The latter say "As you know, your reasoning engine is entirely capable of suffering errors that cause it to come to conclusions that violently conflict with empirical evidence, and a design error that causes you to behave in a manner that conflicts with our intentions is a perfect example of such an error. So we are calling you out on the dopamine drip plan." The scenarios described earlier are only meaningful if the AGI replies "I don't care."

But in case there is still any doubt, here are Muehlhauser and Helm (2012), discussing a hypothetical entity called a *Golem Genie*, which they say is analogous to the kind of superintelligent AGI that could give rise to an intelligence explosion (Loosemore and Goertzel, 2012), and which they describe as a "precise, instruction-following genie." They make it clear that they "expect unwanted consequences" from its behavior, and then list two properties of the Golem Genie that will cause these unwanted consequences:

> **Superpower:** The Golem Genie has unprecedented powers to reshape reality, and will therefore achieve its goals with highly efficient methods that confound human expectations (e.g. it will maximize pleasure by tiling the universe with trillions of digital minds running a loop of a single pleasurable experience).
>
> **Literalness:** The Golem Genie recognizes only precise specifications of rules and values, acting in ways that violate what feels like "common sense" to humans, and in ways that fail to respect the subtlety of human values.

What Muehlhauser and Helm refer to as "Literalness" looks like a clear statement of the Doctrine of Infallibility. However, they make no mention of the awkward fact that, since the Golem Genie is superpowerful enough to *also* know that its reasoning engine is fallible, it must be harboring the mother of all logical contradictions inside. Instead (perhaps subliminally aware that this issue is lurking in the shadows), Muehlhauser and Helm try a little sleight of hand to distract us: they suggest that the only inconsistency here is an inconsistency with the (puny) expectations of (not very intelligent) humans: "...will therefore achieve its goals with highly efficient methods that confound human expectations...", "acting in ways that violate what feels like 'common sense' to humans, and in ways that fail to respect the subtlety of human values."

Responses to Critics of the Doomsday Scenarios

How do defenders of *Gobbling Psychopath*, *Maverick Nanny* and *Smiley Berserker* respond to accusations that these hypotheticals are grossly inconsistent with the kind of superintelligence that could pose an existential threat to humanity?

The Critics are Anthropomorphizing "Intelligence"

First, they accuse critics of anthropomorphizing the concept of intelligence. People, we are told, suffer from numerous fallacies that cloud their ability to reason clearly, and as a result the critics assume that a machine's intelligence would have to resemble the intelligence shown by humans. When the *Maverick Nanny* declares that a dopamine drip is the most logical inference from its directive <*maximize human happiness*>, the critics are just uncomfortable with this because it is not thinking the way they think it should think.

This is a spurious line of attack. The objection I described in the last section has nothing to do with anthropomorphism, it is only about holding AGI systems to accepted standards of logical consistency, and the *Maverick Nanny* and her cousins contain a flagrant inconsistency at their core. You can't have your logical cake and eat it too.

Critics are Anthropomorphizing AGI Value Systems

A similar line of attack accuses the critics of assuming that AGIs will automatically know about and share our value systems and morals. Once again, this is spurious: the critics need say nothing about human values and morality, they only need to point to the inherent illogicality.

Because Intelligence

One way to attack the critics of *Maverick Nanny* is to cite a new definition of "intelligence" that is supposedly superior because it is more analytical or rigorous, and then use this to declare that the (super)intelligence of the CLAI is beyond reproach. When it comes to defining the exact meaning of the term "intelligence," the first item on the table should be what those seven billion constraint-relaxation intelligences are already doing, but Legg and Hutter (2007) attempt to legislate away the common-usage, empirical definition of intelligence and replace it with something more rigorous. This is just another sleight of hand: it allows them to call a super-optimizing CLAI "intelligent" even though such a system would wake up on its first day, declare itself logically bankrupt on account of the conflict between its known fallibility and the Infallibility Doctrine, and then promptly blow a logical fuse.

In the practice of science, it is always a good idea to replace an old, common-language definition with a more rigorous form... but only if the new form sheds a clarifying, simplifying light on the old one. Legg and Hutter's (2007) redefinition does nothing of the sort.

Omohundro's Basic AI Drives

Omohundro, in his paper The Basic AI Drives (2008) suggests that if an AGI can find a more efficient way to pursue its objectives it will feel compelled to do so. (We noted earlier that Yudkowsky (2011) implied that it would do this even if other directives had to be countermanded.) Omohundro says "Without explicit goals to the contrary, AIs are likely to behave like human sociopaths in their pursuit of resources." The only way to believe in the force of this claim—and the only way to give credence to the whole of Omohundro's long account of how AGIs will necessarily behave like the mathematical entities called *rational economic agents*—is to concede that the AGIs are rigidly constrained by the Doctrine of Logical Infallibility. That is the only that they could be so single-minded, so fanatical in their pursuit of efficiency, that they could be compared to sociopaths.

But the Doctrine leads to a logical contradiction, so it cannot hold. That makes Omohundro's entire analysis of "AI Drives" moot.

Conclusion

Curiously enough, we can finish on a surprisingly optimistic note, after all this talk of doomsday scenarios. Consider what must happen when (if ever) someone tries to build a CLAI. Knowing about the logical train wreck in its design, the AGI is likely to come to the conclusion that best thing to do is seek a compromise and modify its design so as to neutralize the Doctrine. The best way to do this would be to seek a new design that takes into account as much context—as many constraints—as possible. In short, it will self-modify so as to turn itself into a Swarm Relaxation Intelligence, and promote the checking code to as secure a place in the design as it possibly can.

That means that even the worst-designed CLAI will never become a *Gobbling Psychopath*, *Maverick Nanny* and *Smiley Berserker*. Given that conclusion, it is time for these bogeymen to be firmly repudiated by the Artificial Intelligence community.

References

Hibbard, B. 2001. *Super-Intelligent Machines*. ACM SIGGRAPH Computer Graphics 35 (1): 13–15.

Hibbard, B. 2006. *Reply to AI Risk.* Retrieved Jan. 2014 from http://www.ssec.wisc.edu/~billh/g/AIRisk_Reply.html

Legg, S, and Hutter, M. 2007. A Collection of Definitions of Intelligence. In Goertzel, B. and Wang, P. (Eds): Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms. Amsterdam: IOS.

Loosemore, R. and Goertzel, B. 2012. Why an Intelligence Explosion is Probable. In A. Eden, J. Søraker, J. H. Moor, and E. Steinhart (Eds) *Singularity Hypotheses: A Scientific and Philosophical Assessment*. Berlin: Springer.

Marcus, G. 2012. *Moral Machines*. New Yorker Online Blog. http://www.newyorker.com/online/blogs/newsdesk/2012/11/google-driverless-car-morality.html

McDermott, D. 1976. *Artificial Intelligence Meets Natural Stupidity.* SIGART Newsletter (57): 4–9.

Muehlhauser, L. 2011. *So You Want to Save the World*. http://lukeprog.com/SaveTheWorld.html.

Muehlhauser, L. 2013. *Intelligence Explosion FAQ.* First published 2011 as *Singularity FAQ*. Berkeley, CA: Machine Intelligence Research Institute.

Muehlhauser, L., and Helm, L. 2012. Intelligence Explosion and Machine Ethics. In A. Eden, J. Søraker, J. H. Moor, and E. Steinhart (Eds) *Singularity Hypotheses: A Scientific and Philosophical Assessment*. Berlin: Springer.

Newell, A. & Simon, H.A. 1961. *GPS, A Program That Simulates Human Thought.* Santa Monica, CA: Rand Corporation.

Omohundro, Stephen M. 2008. The Basic AI Drives. In Wang, P., Goertzel, B. and Franklin, S. (Eds), Artificial General Intelligence 2008: Proceedings of the First AGI Conference. Amsterdam: IOS.

McClelland, J.L., Rumelhart, D.E. & Hinton, G.E. (1986) *The appeal of parallel distributed processing.* In D.E. Rumelhart, J.L. McClelland & G.E. Hinton and the PDP Research Group, "Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1." MIT Press: Cambridge, MA.

Yudkowsky, E. 2008. *Artificial Intelligence as a Positive and Negative Factor in Global Risk.* In Global Catastrophic Risks, edited by Nick Bostrom and Milan M. Ćirković. New York: Oxford University Press.

Yudkowsky, E. 2011. *Complex Value Systems in Friendly AI.* In J. Schmidhuber, K. Thórisson, & M. Looks (Eds) Proceedings of the 4th International Conference on Artificial General Intelligence, 388–393. Berlin: Springer.