Intelligence Explosion FAQ

Luke Muehlhauser
Machine Intelligence Research Institute

Abstract

The Machine Intelligence Research Institute is one of the leading research institutes on intelligence explosion. Below are short answers to common questions we receive.

Muehlhauser, Luke. 2013. "Intelligence Explosion FAQ." First published 2011 as "Singularity FAQ." Machine Intelligence Research Institute, Berkeley, CA

Contents

1. Basics

1.1. What is an intelligence explosion?

The intelligence explosion idea was expressed by statistician I. J. Good in 1965:

> Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an "intelligence explosion," and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make. (Good 1965)

The argument is this: Every year, computers surpass human abilities in new ways. A program written in 1956 was able to prove mathematical theorems and found a more elegant proof for one of them than Russell and Whitehead had given in *Principia Mathematica* (Mackenzie 1995). By the late 1990s, "expert systems" had surpassed human skill for a wide range of tasks. In 1997, IBM's Deep Blue computer beat the world chess champion (Nilsson 2009), and in 2011, IBM's Watson beat the best human players at a much more complicated game: *Jeopardy!* (Markoff 2011). Recently a robot named Adam was programmed with our scientific knowledge about yeast, then posed its own hypotheses, tested them, and assessed the results (King et al. 2009; King 2011).

Computers remain far short of human intelligence, but resources that aid AI design are accumulating (including hardware, large datasets, neuroscientific knowledge, and AI theory). We may one day design a machine that surpasses human skill *at designing AIs*. After that, this machine could improve its own intelligence faster and better than humans could, which would make it even *more* skilled at improving its own intelligence. This could continue in a positive feedback loop such that the machine quickly becomes vastly more intelligent than the smartest human being on Earth: an "intelligence explosion" resulting in a machine superintelligence.

Further readings:

- "The Coming Technological Singularity" (Vinge 1993)
- "Technological Singularity" (*Wikipedia*)
- "The Singularity: A Philosophical Analysis" (Chalmers 2010)

2. How Likely Is an Intelligence Explosion?

2.1. How is "intelligence" defined?

Artificial intelligence researcher Shane Legg defines intelligence like this:

> Intelligence measures an agent's ability to achieve goals in a wide range of environments. (Legg 2008)

This is a bit vague, but it will serve as the working definition of "intelligence" for this FAQ.

Further readings:

- "Intelligence" (*Wikipedia*)
- "Intelligence: Knowns and Unknowns" (Neisser et al. 1996)
- *Comparative Cognition: Experimental Explorations of Animal Intelligence* (Wasserman and Zentall 2006)

2.2. What is greater-than-human intelligence?

Machines are already smarter than humans are at many specific tasks: performing calculations, playing chess, searching large data banks, detecting underwater mines, and more (Nilsson 2009). But one thing that makes humans special is their *general* intelligence. Humans can intelligently adapt to radically new problems in the urban jungle or outer space for which evolution could not have prepared them. Humans can solve problems for which their brain hardware and software were never trained. Humans can even examine the processes that produce their own intelligence (cognitive neuroscience) and design new kinds of intelligence never seen before (artificial intelligence).

To possess greater-than-human intelligence, a machine must be able to achieve goals more effectively than humans can, in a wider range of environments than humans can. This kind of intelligence involves the capacity not just to do science and play chess but also to manipulate the social environment.

Computer scientist Marcus Hutter has described a formal model called AIXI that he says possesses the greatest general intelligence possible (Hutter 2005). But to implement it would require more computing power than all the matter in the universe can provide. Several projects try to approximate AIXI while still being computable—for example, MC-AIXI (Veness et al. 2011).

Still, there remains much work to be done before greater-than-human intelligence can be achieved in machines. Greater-than-human intelligence need not be achieved by

directly programming a machine to be intelligent. It could also be achieved by whole-brain emulation, by biological cognitive enhancement, or by brain-computer interfaces (see below).

Further readings:

- *Artificial General Intelligence* (Goertzel and Pennachin 2007)
- *Whole Brain Emulation: A Roadmap* (Sandberg and Bostrom 2008)
- "Cognitive Enhancement: Methods, Ethics, Regulatory Challenges" (Bostrom and Sandberg 2009)
- "Brain-Computer Interface" (*Wikipedia*)

2.3. What is whole-brain emulation?

Whole-brain emulation (WBE), or "mind uploading," consists of computer emulation of all the cells and connections in a human brain. Even if the underlying principles of general intelligence prove difficult to discover, we might still emulate an entire human brain and make it run at a million times its normal speed (computer circuits communicate *much* faster than neurons do). So this would not lead immediately to smarter-than-human intelligence, but it would lead to faster-than-human intelligence. A WBE could be backed up (leading to a kind of immortality), and it could be copied so that hundreds or millions of WBEs could work on separate problems in parallel. If WBEs are created, they may therefore be able to solve scientific problems far more rapidly than ordinary humans, accelerating further technological progress.

Further readings:

- *Whole Brain Emulation: A Roadmap* (Sandberg and Bostrom 2008)
- The Blue Brain Project

2.4. What is biological cognitive enhancement?

There may be genes or molecules that can be modified to improve general intelligence. Researchers have already done this in mice: they overexpressed the NR2B gene, which improved those mice's memory beyond that of any other mice of any species (Tang et al. 1999). Biological cognitive enhancement in humans may cause an intelligence explosion to occur more quickly than it otherwise would.

Further reading:

- "Cognitive Enhancement: Methods, Ethics, Regulatory Challenges" (Bostrom and Sandberg 2009)

2.5. What are brain-computer interfaces?

A brain-computer interface (BCI) is a direct communication pathway between the brain and a computer device. BCI research is heavily funded and has already met with dozens of successes. Three successes in human BCIs are a device that restores (partial) sight to the blind, cochlear implants that restore hearing to the deaf, and a device that allows use of an artificial hand by direct thought (Hochberg et al. 2006).

Such devices restore impaired functions, but many researchers also expect to augment and improve normal human abilities with BCIs. Ed Boyden is researching these opportunities as the lead of the Synthetic Neurobiology Group at MIT. Such devices might hasten the arrival of the intelligence explosion, if only by improving human intelligence so that the hard problems of AI can be solved more rapidly.

Further reading:

- "Brain-Computer Interface" (*Wikipedia*)

2.6. How could general intelligence be programmed into a machine?

There are many paths to artificial general intelligence. One path is to imitate the human brain by using neural nets or evolutionary algorithms to build dozens of separate components which can then be pieced together (Martinetz and Schulten 1991; Grossberg 1992; de Garis 2007). Another path is to start with a formal model of perfect general intelligence and try to approximate that (Hutter 2007; Schmidhuber 2007). A third path is to focus on developing a "seed AI" that can recursively self-improve, such that it can learn to be intelligent on its own without needing to first achieve human-level general intelligence (Yudkowsky 2007a). Eurisko is a self-improving AI in a limited domain but is not able to achieve human-level general intelligence.

Further reading:

- "Contemporary Approaches to Artificial General Intelligence" (Pennachin and Goertzel 2007) in *Artificial General Intelligence* (Goertzel and Pennachin 2007)

2.7. What is superintelligence?

Nick Bostrom defined "superintelligence" as

> an intellect that is much smarter than the best human brains in practically every field, including scientific creativity, general wisdom and social skills. (Bostrom 1998)

This definition includes vague terms like "much" and "practically," but it will serve as the working definition for superintelligence in this FAQ. An intelligence explosion would lead to machine superintelligence, and some believe that an intelligence explosion is the most likely path to superintelligence.

Further reading:

- "How Long Before Superintelligence?" (Bostrom 1998)
- *Machine Super Intelligence* (Legg 2008)

2.8. When will the intelligence explosion happen?

Predicting the future is risky business. There are many philosophical, scientific, technological, and social uncertainties relevant to the arrival of the intelligence explosion. Because of this, experts disagree on when the intelligence explosion will occur. Here are some of their predictions:

- Futurist Ray Kurzweil predicts that machines will reach human-level intelligence by 2030 and that we will reach "a profound and disruptive transformation in human capability" by 2045 (Kurzweil 2005).
- Intel's chief technology officer, Justin Rattner, expects "a point when human and artificial intelligence [merge] to create something bigger than itself" by 2048 (Rattner 2008).
- AI researcher Eliezer Yudkowsky expects the intelligence explosion by 2060 (Yudkowsky 2011).
- Philosopher David Chalmers has over 50% credence in the intelligence explosion occurring by 2100 (Chalmers 2010).
- Quantum computing expert Michael Nielsen estimates that the probability of the intelligence explosion occurring by 2100 is between 0.2% and about 70% (Nielsen 2011).
- In 2009, at the AGI-09 conference, experts were asked when AI might reach superintelligence with massive new funding. The median estimates were that machine superintelligence could be achieved by 2045 (with 50% confidence) or by 2100

(with 90% confidence). Of course, attendees to this conference were self-selected to think that near-term artificial general intelligence is plausible (Baum, Goertzel, and Goertzel 2011).

- iRobot CEO Rodney Brooks and cognitive scientist Douglas Hofstadter allow that the intelligence explosion may occur in the future, but probably not in the twenty-first century.
- Roboticist Hans Moravec predicts that AI will surpass human intelligence "well before 2050" (Moravec 1999).
- In a 2005 survey of twenty-six contributors to a series of reports on emerging technologies, the median estimate for machines reaching human-level intelligence was 2085 (Bainbridge 2006).
- Participants in a 2011 intelligence conference at Oxford gave a median estimate of 2050 for when there will be a 50% of human-level machine intelligence, and a median estimate of 2150 for when there will be a 90% chance of human-level machine intelligence (Sandberg and Bostrom 2011).
- On the other hand, 41% of the participants in the AI50 conference (in 2006) stated that machine intelligence would *never* reach the human level (Goertzel, Baum, and Goertzel 2010).

Further reading:

- "How Long Until Human-Level AI? Results from an Expert Assessment" (Baum, Goertzel, and Goertzel 2011)

2.9. Might an intelligence explosion never occur?

Dreyfus (1972) and Penrose (1994) have argued that human cognitive abilities can't be emulated by a computational machine. Searle (1980) and Block (1981) argue that certain kinds of machines cannot have a mind (consciousness, intentionality, etc.). But these objections need not concern those who predict an intelligence explosion (Chalmers 2010).

We can reply to Dreyfus and Penrose by noting that the intelligence explosion idea does not require an AI to be a classical computational system. And we can reply to Searle and Block by noting that the intelligence explosion does not depend on machines having consciousness or other properties of "mind," only that they be able to solve problems better than humans can in a wide variety of unpredictable environments. As Edsger

Dijkstra once said, the question of whether a machine can *really* think is "no more interesting than the question of whether a submarine can swim."

Others who are pessimistic about the intelligence explosion occurring within the next few centuries don't have a specific objection but instead think there are hidden obstacles that will reveal themselves and slow or halt progress toward machine superintelligence (Baum, Goertzel, and Goertzel 2011).

Finally, a global catastrophe like nuclear war or a large asteroid impact could so damage human civilization that the intelligence explosion never occurs. Or a stable and global totalitarianism could prevent the technological development required for an intelligence explosion to occur (Caplan 2008).

3. Consequences of an Intelligence Explosion

3.1. Why would great intelligence produce great power?

Intelligence is powerful (Yudkowsky 2007c; Legg 2008). One might say that intelligence is no match for a gun, or for someone with lots of money, but both guns and money were produced by intelligence. If not for our intelligence, humans would still be foraging the savannah for food.

Intelligence is what caused humans to dominate the planet in the blink of an eye (on evolutionary timescales). Intelligence is what allows us to eradicate diseases, and what gives us the potential to eradicate ourselves with nuclear war. Intelligence gives us superior strategic skills, superior social skills, superior economic productivity, and the power of invention.

A machine with superintelligence would be able to hack into vulnerable networks via the Internet, commandeer those resources for additional computing power, take over mobile machines connected to networks connected to the Internet, use them to build additional machines, perform scientific experiments to understand the world better than humans can, invent quantum computing and nanotechnology, manipulate the social world better than we can, and do whatever it could to give itself more power to achieve its goals—all at a speed much faster than humans could respond to.

3.2. How could an intelligence explosion be useful?

A machine superintelligence, if programmed with the right motivations, could potentially solve all the problems that humans are trying to solve but haven't had the ingenuity or processing speed to solve yet. A superintelligence might cure disabilities and diseases, achieve world peace, give humans vastly longer and healthier lives, eliminate food and energy shortages, boost scientific discovery and space exploration, and so on.

Furthermore, humanity faces several existential risks in the twenty-first century, including global nuclear war, bioweapons, superviruses, and more (Bostrom and Ćirković 2008). A superintelligent machine would be more capable of solving those problems than humans are.

Further reading:

- "Artificial Intelligence as a Positive and Negative Factor in Global Risk" (Yudkowsky 2008)

3.3. How might an intelligence explosion be dangerous?

If programmed with the wrong motivations, a machine could be malevolent toward humans and intentionally exterminate our species. More likely, it could be designed with motivations that initially appear safe (and easy to program) to its designers, but that turn out to be best fulfilled (given sufficient power) by reallocating resources from sustaining human life to other projects (Omohundro 2008). As Yudkowsky writes, "the AI does not hate you, nor does it love you, but you are made out of atoms which it can use for something else."

Since weak AIs with many different motivations could best achieve their goals by faking benevolence until they are powerful, safety testing to avoid this could be very challenging. Alternatively, competitive pressures, both economic and military, might lead AI designers to try to use other methods to control AIs with undesirable motivations. As those AIs became more sophisticated this could eventually lead to one risk too many.

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.

Further readings:

- "Artificial intelligence as a positive and negative factor in global risk" (Yudkowsky 2008)
- "The Singularity: A Philosophical Analysis" (Chalmers 2010)

4. Friendly AI

4.1. What is Friendly AI?

A Friendly Artificial Intelligence (Friendly AI) is an artificial intelligence that is *friendly* to humanity—one that has a good rather than bad effect on humanity.

AI researchers continue to make progress with machines that make their own decisions, and there is a growing awareness that we need to design machines to act safely and ethically. This research program goes by many names: "machine ethics" (McLaren 2005; Powers 2005; Anderson and Anderson 2006, 2011), "machine morality" (Wallach, Allen, and Smit 2008), "artificial morality" (Danielson 1992), "computational ethics" (Allen 2002), "computational metaethics" (Lokhorst 2011), "Friendly AI" (Yudkowsky 2001), and "robo-ethics" or "robot ethics" (Capurro et al. 2006; Sawyer 2007).

The most immediate concern may be in battlefield robots; the US Department of Defense contracted Ronald Arkin to design a system for ensuring ethical behavior in autonomous battlefield robots (Arkin 2009). The US Congress has declared that a third of America's ground systems must be robotic by 2025, and by 2030 the US Air Force plans to have swarms of bird-sized flying robots that operate semi-autonomously for weeks at a time.

But Friendly AI research is not only concerned with battlefield robots or machine ethics in general. It is concerned with a problem of a much larger scale: designing AI that would remain safe and friendly after the intelligence explosion.

A machine superintelligence would be enormously powerful. Successful implementation of Friendly AI could mean the difference between a solar system of unprecedented happiness and a solar system in which all available matter has been converted into parts for achieving the superintelligence's goals.

It must be noted that Friendly AI is a harder project than often supposed. As explored below, commonly suggested solutions for Friendly AI are likely to fail because of two features possessed by any superintelligence:

1. *Superpower*: A superintelligent machine will have unprecedented powers to reshape reality, and therefore will achieve its goals with highly efficient methods that confound human expectations and desires.

2. *Literalness*: A superintelligent machine will make decisions based on the mechanisms it is designed with, not the hopes its designers had in mind when they programmed those mechanisms. It will act only on precise specifications of rules and values, and will do so in ways that need not respect the complexity and subtlety (Schroeder 2004; Yudkowsky 2007b; Kringelbach and Berridge 2009) of what humans value. A demand like "maximize human happiness" sounds simple to us

because it contains few words, but philosophers and scientists have failed for centuries to explain *exactly* what this means and certainly have not translated it into a form sufficiently rigorous for AI programmers to use.

Further readings:

- "Friendly Artificial Intelligence" (*Wikipedia*)
- "The Singularity: Humanity's Last Invention?" (Kaste 2011)
- "What is Friendly AI?" (SIAI)
- "A Review of Proposals toward Safe AI" (Fox 2011)
- "Friendly AI: A Bibliography" (Muehlhauser 2011)

4.2. What can we expect the motivations of a superintelligent machine to be?

Except in the case of whole-brain emulation, there is no reason to expect a superintelligent machine to have motivations anything like those of humans. Human minds represent a tiny dot in the vast space of all possible mind designs, and very different kinds of minds are unlikely to share to complex motivations unique to humans and other mammals.

Whatever its goals, a superintelligence would tend to commandeer resources that could help it achieve its goals, including the energy and elements on which human life depends. It would not stop because of a concern for humans or other intelligences that is built into all possible mind designs. Rather, it would pursue its particular goal and give no thought to concerns that seem natural to that particular species of primate called *Homo sapiens*.

There are, however, some basic instrumental motivations we can expect superintelligent machines to display, because they are useful for achieving its goals, no matter what its goals are. For example, an AI will "want" to self-improve, to be optimally rational, to retain its original goals, to acquire resources, and to protect itself—because all these things help it achieve the goals with which it was originally programmed.

Further readings:

- "The Basic AI Drives" (Omohundro 2008) in *Artificial General Intelligence 2008: Proceedings of the First AGI Conference* (Wang, Goertzel, and Franklin 2008)
- *Basic AI Drives and Catastrophic Risks* (Shulman 2010)

4.3. Can't we just keep the superintelligence in a box, with no access to the Internet?

"AI-boxing" is a common suggestion: why not use a superintelligent machine as a kind of question-answering oracle and never give it access to the Internet or any motors with which to move itself and acquire resources beyond what we give it? There are several reasons to suspect that AI-boxing will not work in the long run:

1. Whatever goals the creators designed the superintelligence to achieve, it will be more able to achieve those goals if given access to the Internet and other means of acquiring additional resources. So there will be tremendous temptation to let the AI out of its box.

2. Preliminary experiments in AI-boxing do not inspire confidence (Yudkowsky 2002). And a superintelligence will generate far more persuasive techniques than we can imagine for getting humans to let it out of the box.

3. If one superintelligence has been created, then other labs or even independent programmers will be only weeks or decades away from creating a second superintelligence, and then a third, and then a fourth. You cannot hope to successfully contain all superintelligences created around the world by hundreds of people for hundreds of different purposes.

4.4. Can't we just program the superintelligence not to harm us?

Science fiction author Isaac Asimov told stories about robots programmed with the Three Laws of Robotics (Asimov 1942): (1) a robot may not injure a human being or, through inaction, allow a human being to come to harm; (2) a robot must obey any orders given to it by human beings, except where such orders would conflict with the First Law; and (3) a robot must protect its own existence as long as such protection does not conflict with the First or Second Law. But Asimov's stories tended to illustrate why such rules would go wrong (Anderson 2008).

Still, could we program constraints into a superintelligence that would keep it from harming us? Probably not.

One approach would be to implement constraints as rules or mechanisms that prevent a machine from taking actions that it would normally take to fulfill its goals: perhaps "filters" that intercept and cancel harmful actions, or "censors" that detect and suppress potentially harmful plans within a superintelligence.

Constraints of this kind, no matter how elaborate, are nearly certain to fail for a simple reason: they pit human design skills against superintelligence. A superintelligence would correctly see these constraints as obstacles to the achievement of its goals and would do everything in its power to remove or circumvent them. Perhaps it would

delete the section of its source code that contains the constraint. If we were to block this by adding another constraint, it could create new machines that didn't have the constraint written into them, or fool us into removing the constraints ourselves. Further constraints may seem impenetrable to humans, but would likely be defeated by a superintelligence. Counting on humans to out-think a superintelligence is not a viable solution.

If constraints *on top of* goals are not feasible, could we put constraints *inside of* goals? If a superintelligence had a goal of avoiding harm to humans, it would not be motivated to remove this constraint, avoiding the problem we pointed out above. Unfortunately, the intuitive notion of "harm" is very difficult to specify in a way that doesn't lead to very bad results when used by a superintelligence. If "harm" is defined in terms of human pain, a superintelligence could rewire humans so that they don't feel pain. If "harm" is defined in terms of thwarting human desires, it could rewire human desires. And so on.

If, instead of trying to fully specify a term like "harm," we decide to explicitly list all of the actions a superintelligence ought to avoid, we run into a related problem: human value is complex and subtle, and it's unlikely we can come up with a list of all the things we *don't* want a superintelligence to do. This would be like writing a recipe for a cake that reads: "Don't use avocados. Don't use a toaster. Don't use vegetables . . ." and so on. Such a list can never be long enough.

4.5. Can we program the superintelligence to maximize human pleasure or desire satisfaction?

Let's consider the likely consequences of some utilitarian designs for Friendly AI.

An AI designed to minimize human suffering might simply kill all humans: no humans, no human suffering (Smart 1958; Russell and Norvig 2010).

Or consider an AI designed to maximize human pleasure. Rather than build an ambitious utopia that caters to the complex and demanding wants of humanity for billions of years, it could achieve its goal more efficiently by wiring humans into Nozick's experience machines. Or it could rewire the "liking" component of the brain's reward system so that whichever hedonic hotspot (Smith et al. 2009) paints sensations with a "pleasure gloss" (Aldridge and Berridge 2009; Frijda 2009) is wired to maximize pleasure when humans sit in jars. That would be an easier world for the AI to build than one that caters to the complex and nuanced set of world states currently painted with the pleasure gloss by most human brains.

Likewise, an AI motivated to maximize objective desire satisfaction or reported subjective well-being could rewire human neurology so that both ends are realized whenever humans sit in jars. Or it could kill all humans (and animals) and replace them with beings made from scratch to attain objective desire satisfaction or subjective well-being

when sitting in jars. Either option might be easier for the AI to achieve than maintaining a utopian society catering to the complexity of human (and animal) desires. Similar problems afflict other utilitarian AI designs.

It's not just a problem of specifying goals, either. It is hard to predict how goals will change in a self-modifying agent. No current mathematical decision theory can process the decisions of a self-modifying agent.

So, while it may be *possible* to design a superintelligence that would do what we want, it's harder than one might initially think.

4.6. Can we teach a superintelligence a moral code with machine learning?

Some have proposed (Rzepka and Araki 2005; Anderson, Anderson, and Armen 2005b; Guarini 2006; Honarvar and Ghasem-Aghaee 2009) that we teach machines a moral code with case-based machine learning. The basic idea is this: Human judges would rate thousands of actions, character traits, desires, laws, or institutions as having varying degrees of moral acceptability. The machine would then find the connections between these cases and *learn* the principles behind morality, such that it could apply those principles to determine the morality of new cases not encountered during its training. This kind of machine learning has already been used to design machines that can, for example, detect underwater mines (Gorman and Sejnowski 1988) after being fed hundreds of cases of mines and non-mines.

There are several reasons machine learning does not present an easy solution for Friendly AI. The first is that, of course, humans themselves hold deep disagreements about what is moral and immoral. But even if humans could be made to agree on all the training cases, at least two problems would remain.

The first problem is that training on cases from our present reality may not result in a machine that will make correct ethical decisions in a world radically reshaped by superintelligence.

The second problem is that a superintelligence may generalize the wrong principles due to coincidental patterns in the training data (Yudkowsky 2008). Consider the parable of the machine trained to recognize camouflaged tanks in a forest. Researchers take a hundred photos of camouflaged tanks and a hundred photos of trees. They then train the machine on fifty photos of each, so that it learns to distinguish camouflaged tanks from trees. As a test, they show the machine the remaining fifty photos of each, and it classifies each one correctly. Success! However, later tests show that the machine classifies additional photos of camouflaged tanks and trees poorly. The problem turns out to be that the researchers' photos of camouflaged tanks had been taken on cloudy days, while their photos of trees had been taken on sunny days. The machine had learned to distinguish cloudy days from sunny days, not camouflaged tanks from trees.

Thus, it seems that trustworthy Friendly AI design must involve detailed models of the underlying processes generating human moral judgments, not only surface similarities of cases.

Further reading:

- "Artificial Intelligence as a Positive and Negative Factor in Global Risk" (Yudkowsky 2008)

4.7. What is coherent extrapolated volition?

Eliezer Yudkowsky has proposed coherent extrapolated volition as a solution to at least two problems facing Friendly AI design (Yudkowsky 2004):

1. *The fragility of human values*: Yudkowsky writes that "any future not shaped by a goal system with detailed reliable inheritance from human morals and metamorals will contain almost nothing of worth." The problem is that what humans value is complex and subtle, and difficult to specify. Consider the seemingly minor value of *novelty*. If a human-like value of novelty is not programmed into a superintelligent machine, it might explore the universe for valuable things up to a certain point and then maximize the most valuable thing it finds (the exploration-exploitation tradeoff [Azoulay-Schwartz, Kraus, and Wilkenfeld 2004])—tiling the solar system with brains in vats wired into happiness machines, for example. When a superintelligence is in charge, you have to get its motivational system *exactly right* in order to *not* make the future undesirable.
2. *The locality of human values*: Imagine if the Friendly AI problem had faced the ancient Greeks, and they had programmed it with the most progressive moral values of their time. That would have led the world to a rather horrifying fate. But why should we think that humans have, in the twenty-first century, arrived at the apex of human morality? We can't risk programming a superintelligent machine with the moral values we happen to hold today. But then which moral values *do* we give it?

Yudkowsky suggests that we build a "seed AI" to discover and then extrapolate the "coherent extrapolated volition" of humanity:

> In poetic terms, our coherent extrapolated volition is our wish if we knew more, thought faster, were more the people we wished we were, had grown up farther together; where the extrapolation converges rather than diverges, where our wishes cohere rather than interfere; extrapolated as we wish that extrapolated, interpreted as we wish that interpreted. (Yudkowsky 2004)

The seed AI would use the results of this examination and extrapolation of human values to program the motivational system of the superintelligence that would determine the fate of the galaxy.

However, some worry that the collective will of humanity won't converge on a coherent set of goals. Others believe that guaranteed friendliness is not possible, even by such elaborate and careful means.

Further reading:

- "Coherent Extrapolated Volition" (Yudkowsky 2004)

4.8. Can we add friendliness to any artificial intelligence design?

Many AI designs that would generate an intelligence explosion would not have a "slot" in which a goal (such as "be friendly to human interests") could be placed. For example, if AI is made via whole-brain emulation, or evolutionary algorithms, or neural nets, or reinforcement learning, the AI will end up with some goal as it self-improves, but that eventual stable goal may be very difficult to predict in advance.

Thus, in order to design a Friendly AI, it is not sufficient to determine what "friendliness" is (and to specify it clearly enough that even a superintelligence will interpret it the way we want it to). We must also figure out how to build a general intelligence that satisfies a goal at all, and that stably retains that goal as it edits its own code to make itself smarter. This task is perhaps the primary difficulty in designing Friendly AI.

4.9. Who is working on the Friendly AI problem?

Today, Friendly AI research is being explored by the Machine Intelligence Research Institute (in Berkeley, California), by the Future of Humanity Institute (in Oxford, U.K.), and by a few other researchers such as David Chalmers. Machine ethics researchers occasionally touch on the problem, for example Wendell Wallach and Colin Allen in *Moral Machines* (Wallach and Allen 2009).

References

Aldridge, J. Wayne, and Kent C. Berridge. 2009. "Neural Coding of Pleasure: 'Rose-Tinted Glasses' of the Ventral Pallidum." In Kringelbach and Berridge 2009, 62–73.

Allen, Colin. 2002. "Calculated Morality: Ethical Computing in the Limit." In *Cognitive, Emotive and Ethical Aspects of Decision Making & Human Action,* edited by Iva Smit and George E. Lasker, 19–23. Vol. 1. Windsor, ON: International Institute for Advanced Studies in Systems Research / Cybernetics.

Anderson, Michael, and Susan Leigh Anderson, eds. 2006. "Machine Ethics." Special issue, *IEEE Intelligent Systems* 21 (4). doi:10.1109/MIS.2006.69.

———, eds. 2011. *Machine Ethics.* New York: Cambridge University Press.

Anderson, Michael, Susan Leigh Anderson, and Chris Armen, eds. 2005a. *Machine Ethics: Papers from the 2005 AAAI Fall Symposium.* Technical Report, FS-05-06. AAAI Press, Menlo Park, CA. http://www.aaai.org/Library/Symposia/Fall/fs05-06.

———. 2005b. "Towards Machine Ethics: Implementing Two Action-Based Ethical Theories." In Anderson, Anderson, and Armen 2005a, 1–7.

Anderson, Susan Leigh. 2008. "Asimov's 'Three Laws of Robotics' and Machine Metaethics." *AI & Society* 22 (4): 477–493. doi:10.1007/s00146-007-0094-5.

Arkin, Ronald C. 2009. *Governing Lethal Behavior in Autonomous Robots.* Boca Raton, FL: CRC Press.

Asimov, Isaac. 1942. "Runaround." *Astounding Science-Fiction,* March, 94–103.

Azoulay-Schwartz, Rina, Sarit Kraus, and Jonathan Wilkenfeld. 2004. "Exploitation vs. Exploration: Choosing a Supplier in an Environment of Incomplete Information." *Decision Support Systems* 38 (1): 1–18. doi:10.1016/S0167-9236(03)00061-7.

Bainbridge, William Sims. 2006. "Survey of NBIC Applications." In *Managing Nano-Bio-Info-Cogno Innovations: Converging Technologies in Society,* edited by William Sims Bainbridge and Mihail C. Roco, 337–346. Dordrecht, The Netherlands: Springer.

Baum, Seth D., Ben Goertzel, and Ted G. Goertzel. 2011. "How Long Until Human-Level AI? Results from an Expert Assessment." *Technological Forecasting and Social Change* 78 (1): 185–195. doi:10.1016/j.techfore.2010.09.006.

Block, Ned. 1981. "Psychologism and Behaviorism." *Philosophical Review* 90 (1): 5–43. doi:10.2307/2184371.

Bostrom, Nick. 1998. "How Long Before Superintelligence?" *International Journal of Futures Studies* 2.

Bostrom, Nick, and Milan M. Ćirković, eds. 2008. *Global Catastrophic Risks.* New York: Oxford University Press.

Bostrom, Nick, and Anders Sandberg. 2009. "Cognitive Enhancement: Methods, Ethics, Regulatory Challenges." *Science and Engineering Ethics* 15 (3): 311–341. doi:10.1007/s11948-009-9142-5.

Caplan, Bryan. 2008. "The Totalitarian Threat." In Bostrom and Ćirković 2008, 504–519.

Capurro, Rafael, Thomas Hausmanninger, Karsten Weber, and Felix Weil, eds. 2006. "Ethics in Robotics." Special issue, *International Review of Information Ethics* 6. http://www.i-r-i-e.net/issue6.htm.

Chalmers, David John. 2010. "The Singularity: A Philosophical Analysis." *Journal of Consciousness Studies* 17 (9–10): 7–65. http://www.ingentaconnect.com/content/imp/jcs/2010/00000017/f0020009/art00001.

Danielson, Peter. 1992. *Artificial Morality: Virtuous Robots for Virtual Games.* New York: Routledge.

de Garis, Hugo. 2007. "Artificial Brains." In Goertzel and Pennachin 2007, 159–174.

Dreyfus, Hubert L. 1972. *What Computers Can't Do: A Critique of Artificial Reason.* New York: Harper & Row.

Fox, Joshua. 2011. "A Review of Proposals Toward Safe AI." *Adarti* (blog), April 5. http://blog.joshuafox.com/2011/04/review-of-proposals-toward-safe-ai.html.

Frijda, Nico H. 2009. "On the Nature and Function of Pleasure." In Kringelbach and Berridge 2009, 99–112.

Goertzel, Ben, Seth Baum, and Ted Goertzel. 2010. "How Long Till Human-Level AI?" *H+ Magazine,* February 5. http://hplusmagazine.com/2010/02/05/how-long-till-human-level-ai/.

Goertzel, Ben, and Cassio Pennachin, eds. 2007. *Artificial General Intelligence.* Cognitive Technologies. Berlin: Springer. doi:10.1007/978-3-540-68677-4.

Good, Irving John. 1965. "Speculations Concerning the First Ultraintelligent Machine." In *Advances in Computers,* edited by Franz L. Alt and Morris Rubinoff, 31–88. Vol. 6. New York: Academic Press. doi:10.1016/S0065-2458(08)60418-0.

Gorman, R. Paul, and Terrence J. Sejnowski. 1988. "Analysis of Hidden Units in a Layered Network Trained to Classify Sonar Targets." *Neural Networks* 1 (1): 75–89. doi:10.1016/0893-6080(88)90023-8.

Grossberg, Stephen, ed. 1992. *Neural Networks and Natural Intelligence.* Bradford Books. Cambridge, MA: MIT Press.

Guarini, Marcello. 2006. "Particularism and the Classification and Reclassification of Moral Cases." *IEEE Intelligent Systems* 21 (4): 22–28. doi:10.1109/MIS.2006.76.

Hochberg, Leigh R., Mijail D. Serruya, Gerhard M. Friehs, Jon A. Mukand, Maryam Saleh, Abraham H. Caplan, Almut Branner, David Chen, Richard D. Penn, and John P. Donoghue. 2006. "Neuronal Ensemble Control of Prosthetic Devices by a Human with Tetraplegia." *Nature* 442 (7099): 164–171. doi:10.1038/nature04970.

Honarvar, Ali Reza, and Nasser Ghasem-Aghaee. 2009. "An Artificial Neural Network Approach for Creating an Ethical Artificial Agent." In *2009 IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA),* 290–295. Piscataway, NJ: IEEE Press. doi:10.1109/CIRA.2009.5423190.

Hutter, Marcus. 2005. *Universal Artificial Intelligence: Sequential Decisions Based On Algorithmic Probability.* Texts in Theoretical Computer Science. Berlin: Springer. doi:10.1007/b138233.

———. 2007. "Universal Algorithmic Intelligence: A Mathematical Top→Down Approach." In Goertzel and Pennachin 2007, 227–290.

Kaste, Martin. 2011. "The Singularity: Humanity's Last Invention?" *NPR,* All Things Considered (January 11). Accessed November 4, 2012. http://www.npr.org/2011/01/11/132840775/The-Singularity-Humanitys-Last-Invention.

King, Ross D. 2011. "Rise of the Robo Scientists." *Scientific American* 304 (1): 72–77. doi:10.1038/scientificamerican0111-72.

King, Ross D., Jem Rowland, Stephen G. Oliver, Michael Young, Wayne Aubrey, Emma Byrne, Maria Liakata, et al. 2009. "The Automation of Science." *Science* 324 (5923): 85–89. doi:10.1126/science.1165620.

Kringelbach, Morten L., and Kent C. Berridge, eds. 2009. *Pleasures of the Brain.* Series in Affective Science. New York: Oxford University Press.

Kurzweil, Ray. 2005. *The Singularity Is Near: When Humans Transcend Biology.* New York: Viking.

Legg, Shane. 2008. "Machine Super Intelligence," University of Lugano. http://www.vetta.org/documents/Machine_Super_Intelligence.pdf.

Lokhorst, Gert-Jan C. 2011. "Computational Meta-Ethics: Towards the Meta-Ethical Robot." *Minds and Machines* 21 (2): 261–274. doi:10.1007/s11023-011-9229-z.

Mackenzie, Donald. 1995. "The Automation of Proof: A Historical and Sociological Exploration." *IEEE Annals of the History of Computing* 17 (3): 7–29. doi:10.1109/85.397057.

Markoff, John. 2011. "Computer Wins on 'Jeopardy!': Trivial, It's Not." *New York Times,* February 16. http://www.nytimes.com/2011/02/17/science/17jeopardy-watson.html.

Martinetz, Thomas, and Klaus Schulten. 1991. "A 'Neural-Gas' Network Learns Topologies." In *Artificial Neural Networks,* edited by Teuvo Kohonen, Kai Mäkisara, Olli Simula, and Jari Kangas, 397–402. Amsterdam: Elsevier.

McLaren, Bruce M. 2005. "Lessons in Machine Ethics from the Perspective of Two Computational Models of Ethical Reasoning." In Anderson, Anderson, and Armen 2005a, 70–77.

Moravec, Hans P. 1999. "Rise of the Robots." *Scientific American,* December, 124–135.

Muehlhauser, Luke. 2011. "Friendly AI: A Bibliography." *Common Sense Atheism* (blog), February 4. http://commonsenseatheism.com/?p=12147.

Neisser, Ulric, Gwyneth Boodoo, Thomas J. Bouchard Jr., A. Wade Boykin, Nathan Brody, Stephen J. Ceci, Diane F. Halpern, et al. 1996. "Intelligence: Knowns and Unknowns." *American Psychologist* 51 (2): 77–101. doi:10.1037/0003-066X.51.2.77.

Nielsen, Michael. 2011. "What Should a Reasonable Person Believe about the Singularity?" *Michael Nielsen* (blog), January 12. http://michaelnielsen.org/blog/what-should-a-reasonable-person-believe-about-the-singularity/.

Nilsson, Nils J. 2009. *The Quest for Artificial Intelligence: A History of Ideas and Achievements.* New York: Cambridge University Press.

Omohundro, Stephen M. 2008. "The Basic AI Drives." In Wang, Goertzel, and Franklin 2008, 483–492.

Pennachin, Cassio, and Ben Goertzel. 2007. "Contemporary Approaches to Artificial General Intelligence." In Goertzel and Pennachin 2007, 1–30.

Penrose, Roger. 1994. *Shadows of the Mind: A Search for the Missing Science of Consciousness.* New York: Oxford University Press.

Powers, Thomas M. 2005. "Deontological Machine Ethics." In Anderson, Anderson, and Armen 2005a, 79–84.

Rattner, Justin. 2008. *Crossing the Chasm between Humans and Machines.* Keynote presentation at Intel Developer Forum. San Francisco, CA, August 21. http://download.intel.com/pressroom/kits/events/idffall_2008/JustinRattner_keynote_transcript.pdf.

Russell, Stuart J., and Peter Norvig. 2010. *Artificial Intelligence: A Modern Approach*. 3rd ed. Upper Saddle River, NJ: Prentice-Hall.

Rzepka, Rafal, and Kenji Araki. 2005. "What Statistics Could Do for Ethics? The Idea of Common Sense Processing Based Safety Valve." In Anderson, Anderson, and Armen 2005a, 85–86.

Sandberg, Anders, and Nick Bostrom. 2008. *Whole Brain Emulation: A Roadmap*. Technical Report, 2008-3. Future of Humanity Institute, University of Oxford. http://www.fhi.ox.ac.uk/Reports/2008-3.pdf.

———. 2011. *Machine Intelligence Survey*. Technical Report, 2011-1. Future of Humanity Institute, University of Oxford. www.fhi.ox.ac.uk/reports/2011-1.pdf.

Sawyer, Robert J. 2007. "Robot Ethics." *Science* 318 (5853): 1037. doi:10.1126/science.1151606.

Schmidhuber, Jürgen. 2007. "Gödel Machines: Fully Self-Referential Optimal Universal Self-Improvers." In Goertzel and Pennachin 2007, 199–226.

Schroeder, Timothy. 2004. *Three Faces of Desire*. Philosophy of Mind Series. New York: Oxford University Press. doi:10.1093/acprof:oso/9780195172379.001.0001.

Searle, John R. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3 (03): 417–424. doi:10.1017/S0140525X00005756.

Shulman, Carl. 2010. *Omohundro's "Basic AI Drives" and Catastrophic Risks*. The Singularity Institute, San Francisco, CA. http://intelligence.org/files/BasicAIDrives.pdf.

Smart, R. N. 1958. "Negative Utilitarianism." *Mind*, n.s., 67 (268): 542–543. http://www.jstor.org/stable/2251207.

Smith, Kyle, Stephen V. Mahler, Susana Pecina, and Kent C. Berridge. 2009. "Hedonic Hotspots: Generating Sensory Pleasure in the Brain." In Kringelbach and Berridge 2009, 27–49.

Tang, Y. P., E. Shimizu, G. R. Dube, C. Rampon, G. A. Kerchner, M. Zhuo, G. Liu, and J. Z. Tsien. 1999. "Genetic Enhancement of Learning and Memory in Mice." *Nature* 401 (6748): 63–69. doi:10.1038/43432.

Veness, Joel, Kee Siong Ng, Marcus Hutter, William Uther, and David Silver. 2011. "A Monte-Carlo AIXI Approximation." *Journal of Artificial Intelligence Research* 40:95–142. doi:10.1613/jair.3125.

Vinge, Vernor. 1993. "The Coming Technological Singularity: How to Survive in the Post-Human Era." In *Vision-21: Interdisciplinary Science and Engineering in the Era of Cyberspace*, 11–22. NASA Conference Publication 10129. NASA Lewis Research Center. http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19940022855_1994022855.pdf.

Wallach, Wendell, and Colin Allen. 2009. *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press. doi:10.1093/acprof:oso/9780195374049.001.0001.

Wallach, Wendell, Colin Allen, and Iva Smit. 2008. "Machine Morality: Bottom-Up and Top-Down Approaches for Modelling Human Moral Faculties." In "Ethics and Artificial Agents." Special issue, *AI & Society* 22 (4): 565–582. doi:10.1007/s00146-007-0099-0.

Wang, Pei, Ben Goertzel, and Stan Franklin, eds. 2008. *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*. Frontiers in Artificial Intelligence and Applications 171. Amsterdam: IOS.

Wasserman, Edward, and Thomas Zentall, eds. 2006. *Comparative Cognition: Experimental Explorations of Animal Intelligence*. New York: Oxford University Press.

Yudkowsky, Eliezer. 2001. *Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures.* The Singularity Institute, San Francisco, CA, June 15. http://intelligence.org/files/CFAI.pdf.

———. 2002. "The AI-Box Experiment." Accessed January 15, 2012. http://yudkowsky.net/singularity/aibox.

———. 2004. *Coherent Extrapolated Volition.* The Singularity Institute, San Francisco, CA, May. http://intelligence.org/files/CEV.pdf.

———. 2007a. "Levels of Organization in General Intelligence." In Goertzel and Pennachin 2007, 389–501.

———. 2007b. "The Hidden Complexity of Wishes." *LessWrong* (blog), November 24. http://lesswrong.com/lw/ld/the_hidden_complexity_of_wishes/.

———. 2007c. "The Power of Intelligence." Accessed March 8, 2013. http://yudkowsky.net/singularity/power.

———. 2008. "Artificial Intelligence as a Positive and Negative Factor in Global Risk." In Bostrom and Ćirković 2008, 308–345.

———. 2011. "Becoming a Rationalist." Interview by Luke Muehlhauser, *Conversations from the Pale Blue Dot* (February 5). http://commonsenseatheism.com/?p=12147.