Intelligence Explosion FAQ

(PDF Version Available (/files/IE-FAQ.pdf))

---

1. Basics

1.1. What is the intelligence explosion?

The intelligence explosion idea was expressed by statistician I.J. Good in 1965[13]:

> 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.

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*[14]. By the late 1990s, 'expert systems' had surpassed human skill for a wide range of tasks.[15] In 1997, IBM's Deep Blue computer beat the world chess champion[16], and in 2011, IBM's Watson computer beat the best human players at a much more complicated game: *Jeopardy!*[17]. Recently, a robot named Adam was programmed with our scientific knowledge about yeast, then posed its own hypotheses, tested them, and assessed the results.[18][19]

Computers remain far short of human intelligence, but the resources that aid AI design are accumulating (including hardware, large datasets, neuroscience knowledge, and AI theory). We may one day design a machine that surpasses human skill *at designing artificial intelligences*. After that, this machine could improve its own intelligence faster and better

than humans can, 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.

This is what is meant by the 'intelligence explosion' in this FAQ.

See also:

- Vinge, The Coming Technological Singularity (http://www-rohan.sdsu.edu/faculty/vinge/misc/singularity.html)
- Wikipedia, Technological Singularity (http://en.wikipedia.org/wiki/Technological_singularity)
- Chalmers, The Singularity: A Philosophical Analysis (http://commonsenseatheism.com/wp-content/uploads/2011/01/Chalmers-The-Singularity-a-philosophical-analysis.pdf)

2. How Likely is an Intelligence Explosion?

2.1. How is 'intelligence' defined?

Artificial intelligence researcher Shane Legg defines[20] intelligence like this:

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

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

See also:

- Wikipedia, Intelligence (http://en.wikipedia.org/wiki/Intelligence)
- Neisser et al., Intelligence: Knowns and Unknowns (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.134.1282&rep=rep1&type=pdf)
- Wasserman & Zentall (eds.), *Comparative Cognition:*

Experimental Explorations of Animal Intelligence (http://www.amazon.com/Comparative-Cognition-Experimental-Explorations-Intelligence/dp/019537780X/)

- Legg, Definitions of Intelligence (http://www.vetta.org/definitions-of-intelligence/)

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 databanks, detecting underwater mines, and more.[15] 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 was never trained. Humans can even examine the processes that produce their own intelligence (cognitive neuroscience (http://en.wikipedia.org/wiki/Cognitive_neuroscience)), and design new kinds of intelligence never seen before (artificial intelligence (http://en.wikipedia.org/wiki/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[21] a formal model called AIXI that he says possesses the greatest general intelligence possible. 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.[22]

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).

See also:

- Goertzel & Pennachin (eds.), *Artificial General Intelligence (http://www.amazon.com/dp/3642062679/)*
- Sandberg & Bostrom, *Whole Brain Emulation: A Roadmap (http://www.philosophy.ox.ac.uk/__data/assets/pdf_file/0019/3853/brain-emulation-roadmap-report.pdf)*
- Bostrom & Sandberg, Cognitive Enhancement: Methods, Ethics, Regulatory Challenges (http://www.nickbostrom.com/cognitive.pdf)
- Wikipedia, Brain-computer interface (http://en.wikipedia.org/wiki/Brain%E2%80%93computer_interface)

2.3. What is whole brain emulation?

Whole Brain Emulation (WBE) or 'mind uploading' is a computer emulation of all the cells and connections in a human brain. So 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). Such a WBE could do more thinking in one second than a normal human can in 31 years. 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.

See also:

- Sandberg & Bostrom, *Whole Brain Emulation: A Roadmap (http://www.philosophy.ox.ac.uk/__data/assets/pdf_file/0019/3853/brain-emulation-roadmap-report.pdf)*
- Blue Brain Project (http://bluebrain.epfl.ch/)

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 over-expressed the NR2B gene, which improved those mice's memory beyond that of any other mice of any mouse species.[23] Biological cognitive enhancement in humans may cause an intelligence explosion to occur more quickly than it otherwise would.

See also:

- Bostrom & Sandberg, Cognitive Enhancement: Methods, Ethics, Regulatory Challenges (http://www.nickbostrom.com/cognitive.pdf)

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 dozens of successes. Three successes in human BCIs are a device (http://archives.cnn.com/2002/HEALTH/06/13/cov.bionic.eye/index.html) that restores (partial) sight to the blind, cochlear implants (http://en.wikipedia.org/wiki/Cochlear_implant) that restore hearing to the deaf, and a device that allows use of an artificial hand by direct thought.[24]

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

See also:

- Wikipedia, Brain-computer interface

(http://en.wikipedia.org/wiki/Brain%E2%80%93computer_interface)

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

There are many paths to artificial general intelligence (AGI). 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.[29][30][31] Another path is to start with a formal model of perfect general intelligence and try to approximate that.[32][33] 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.[34] Eurisko (http://en.wikipedia.org/wiki/Eurisko) is a self-improving AI in a limited domain, but is not able to achieve human-level general intelligence.

See also:

- Pennachin & Goertzel, Contemporary Approaches to Artificial General Intelligence (http://commonsenseatheism.com/wp-content/uploads/2011/03/Pennachin-Goertzel-Contemporary-Approaches-to-Artificial-General-Intelligence.pdf)

2.7. What is superintelligence?

Nick Bostrom defined[25] '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.

This definition includes vague terms like 'much' and 'practically', but it will serve as a 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.

See also:

- Bostrom, How Long Before Superintelligence? (http://www.nickbostrom.com/superintelligence.html)
- Legg, *Machine Super Intelligence (http://www.vetta.org/documents/Machine_Super_Intelligence.pdf)*

2.8. When will an intelligence explosion happen?

Predicting the future is risky business. There are many philosophical, scientific, technological, and social uncertainties relevant to the arrival of an intelligence explosion. Because of this, experts disagree on when this event might 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.[26]
- Intel's chief technology officer, Justin Rattner, expects (http://www.techwatch.co.uk/2008/08/22/intel-predicts-singularity-by-2048/) "a point when human and artificial intelligence merges to create something bigger than itself" by 2048.
- AI researcher Eliezer Yudkowsky expects (http://commonsenseatheism.com/?p=12147) the intelligence explosion by 2060.
- Philosopher David Chalmers has over 1/2 credence in the intelligence explosion occurring by 2100.[27]
- Quantum computing expert Michael Nielsen estimates (http://michaelnielsen.org/blog/what-should-a-reasonable-person-believe-about-the-singularity/) that the probability of the intelligence explosion occurring by 2100 is between 0.2% and about 70%.
- 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.[28]

- iRobot CEO Rodney Brooks (http://itc.conversationsnetwork.org/shows/detail3400.html) and cognitive scientist Douglas Hofstadter (http://video.google.com/videoplay?docid=8832143373632003914) allow that the intelligence explosion may occur in the future, but probably not in the 21st century.
- Roboticist Hans Moravec predicts that AI will surpass human intelligence "well before 2050 (http://www.scientificamerican.com/article.cfm?id=rise-of-the-robots&print=true)."
- In a 2005 survey of 26 contributors to a series of reports on emerging technologies, the median estimate for machines reaching human-level intelligence was 2085.[61]
- 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.[62]
- On the other hand, 41% of the participants in the AI@50 conference (in 2006) stated (http://www.engagingexperience.com/ai50/) that machine intelligence would *never* reach the human level.

See also:

- Baum, Goertzel, & Goertzel, How Long Until Human-Level AI? Results from an Expert Assessment (http://sethbaum.com/ac/2011_AI-Experts.pdf)

2.9. Might an intelligence explosion never occur?

Dreyfus[35] and Penrose[36] have argued that human cognitive abilities can't be emulated by a computational machine. Searle[37] and Block[38] 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.[27]

We can reply to Dreyfus and Penrose by noting that an intelligence explosion does not require an AI to be a classical computational system. And we can reply to Searle and Block by noting that an intelligence explosion does not depend on machines having consciousness or other properties of 'mind', only that it 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 an 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.[28]

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.[59]

3. Consequences of an Intelligence Explosion

3.1. Why would great intelligence produce great power?

Intelligence is powerful.[60][20] 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 can to give itself more power to achieve its goals — all at a speed much faster than humans can 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 21st century, including global nuclear war, bioweapons, superviruses, and more.[56] A superintelligent machine would be more capable of solving those problems than humans are.

See also:

- Yudkowsky, Artificial intelligence as a positive and negative factor in global risk (/files/AIPosNegFactor.pdf)

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 appeared 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.[55] As Yudkowsky 55 (/files/AIPosNegFactor.pdf)] As Yudkowsky writes (/files/AIPosNegFactor.pdf',100])), “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 better achieve their goal 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 superficially benevolent motivations could easily go awry when it discovers 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.

See also:

- Yudkowsky, Artificial intelligence as a positive and negative factor in global risk (/files/AIPosNegFactor.pdf)
- Chalmers, The Singularity: A Philosophical Analysis (http://consc.net/papers/singularity.pdf)

4. Friendly AI

4.1. What is Friendly AI?

A Friendly Artificial Intelligence (Friendly AI or FAI) 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’[2][3][8][9], ‘machine morality’[11], ‘artificial morality’[6], ‘computational ethics’[12] and ‘computational metaethics’[7], ‘friendly AI’[1], and ‘robo-ethics’ or ‘robot ethics’.[5][10]

The most immediate concern may be in battlefield robots; the U.S. Department of Defense contracted Ronald Arkin to design a system for ensuring ethical behavior in autonomous battlefield robots[4]. The U.S. Congress has declared that a third of America’s ground systems must be robotic by 2025, and by 2030 the U.S. Air Force plans (http://commonsenseatheism.com/?p=14918) to have swarms of bird-sized flying robots that operate semi-autonomously for weeks at a time.

But Friendly AI research is not 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[41][42][43] 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.

See also:

- Wikipedia, Friendly Artificial Intelligence (http://en.wikipedia.org/wiki/Friendly_artificial_intelligence).
- *All Things Considered*, The Singularity: Humanity’s Last Invention? (http://www.npr.org/2011/01/11/132840775/The-Singularity-Humanitys-Last-Invention)
- Fox, A review of proposals toward safe AI (http://adarti.blogspot.com/2011/04/review-of-proposals-toward-safe-ai.html)
- Muehlhauser, Friendly AI: A Bibliography (http://commonsenseatheism.com/?p=14047)

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 can 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 in’ to 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.

See also:

- Omohundro, The Basic AI Drives (http://selfawaresystems.files.wordpress.com/2008/01/ai_drives_final.pdf)
- Shulman, Basic AI Drives and Catastrophic Risks (/files/BasicAIDrives.pdf)

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 (http://yudkowsky.net/singularity/aibox) in AI-boxing do not inspire confidence. And, a superintelligence will generate far more persuasive techniques for getting humans to "let it out of the box" than we can imagine.
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[39]: (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.[40]

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 don'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 (http://lesswrong.com/lw/ld/the_hidden_complexity_of_wishes/), 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 (http://lesswrong.com/lw/4qh/why_not_just_write_failsafe_rules_into_the/3nkv): "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 (http://en.wikipedia.org/wiki/Utilitarianism) designs for Friendly AI.

An AI designed to minimize human suffering might simply kill all humans: no humans, no human suffering.[44][45]

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 (http://en.wikipedia.org/wiki/Experience_machine). Or, it could rewire the 'liking' component

(http://lesswrong.com/lw/4yq/the_neuroscience_of_pleasure/) of the brain's reward system (http://www.scholarpedia.org/article/Reward) so that whichever hedonic hotspot[48] paints sensations with a 'pleasure gloss'[46][47] 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[49][50][51][52] 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[53] after feeding the machine hundreds of cases of mines and not-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 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.[54] Consider the parable of the machine trained to recognize camouflaged tanks in a forest. Researchers take 100 photos of camouflaged tanks and 100 photos of trees. They then train the machine on 50 photos of each, so that it learns to distinguish camouflaged tanks from trees. As a test, they show the machine the remaining 50 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.

See also:

- Yudkowsky, Artificial intelligence as a positive and negative factor in global risk (/files/AIPosNegFactor.pdf)

4.7. What is Coherent Extrapolated Volition?

Eliezer Yudkowsky has proposed[57] Coherent Extrapolated Volition as a solution to at least two problems facing Friendly AI design:

1. *The fragility of human values*: Yudkowsky writes (http://lesswrong.com/lw/y3/value_is_fragile/) 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[58]) — 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 21st 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 (/files/CEV.pdf) 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.

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 (http://multiverseaccordingtoben.blogspot.com/2010/10/singularity-institutes-scary-idea-and.html) that guaranteed Friendliness is not possible, even by such elaborate and careful means.

- Yudkowsky, Coherent Extrapolated Volition (/files/CEV.pdf)

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 stable eventual 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 (https://intelligence.org/) (in Berkeley, California), by the Future of Humanity Institute (http://www.fhi.ox.ac.uk/) (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 (http://www.amazon.com/Moral-Machines-Teaching-Robots-Right/dp/0199737975/)*.

References

[1] Yudkowsky (2001). *Creating Friendly AI 1.0 (/files/CFAI.pdf)*. Machine Intelligence Research Institute.

[2] Anderson & Anderson, eds. (2006). *IEEE Intelligent Systems, 21(4)*.

[3] Anderson & Anderson, eds. (2011). *Machine Ethics (http://www.amazon.com/dp/0521112354/)*. Cambridge University Press.

[4] Arkin (2009). *Governing Lethal Behavior in Autonomous Robots (http://www.amazon.com/dp/1420085948/)*. Chapman and Hall.

[5] Capurro, Hausmanninger, Weber, Weil, Cerqui, Weber, & Weber (2006). *International Review of Information Ethics, Vol. 6: Ethics in Robots (http://commonsenseatheism.com/wp-content/uploads/2011/03/Capurro-International-Review-of-Information-Ethics-Vol.-6-Ethics-in-Robotics.pdf)*.

[6] Danielson (1992). *Artificial morality: Virtuous robots for virtual games (http://www.amazon.com/dp/0415076919/)*. Routledge.

[7] Lokhorst (2011). Computational meta-ethics: Towards the meta-ethical robot (http://commonsenseatheism.com/wp-content/uploads/2011/03/Lokhorst-Computational-meta-ethics-toward-the-meta-ethical-robot.pdf). *Minds and Machines*.

[8] McLaren (2005). Lessons in Machine Ethics from the Perspective of Two Computational Models of Ethical Reasoning. *AAAI Technical Report FS-05-06*: 70-77.

[9] Powers (2005). Deontological Machine Ethics. *AAAI Technical Report FS-05-06*: 79-86.

[10] Sawyer (2007). Robot ethics. *Science, 318(5853)*: 1037.

[11] Wallach, Allen, & Smit (2008). Machine morality: Bottom-up and top-down approaches for modeling human moral faculties. *AI and Society, 22(4)*: 565–582.

[12] Allen (2002). Calculated morality: Ethical computing in the limit. In Smit & Lasker, eds., *Cognitive, emotive and ethical aspects of decision making and human action, vol I*. Baden/IIAS.

[13] Good (1965). Speculations concerning the first ultraintelligent machine (http://commonsenseatheism.com/wp-content/uploads/2011/01/Good-Speculations-Concerning-the-First-UltraIntelligent-Machine.pdf). *Advanced in Computers, 6*: 31-88.

[14] MacKenzie (1995). The Automation of Proof: A Historical and Sociological Exploration. *IEEE Annals, 17(3)*: 7-29.

[15] Nilsson (2009). *The Quest for Artificial Intelligence (http://www.amazon.com/dp/0521122937/)*. Cambridge University Press.

[16] Campbell, Hoane, & Hsu (2002). Deep Blue. *Artificial Intelligence, 134*: 57-83.

[17] Markoff (2011). Computer Wins on 'Jeopardy!'; Trivial, it's Not (http://www.nytimes.com/2011/02/17/science/17jeopardy-watson.html?_r=2&ref=homepage&src=me&pagewanted=all). *New York Times, February 17th 2011*: A1.

[18] King et al. (2009). The automation of science (http://commonsenseatheism.com/wp-content/uploads/2011/02/King-The-Automation-of-Science.pdf). *Science, 324*: 85-89.

[19] King (2011). Rise of the robo scientists (http://commonsenseatheism.com/wp-content/uploads/2011/02/King-Rise-of-the-Robo-Scientists.pdf). *Scientific American, January 2011*.

[20] Legg (2008). *Machine Super Intelligence (http://www.vetta.org/documents/Machine_Super_Intelligence.pdf)*. PhD Thesis. IDSIA.

[21] Hutter (2005). *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability (http://www.amazon.com/dp/3642060528/)*. Springer.

[22] Veness, Ng, Hutter, & Silver (2011). A Monte Carlo AIXI Approximation (http://arxiv.org/PS_cache/arxiv/pdf/0909/0909.0801v1.pdf). *Journal of Artificial Intelligence Research, 40*: 95-142.

[23] Tang, Shimizu, Dube, Rampon, Kerchner, Zhuo, Liu, & Tsien (1999). Genetic enhancement of learning and memory in mice. *Nature, 401*: 63–69.

[24] Hochberg, Serruya, Friehs, Mukand, Saleh, Caplan, Branner, Chen, Penn, & Donoghue (2006). Neuronal ensemble control of prosthetic devices by a human with tetraplegia. *Nature 442*: 164-171.

[25] Bostrom (1998). How long before superintelligence? (http://www.nickbostrom.com/superintelligence.html) *International Journal of Future Studies, 2*.

[26] Kurzweil (2005). *The Singularity is Near (http://www.amazon.com/dp/0143037889/)*. Viking.

[27] Chalmers (2010). The Singularity: A Philosophical Analysis (http://consc.net/papers/singularity.pdf). *Journal of Consciousness Studies, 17*: 7-65.

[28] Baum, Goertzel, & Goertzel (forthcoming). How Long Until Human-Level AI? Results from an Expert Assessment (http://sethbaum.com/ac/2011_AI-Experts.pdf). *Technological and Forecasting Change*.

[29] Grossberg (1992). *Neural Networks and Natural Intelligence (http://www.amazon.com/dp/0262570912/)*. MIT Press.

[30] Martinetz & Schulten (1991). A 'neural-gas' network learns topologies. In Kohonen, Makisara, Simula, & Kangas (eds.), *Artificial Neural Networks* (pp. 397-402). North Holland.

[31] de Garis (2010). Artificial Brains. In Goertzel & Pennachin (eds.), *Artificial General Intelligence* (pp. 159-174). Springer.

[32] Schmidhuber (2010). Gödel Machines: Fully Self-Referential Optimal Universal Self-Improvers. In Goertzel & Pennachin (eds.), *Artificial General Intelligence* (pp. 199-223). Springer.

[33] Hutter (2010). Universal Algorithmic Intelligence: A Mathematical Top-Down Approach. In Goertzel & Pennachin (eds.), *Artificial General Intelligence* (pp. 227-287). Springer.

[34] Yudkowsky (2010). Levels of Organization in General Intelligence (/files/LOGI.pdf). In Goertzel & Pennachin (eds.), *Artificial General Intelligence* (pp. 389-496). Springer.

[35] Dreyfus (1972). *What Computers Can't Do (http://www.amazon.com/dp/0060110821/)*. Harper & Row.

[36] Penrose (1994). *Shadows of the Mind (http://www.amazon.com/dp/0195106466/)*. Oxford University Press.

[37] Searle (1980). Minds, brains, and programs (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.120.749&rep=rep1&type=pdf). *Behavioral and Brain Sciences, 3*: 417-457.

[38] Block (1981). Psychologism and behaviorism (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.4.5828&rep=rep1&type=pdf). *Philosophical Review, 90*: 5-43.

[39] Asimov (1942). Runaround (http://www.rci.rutgers.edu/~cfs/472_html/Intro/NYT_Intro/History/Runaround.html). *Astounding Science Fiction, March 1942*. Street & Smith.

[40] Anderson (2008). Asimov's 'three laws of robotics' and machine metaethics. *AI & Society, 22(4)*: 477-493.

[41] Kringelbach & Berridge, eds. (2009). *Pleasures of the Brain (http://www.amazon.com/dp/0195331028/)*. Oxford University Press.

[42] Schroeder (2004). *Three Faces of Desire (http://www.amazon.com/dp/019517237X/)*. Oxford University Press.

[43] Yudkowsky (2007). The hidden complexity of wishes (http://lesswrong.com/lw/ld/the_hidden_complexity_of_wishes/).

[44] Smart (1958). Negative utilitarianism. *Mind, 67*: 542-543.

[45] Russell & Norvig (2009). *Artificial Intelligence: A Modern Approach, 3rd edition (http://www.amazon.com/dp/0136042597/)*. Prentice Hall. (see page 1037)

[46] Frijda (2009). On the nature and function of pleasure. In Kringelbach & Berridge (eds.), *Pleasures of the brain* (pp. 99-112). Oxford University Press.

[47] Aldridge & Berridge (2009). Neural coding of pleasure: 'rose-tinted glasses' of the ventral pallidum. In Kringelbach & Berridge (eds.), *Pleasures of the brain* (pp. 62-73). Oxford University Press.

[48] Smith, Mahler, Pecina, & Berridge (2009). Hedonic hotspots: generating sensory pleasure in the brain. In Kringelbach & Berridge (eds.), *Pleasures of the brain* (pp. 27-49). Oxford University Press.

[49] Guarini, (2006). Particularism and classification and reclassification of moral cases. IEEE Intelligent Systems 21(4): 22–28.

[50] Anderson, Anderson, & Armen (2005). Toward machine ethics: Implementing two action-based ethical theories. In *Proceedings of the AAAI Fall 2005 Symposium on Machine Ethics*, Arlington, Virginia, November.

[51] Honarvar & Ghasem-Aghaee (2009). An artificial neural network approach for creating an ethical artificial agent. *Proceedings of the 8th IEEE international conference on Computational intelligence in robotics and automation*: 290-295.

[52] Rzepka & Araki (2005). What statistics could do for ethics? – The idea of common sense processing based safety valve. In *Machine ethics: papers from the AAAI fall symposium*. American Association of Artificial Intelligence.

[53] Gorman & Sejnowski (1988). Analysis of hidden units in a layered network trained to classify sonar targets (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.84.6963&rep=rep1&type=pdf). *Neural Networks, 1*: 75-89.

[54] Yudkowsky (2008). Artificial intelligence as a positive and negative factor in global risk (/files/AIPosNegFactor.pdf). In Bostrom & Cirkovic (eds.), *Global Catastrophic Risks*. Oxford University Press.

[55] Omohundro (2008). The Basic AI Drives (http://selfawaresystems.files.wordpress.com/2008/01/ai_drives_final.pdf).

[56] Bostrom & Cirkovic, eds. (2008). *Global Catastrophic Risks (http://www.amazon.com/dp/0198570503/)*. Oxford University Press.

[57] Yudkowsky (2004). Coherent extrapolated volition (/files/CEV.pdf). Machine Intelligence Research Institute.

[58] Azoulay-Schwartz, Kraus, & Wilkenfeld (2004). Exploitation vs. exploration: choosing a supplier in an environment of incomplete information. *Decision Support Systems, 38*: 1-18.

[59] Caplan (2008). The totalitarian threat. In Bostrom & Cirkovic (eds.), *Global Catastrophic Risks*. Oxford University Press.

[60] Yudkowsky (2007). The Power of Intelligence (http://yudkowsky.net/singularity/power).

[61] Bainbridge (2005). Survey of NBIC Applications. In Bainbridge & Roco (eds.), *Managing nano-bio-info-cogno innovations: Converging technologies in society* (http://www.wtec.org/ConvergingTechnologies/3/NBIC3_report.pdf). Springer.

[62] Sandberg & Bostrom (2011). Machine intelligence survey (http://www.fhi.ox.ac.uk/__data/assets/pdf_file/0015/21516/MI_survey.pdf), Technical Report 2011-1, Future of Humanity Institute, Oxford.

Written by Luke Muehlhauser (http://lukeprog.com/).

This page is up-to-date as of 2013, but may not represent MIRI or Luke Muehlhauser's current views. Last modified November 10, 2015 (original (https://intelligence.org/files/IE-FAQ.pdf)).