FRIENDLY ARTIFICIAL INTELLIGENCE

3.1 INTRODUCTION: A **friendly artificial intelligence** (also **friendly AI** or **FAI**) is a hypothetical artificial general intelligence (AGI) that would have a positive rather than negative effect on humanity. The term was coined by Eliezer Yudkowsky to discuss superintelligent artificial agents that reliably implement human values. Stuart J. Russell and Peter Norvig's leading artificial intelligence textbook, *Artificial Intelligence: A Modern Approach*, describes the idea:

Yudkowsky (2008) goes into more detail about how to design a **Friendly AI**. He asserts that friendliness (a desire not to harm humans) should be designed in from the start, but that the designers should recognize both that their own designs may be flawed, and that the robot will learn and evolve over time. Thus the challenge is one of mechanism design—to define a mechanism for evolving AI systems under a system of checks and balances, and to give the systems utility functions that will remain friendly in the face of such changes.

'Friendly' is used in this context as technical terminology, and picks out agents that are safe and useful, not necessarily ones that are "friendly" in the colloquial sense. The concept is primarily invoked in the context of discussions of recursively self-improving artificial agents that rapidly explode in intelligence, on the grounds that this hypothetical technology would have a large, rapid, and difficult-to-control impact on human society.

3.2 Goals of a friendly AI, and inherent risks

Oxford philosopher Nick Bostrom has said that AI systems with goals that are not perfectly identical to or very closely aligned with human ethics are intrinsically dangerous unless extreme measures are taken to ensure the safety of humanity. He put it this way:

Basically we should assume that a 'superintelligence' would be able to achieve whatever goals it has. Therefore, it is extremely important that the goals we endow it with, and its entire motivation system, is 'human friendly.'

The roots of this concern are very old. Kevin LaGrandeur showed that the dangers specific to AI can be seen in ancient literature concerning artificial humanoid servants such as the golem, or the proto-robots of Gerbert of Aurillac and Roger Bacon. In those stories, the extreme intelligence and power of these humanoid creations clash with their status as slaves (which by nature are seen as sub-human), and cause disastrous conflict.

Ryszard Michalski, a pioneer of machine learning, taught his Ph.D. students decades ago that any truly alien mind, including a machine mind, was unknowable and therefore dangerous to humans.

More recently, Eliezer Yudkowsky has called for the creation of "friendly AI" to mitigate the existential threat of hostile intelligences.

Steve Omohundro says that all advanced AI systems will, unless explicitly counteracted, exhibit a number of basic drives/tendencies/desires because of the intrinsic nature of goal-driven systems and that these drives will, "without special precautions", cause the AI to act in ways that range from the disobedient to the dangerously unethical.

Alex Wissner-Gross says that AIs driven to maximize their future freedom of action (or causal path entropy) might be considered friendly if their planning horizon is longer than a certain threshold, and unfriendly if their planning horizon is shorter than that threshold.

Luke Muehlhauser recommends that machine ethics researchers adopt what Bruce Schneier has called the "security mindset": Rather than thinking about how a system will work, imagine how it could fail. For instance, even an AI that only makes accurate predictions and communicates via a text interface might cause unintended harm.

Coherent Extrapolated Volition

Yudkowsky advances the Coherent Extrapolated Volition (CEV) model. According to him, our coherent extrapolated volition is our choices and the actions we would collectively take if "we knew more, thought faster, were more the people we wished we were, and had grown up closer together."

Rather than a Friendly AI being designed directly by human programmers, it is to be designed by a seed AI programmed to first study human nature and then produce the AI which humanity would want, given sufficient time and insight to arrive at a satisfactory answer.[8] The appeal to an objective though contingent human nature (perhaps expressed, for mathematical purposes, in the form of a utility function or other decision-theoretic formalism), as providing the ultimate criterion of "Friendliness", is an answer to the meta-ethical problem of defining an objective morality; extrapolated volition is intended to be what humanity objectively would want, all things considered, but it can only be defined relative to the psychological and cognitive qualities of present-day, unextrapolated humanity.

Making the CEV concept precise enough to serve as a formal program specification is part of the research agenda of the Machine Intelligence Research Institute.

3.3 Other approaches

Ben Goertzel, an artificial general intelligence researcher, believes that friendly AI cannot be created with current human knowledge. Goertzel suggests humans may instead decide to create an "AI Nanny" with "mildly superhuman intelligence and surveillance powers", to protect the human race from existential risks like nanotechnology and to delay the development of other (unfriendly) artificial intelligences until and unless the safety issues are solved.

Steve Omohundro has proposed a "scaffolding" approach to AI safety, in which one provably safe AI generation helps build the next provably safe generation.

Public policy

James Barrat, author of *Our Final Invention*, suggested that "a public-private partnership has to be created to bring A.I.-makers together to share ideas about security—something like the International Atomic Energy Agency, but in partnership with corporations." He urges AI researchers to convene a meeting similar to the Asilomar Conference on Recombinant DNA, which discussed risks of biotechnology.

John McGinnis encourages governments to accelerate friendly AI research. Because the goalposts of friendly AI aren't necessarily clear, he suggests a model more like the National Institutes of Health, where "Peer review panels of computer and cognitive scientists would sift through projects and choose those that are designed both to advance AI and assure that such advances would be accompanied by appropriate safeguards." McGinnis feels that peer review is better "than regulation to address technical issues that are not possible to capture through bureaucratic mandates". McGinnis notes that his proposal stands in contrast to that of the Machine Intelligence Research Institute, which generally aims to avoid government involvement in friendly AI.

According to Gary Marcus, the annual amount of money being spent on developing machine morality is tiny.

Criticism

Some critics believe that both human-level AI and superintelligence are unlikely, and that therefore friendly AI is unlikely. Writing in *The Guardian*, Alan Winfeld compares human-level artificial intelligence with faster-than-light travel in terms of difficulty, and states that while we need to be "cautious and prepared" given the stakes involved, we "don't need to be obsessing" about the risks of superintelligence.

Some philosophers claim that any truly "rational" agent, whether artificial or human, will naturally be benevolent; in this view, deliberate safeguards designed to produce a friendly AI could be unnecessary or even harmful. Other critics question whether it is possible for an artificial intelligence to be friendly. Adam Keiper and Ari N. Schulman, editors of the technology journal *The New Atlantis*, say that it will be impossible to ever guarantee "friendly" behavior in AIs because problems of ethical complexity will not yield to software advances or increases in computing power. They write that the criteria upon which friendly AI theories are based work "only when one has not only great powers of prediction about the likelihood of myriad possible outcomes, but certainty and consensus on how one values the different outcomes.

3.4 Superintelligence

A **superintelligence**, **hyperintelligence**, or **superhuman intelligence** is a hypothetical agent that possesses intelligence far surpassing that of the brightest and most gifted human minds. ‘’Superintelligence’’ may also refer to the form or degree of intelligence possessed by such an agent.

Technological forecasters and researchers disagree about when human intelligence is likely to be surpassed. Some argue that advances in artificial intelligence (AI) will probably result in general reasoning systems that lack human cognitive limitations. Others believe that humans will evolve or directly modify their biology so as to achieve radically greater intelligence. A number of futures studies scenarios combine elements from both of these possibilities, suggesting that humans are likely to interface with computers, or upload their minds to computers, in a way that enables substantial intelligence amplification.

Experts in AI and biotechnology do not expect any of these technologies to produce a superintelligence in the very near future. A number of scientists and forecasters argue for prioritizing early research into the possible benefits and risks of human and machine cognitive enhancement, because of the potential social impact of such technologies.

Definition

Summarizing the views of intelligence researchers, Linda Gottfredson writes:

Intelligence is a very general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly, and learn from experience. It is not merely book-learning, a narrow academic skill, or test-taking smarts. Rather, it reflects a broader and deeper capability for comprehending our surroundings "catching on," "making sense" of things, or "figuring out" what to do.

There is no agreed-upon way to measure intelligence in all varieties of agent. Intelligence quotient (IQ) tests are used to measure normal human variation in *g* factor, a general skill at cognitive tasks. In machines, one of the oldest operationalizations of intelligence is the Turing test, which judges a system's intelligence by how well it can fool a human interrogator into thinking it is human. However, IQ and Turing tests both focus on ordinary human ability levels; neither extends to provide a definition or measure of superhuman intelligence.

Shane Legg and Marcus Hutter make use of a more abstract definition of *intelligence*, as "an agent's ability to achieve goals in a wide range of environments". On this view, matching or surpassing human-level intelligence is a matter of being able to complete tasks and solve problems in many different domains, regardless of how or why one goes about doing so. "Intelligence is not really the ability to do anything in particular, rather it is a very general ability that affects many kinds of performance." Legg and Hutter argue that this approach makes it possible to define measures of intelligence that are less narrow and human-specific, such as their 'universal intelligence measure', which culminates in the formal agent AIXI.

Oxford futurist Nick Bostrom defines *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." The program Fritz falls short of superintelligence even though it is much better than humans at chess, because Fritz cannot outperform humans in other tasks. Following Hutter and Legg, Bostrom treats superintelligence as general dominance at goal-oriented behavior, leaving open whether an artificial or human superintelligence would possess capacities such as intentionality (cf. the Chinese room argument) or first-person consciousness (cf. the hard problem of consciousness).

Feasibility

Whether superhuman intelligence is possible depends on the feasibility of the particular methods for developing it (see next section), but also on whether humans fall short on various cognitive metrics, such as computational efficiency and speed. Large deficiencies in human thought suggest that more powerful reasoning systems are physically possible.

Bostrom writes, "Biological neurons operate at a peak speed of about 200 Hz, a full seven orders of magnitude slower than a modern microprocessor (~2 GHz)." Moreover, neurons transmit spike signals across axons at no greater than 120 m/s, "whereas existing electronic processing cores can communicate optically at the speed of light". Thus, the simplest example of a superintelligence may be an emulated human mind that's run on much faster hardware than the brain. A human-like reasoner that could think millions of times faster than current humans would have a dominant advantage in most reasoning tasks, particularly ones that require haste or long strings of actions.

Computational resources place another limit on present-day human cognition. A non-human (or modified human) brain could become much larger, like many supercomputers. Bostrom also raises the possibility of *collective superintelligence*: a large enough number of separate reasoning systems, if they communicated and coordinated well enough, could act in aggregate with far greater capabilities than any sub-agent.

There may also be ways to *qualitatively* improve on human reasoning and decision-making. Humans appear to differ from chimpanzees in the ways we think more than we differ in brain size or speed. Humans outperform non-human animals in large part because of new or enhanced reasoning capacities, such as long-term planning and language use. (See evolution of human intelligence and primate cognition.) If there are other possible improvements to human reasoning that would have a similarly large impact, this makes it likelier that an agent can be built that outperforms humans in the same fashion humans outperform chimpanzees. All of the above advantages hold for artificial superintelligence, but it is not clear how many hold for biological superintelligence. Physiological constraints limit the speed and size of biological brains in many ways that are inapplicable to machine intelligence. As such, writers on superintelligence have devoted much more attention to superintelligent AI scenarios.

Artificial superintelligence

Most surveyed AI researchers expect machines to eventually be able to rival humans in intelligence, though there is little consensus on timescales. At the 2006 AI@50 conference, 18% of attendees reported expecting machines to be able "to simulate learning and every other aspect of human intelligence" by 2056; 41% of attendees expected this to happen sometime after 2056; and 41% expected machines to never reach that milestone. In a survey of the 100 most cited authors in AI (as of May 2013, according to Microsoft Academic Search), the median year by which respondents expected machines "that can carry out most human professions at least as well as a typical human" (assuming no global catastrophe occurs) with 10% confidence is 2024 (mean 2034, st. dev. 33 years), with 50% confidence is 2050 (mean 2072, st. dev. 110 years), and with 90% confidence is 2070 (mean 2168, st. dev. 342 years). These estimates exclude the 1.2% of respondents who said no year would ever reach 10% confidence, the 4.1% who said 'never' for 50% confidence, and the 16.5% who said 'never' for 90% confidence. Respondents assigned a median 50% probability to the possibility that machine superintelligence will be invented within 30 years of the invention of approximately human-level machine intelligence.

Philosopher David Chalmers argues that generally intelligent AI — artificial general intelligence — is a very likely path to superhuman intelligence. Chalmers breaks this claim down into an argument that AI can achieve *equivalence* to human intelligence, that it can be *extended* to surpass human intelligence, and that it be further *amplified* to completely dominate humans across arbitrary tasks.

Concerning human-level equivalence, Chalmers argues that the human brain is a mechanical system, and therefore ought to be emulable by synthetic materials. He also notes that human intelligence was able to biologically evolve, making it more likely that human engineers will be able to recapitulate this invention. Evolutionary algorithms in particular should be able to produce human-level AI. Concerning intelligence extension and amplification, Chalmers argues that new AI technologies can generally be improved on, and that this is particularly likely when the invention can assist in designing new technologies.