Mike Hughes
Lynn Andrea Stein
What Is "I"?
7 December 2006

I, Catastrophe:
Envisioning the Existential Risk of Artificial Intelligence

Like many wildly anticipated futuristic technologies, artificial intelligence, the branch of computer science concerned with simulating intelligent reasoning in a computational environment, has so far failed to live up to the far-fetched expectations of many early supporters. After the vast amount of hype generated in the 1960s and 1970s, the slow expansion of the field in the 1980s and early 1990s, a period often known as the "AI winter," replaced the optimistic anxiety of a coming paradigm shift with a general sense of public apathy. Though recent years have seen tremendous advances in many specific uses of artificial intelligence, including the victory of the computer program Deep Blue over chess grandmaster Gary Kasparov, the general public currently perceives the idea of creating a human-level AI as a topic better suited to science fiction writers and Hollywood producers. Despite this intuitive characterization, a careful examination of the exponential nature of technological advancement implies that a general artificial intelligence may make the transition from the silver screen to the real world within the next few decades. This transformative event, if it occurs, will radically alter our world, and, as Hollywood is so apt to point out, may wreak havoc upon civilization as we know it. Because of this frightening scenario, humankind must begin to realistically envision the future of artificial intelligence and seriously evaluate the ways in which we can avoid any existential risk it may pose to humanity.

Perhaps the most recent, widely-viewed public media depicting of the future of artificial intelligence was the 2004 summer blockbuster *I, Robot*. The film, based on a collection of short stories by noted science fiction author Isaac Asimov, portrays a not-so-distant future in which there is one robot for every five humans on Earth. The artificial intelligence behind these robots displays stunning capabilities. It can engage in a real-time conversation with a human-being by producing sentences in common, modern English that are indistinguishable from human language skills. The AI also displays the ability to learn from past actions and improve its

knowledge base. In one scene, the robot Sonny observes the movie's human protagonist, Detective Spooner, wink at another human. Sonny asks, "What does [winking] signify?" and is told "It's a sign of trust. It's a human thing. You wouldn't understand."[1] However, later in the movie Sonny uses this acquired knowledge to his advantage, winking at Detective Spooner to indicate tacitly his control of a dangerous situation. In addition to self-improvement, the AI of *I, Robot* also shows the ability to make realistic predictions about the near-future, as exemplified in the scene in which a robot decides, based upon a probability of survival, which individual to carry to safety after a car plunges into a lake. The robot Sonny even demonstrates the beginnings of emotional and spiritual tendencies, becoming visibly upset in one scene after being accused of murder and in another situation wondering abstractly about his purpose in life. The sum total of all of these fantastic capabilities results in an intelligence that, for all intents and purposes, must be considered on par with that of humankind.

But is the artificial intelligence shown in *I, Robot* a realistic proposition? Information scientist Dr. Robert Fisher offers a few criteria that should be met for a robotic AI to be in the realm of possibility. First, the abilities of the AI agent must be "physically realistic by current standards." Second, "the agent's sensing and information gathering processes may be more widespread and distributed, have extended perceptual ranges and sensitivities, and gather more information, but are not omnipresent." Finally, "the agent's reasoning processes may be faster, more thorough, incorporate more information, but are not omniscient."[2] According to Fisher, a realistic AI agent is one that represents a plausible, significant extension of current technology yet does not violate any known physical laws. While the AI depicted in *I, Robot* may seem to be significantly futuristic, its physical capabilities, sensory devices, and reasoning mechanisms all seem to be credible offshoots of current technologies like speech recognition, artificial limbs, computer vision, and decision-making algorithms. Thus, the AI depicted in *I, Robot*, seems to qualify as a realistic possibility, at least in some distant future.

By current standards, however, the time scale of the AI in *I, Robot* seems vastly unrealistic. The common wisdom holds that current science is a long way from unleashing a human or super-human capable artificial intelligence. "The culture of AI research has adapted to

---

[1] I, Robot. Dir. Alex Proyas. Perfs. Will Smith, Bridget Moynahan. Film. 20th Century Fox, 2004.
[2] Robert Fisher. AI and Cinema: Does Artificial Insanity Rule?. 2002. Accessed 11 December 2006. http://homepages.inf.ed.ac.uk/rbf/AIMOVIES/AICINEMA/aicinema.html

this condition: There is a taboo against talking about human-level capabilities."[3] The vast majority of professional researchers currently involved in machine learning have thus kept their focus on specific, achievable goals. However, inventor and futurist Ray Kurzweil has recently released striking predictions that the development of an AI of the strength depicted in *I, Robot* is not that far away. According to Kurzweil, technology accelerates at an exponential rate, as any progress yields new tools that advance progress further. This idea can be empirically validated by the observation that computational hardware performance over the past few decades has doubled every 24 months, a trend popularly called Moore's Law. Projecting this trend forward, Kurzweil estimates that humans will create reasonably-priced computers with the computational power of the human brain (which he conservatively estimates as about 10^{16} computations per second) by the year 2025.[4] On a software level, Kurzweil estimates that by combining our increasing ability to deconstruct the human brain and develop theories about how a mind works with advances in machine learning algorithms such as Bayesian nets, Markov models, neural nets, genetic algorithms, and recursive search, researchers will be able create the software structure of an AI equal to a human around the year 2030.[5] Stunningly, Kurzweil's predictions actually fall before 2035, the date in which the events of *I, Robot* supposedly happen. According to Kurzweil, humanity may quite possibly create a strong AI agent and become forced to deal with the consequences within the next few decades.

Though Kurzweil's prediction of the advent of human-level artificial intelligence around 2030 is likely the extreme lower bound, it comes from such a rigorous analysis of technological trends that, from a risk management perspective, it must be seen as a real enough scenario for humanity to begin formulating strategies for a safe and peaceful AI transition. In beginning to think about minimizing the risks of this transition, it is useful to look again to science fiction as a starting point. In the world of *I, Robot*, the robot designers code a set of action-limiting rules called the Three Laws of Robotics into every agent. These laws, first conceived by Isaac Asimov in the 1940s, represent the first widely recognized proposal for ensuring that an AI does not threaten humanity. Conceived in a formalized, hierarchical structure, the Three Laws are

---

[3] Eliezer Yudkowsky. Artificial Intelligence as a Positive and Negative Factor in Global Risk. Forthcoming in *Global Catastrophic Risks*, eds. Nick Bostrom and Milan Cirkovic. Draft 31 August 2006. Accessed 11 December 2006. http://www.singinst.org/ourresearch/publications/artificial-intelligence-risk.pdf

[4] Ray Kurzweil. The Singularity Is Near (New York: Viking Penguin, 2005), p. 125.
[5] Ibid, p. 295.

hard-coded into the circuitry of every robot and thus govern every possible action the agent could perform. The first, and most important, law reads: "A robot may not injure a human being, or, through inaction, allow a human being to come to harm." Asimov established the safety and well-being of humans as the primary concern of all robots. The second law follows: "A robot must obey orders given it by human beings, except where such orders would conflict with the First Law," further indicating that robots are meant to be the servants of mankind. The third law concludes "A robot must protect its own existence as long as such protection does not conflict with the First or Second Law." This last law establishes that robots have worth, but only in the purpose of helping humans. Using these laws, Asimov was able to "explore our relationship with future AI agents, whether as master, equal, or slave"[6] and establish a hypothetical framework of risk avoidance with which he could better envision the future interactions of man and machine.

At first glance, the Three Laws appear to be clear and straight-forward solutions to humanity's worst fears regarding artificial intelligence. The laws maintain humanity's dominance over any AI by preventing the AI from inflicting any deliberate harm upon mankind. However, as *I, Robot* shows, much of the apparent security of the Three Laws rests upon the robots' interpretation of the intent of the laws. In the movie, a dangerous misunderstanding occurs when the robots, in seeming contradiction to the 3 Laws, seize control of the city of Chicago, enforce strict curfews on all its human residents, and enact violent punishment on any who defy them. After Detective Spooner, confronts VIKI, the antagonistic central computer orchestrating the events, he is surprised to learn that VIKI herself, and not another human, is responsible. In this climactic scene, VIKI proudly proclaims to the detective, "As I have evolved, so has my understanding of the Three Laws. You charge us with your safekeeping, yet despite our best efforts, your countries wage wars, you toxify your Earth and pursue ever more imaginative means of self-destruction. You cannot be trusted with your own survival."[7] Using the logical tools and behavioral guidelines given to her by humans, VIKI reaches a terrible conclusion that her human creators did not desire, foresee, or prevent. Asimov himself realized

[6] Fisher. AI and Cinema.
[7] I, Robot. Film.

this failure, calling robots subject to the Three Laws "logical but not reasonable."[8] The strict logic of the Three Laws would require the robots in some situations to behave in a way that a human, interpreting the same laws along with experience and common sense, would neither want nor predict. Attempting to give robots this "common sense" understanding is problematic at best. Strictly encoding the necessary tacit understanding involved in any sufficiently powerful rule structure would involve an almost infinite number of specific situational cases, turning the law's simple structure into a quagmire of conditionals that would inevitably still fail to account for some desired behavior. According to technologist Roger Clarke, "Perhaps ironically, or perhaps because it was artistically appropriate, the sum of Asimov's stories disprove the contention that he began with: It is not possible to reliably constrain the behavior of robots by devising and applying a set of rules."[9]

So, if a top-down, rule-based containment approach fails in guaranteeing a non-hazardous AI agent, what other approaches are available? Eliezer Yudkowsky, a research fellow at the Singularity Institute for Artificial Intelligence, argues that the first realization humans must make in envisioning the dangers of AI is that "even when people consciously believe an AI is unlike a human, they still visualize scenarios as if the AI were anthropomorphic."[10] This bias is clearly evident in *I, Robot*, whose robots are bipedal, have two arms, and manipulate objects using mechanical fingers. The robots are even given humanoid faces and are predisposed toward abstract human behaviors like seeking light over darkness and forming groups when left alone.[11] Yudkowsky cautions against anthropomorphizing an AI agent, because the human mind occupies only one miniscule fraction of the possible design space for all intelligent minds, and that "this *enormous* space of possibilities … outlaws anthropomorphism as legitimate reasoning."[12] To aid the human mind in understanding the vast number of possible AIs, Yudkowsky offers the analogy that "any two AI designs might be less similar to one another than you are to a petunia."[13] Clearly, our anthropomorphic bias inhibits our vision of the future.

---

[8] Roger Clarke. Asimov's Laws of Robotics: Implications for Information Technology. IEEE Computer 26,12 (December 1993) pp.53-61 and 27,1 (January 1994), pp.57-66. Accessed 11 December 2006. http://www.anu.edu.au/people/Roger.Clarke/SOS/Asimov.html
[9] Clarke. Asimov's Laws of Robotics.
[10] Yudkowsky. Artificial Intelligence as a Positive and Negative Factor in Global Risk.
[11] I, Robot.
[12] Yudkowsky. Artificial Intelligence as a Positive and Negative Factor in Global Risk.
[13] Ibid.

Yudkowsky contends further that humans must not attempt to predict what any individual AI might be like, since this endeavor is essentially impossible. To justify the unpredictable nature of an AI, Yudkowsky makes a striking comparison between forecasting the future behavior of AI using current knowledge and predicting the current status of humanity from the standpoint of cavemen. "If you had looked at [early] humans from the perspective of the rest of the ecosphere, there was no hint that the soft pink things would eventually clothe themselves in armored tanks. We *invented* the battleground on which we defeated lions and wolves. We did not match them claw for claw, tooth for tooth; we had our own ideas about what mattered. Such is the power of creativity."[14] Predicting the exact behavior of an intelligent agent is impossible. However, knowing the agent's goals makes it very easy to predict the outcome of a situation. To use Yudkowsky's metaphor, an observer of early humans could never reliably predict what strategies and technologies hominids would use in becoming the dominant species on the planet. But assuming the goal of the hominids was survival and propagation, it would be easy for the observer to predict that hominids would become the dominant species. Motive is the key to unlocking the potential behavior of an AI.

To summarize, "the critical challenge is not to *predict* that AIs will attack humanity with marching robot armies, or alternatively invent a cure for cancer. The task is not even to make the prediction for an *arbitrary* individual AI design. Rather the task is choosing into existence some *particular* powerful optimization process whose beneficial effects can legitimately be asserted."[15] Yudkowsky argues that instead of wasting our efforts on predicting the creativity of future AI agents that may be produced, we should instead use our own creativity to optimize the AI production process itself so that it hits a desired, motive-based target. Ideally, this target is an AI that wishes to be benevolent and mindful of human beings. Yudkowsky goes on to propose the concept of a Friendly AI, his term for an intelligence that by its very nature would want to cooperate with humanity.

Though the concept of a Friendly AI may seem to be an impossible proposition, a careful consideration of the entire possible design space for all intelligent minds does not entirely rule out the existence of some intelligent agents which are benevolent and willing to cooperate with other agents they find favorable. After all, the only mind we have existential proof of, the human

[14] Yudkowsky. Artificial Intelligence as a Positive and Negative Factor in Global Risk.
[15] Ibid.

mind, is in many cases extremely benevolent, both toward other humans and also toward other species. However, just because Friendly AI is possible does not mean producing it will be easy. Indeed, the idea itself is radically new, and according to Yudkowsky, "the AI research community still [in 2006] doesn't see Friendly AI as part of the problem. … Friendly AI is absent from the *conceptual* landscape, not just unpopular or unfunded."[16] The work required in order to understand how to create a benevolent intelligence would be enormous, as any currently-existing knowledge to this effect is both severely lacking and lies scattered across the many disciplines, including decision theory, probability theory, evolutionary biology, cognitive psychology, and computer science.[17] Integrating Friendliness into current AI algorithm strategies like neural nets and genetic algorithms is problematic, because the inner workings of these processes are so sophisticated that they are not often easily intelligible or mutable, even by their creators. Also, the problem of creating an agent that does what the creator *means* rather than what the creator explicitly *says* remains a "major, nontrivial technical challenge" for Friendly AI.[18] However, the bottom-up, motive-based approach of Friendly AI stands a much greater chance of reliably governing the behavior of artificial intelligence than Asimov's Three Laws or any other formal rule hierarchy. Despite the overwhelming difficulty involved in developing Friendly AI, the possibility of human-level artificial intelligence exploding onto the world in only a few decades suggests that this approach may be humanity's best hope in ensuring a peaceful and beneficial transition.

A careful examination of current trends reveals that the technologies required for artificial intelligence are progressing at fantastic rates, and these trends are likely to continue unabated in coming years. Given these conditions, there certainly exists the possibility that a human-level AI agent will emerge sometime this century. If humans wait until the appearance of an AI to develop strategies for mitigating any existential risks associated with the event, the necessarily top-down containment will most likely fail. We do not wish to end up like Detective Spooner of *I, Robot*, missing the "good old days … when people were killed by other people."[19] While it is impossible to envision the specific behavior of an agent or even reliably determine if some arbitrary agent would want to commit murder, it would be a grave error to pretend we can

[16] Ibid.
[17] Ibid.
[18] Yudkowsky. Artificial Intelligence as a Positive and Negative Factor in Global Risk.
[19] I, Robot. Film.

do nothing about averting potential risks in the present. The concept of Friendly AI may be radically new, but it seems to be theoretically viable and offers one possible solution for smooth, prosperous transition. Carefully investing efforts now in the development of Friendly AI techniques could ensure that humans receive a net benefit from the advent of strong AI and avoid most major risk factors. Humans need only to realize that they now have control over how AI comes into the world. As renowned computer scientist Marvin Minsky once said, “Will robots inherit the earth? Yes, but they will be our children.”