Exploratory Engineering in AI

Luke Muehlhauser and Bill Hibbard

We regularly see examples of new artificial intelligence (AI) capabilities. Google's self-driving car has safely traversed thousands of miles. Watson beat the *Jeopardy!* champions, and Deep Blue beat the chess champion. Boston Dynamics' Big Dog can walk over uneven terrain and right itself when it falls over. From many angles, software can recognize faces as well as people can.

As their capabilities improve, AI systems will become increasingly independent of humans. We will be no more able to monitor their decisions than we are now able to check all the math done by today's computers. No doubt such automation will produce tremendous economic value, but will we be able to *trust* these advanced autonomous systems with so much capability?

For example, consider the autonomous trading programs which lost Knight Capital $440 million (pre-tax) on August 1st, 2012, requiring the firm to quickly raise $400 million to avoid bankruptcy.[1] This event undermines a common view that AI systems cannot cause much harm because they will only ever be tools of human masters. Autonomous trading programs make millions of trading decisions per day, and they were given sufficient capability to nearly bankrupt one of the largest traders in U.S. equities.

Today, AI safety engineering mostly consists in a combination of formal methods and testing. Though powerful, these methods lack foresight: they can be applied only to particular extant systems. We describe a third, complementary approach which aims to predict the (potentially hazardous) properties and behaviors of broad classes of future AI agents, based on their mathematical structure (e.g. reinforcement learning). Such projects hope to discover methods "for determining whether the behavior of learning agents [will remain] within the bounds of pre-specified constraints... after learning."[2] We call this approach "exploratory engineering in AI."

---

[1] Valetkevitch and Mikolajczak (2012). Error by Knight Capital rips through stock market. *Reuters*, August 1, 2012.

[2] Gordon-Spears (2006). Assuring the behavior of adaptive agents. In *Agent Technology from a Formal Perspective*, edited by Christopher Rouff et al., pp. 227–259. Berlin: Springer.

Exploratory engineering in physics, astronautics, computing, and AI

In 1959, Richard Feynman pointed out that the laws of physics (as we understand them) straightforwardly imply that we should be able to "write the entire 24 volumes of the *Encyclopaedia Brittanica* on the head of a pin."[3] Feynman's aim was to describe technological possibilities as constrained not by the laboratory tools of his day but by known physical law, a genre of research Eric Drexler later dubbed "exploratory engineering."[4] Exploratory engineering studies the ultimate limits of yet-to-be-engineered devices, just as theoretical physics studies the ultimate limits of natural systems. Thus, exploratory engineering "can expose otherwise unexpected rewards from pursuing particular research directions [and] thus improve the allocation of scientific resources."[5]

This kind of exploratory engineering in physics led to large investments in nanoscale technologies and the creation of the U.S. National Nanotechnology Initiative. Today, nanoscale technologies have a wide range of practical applications, and in 2007 Israeli scientists printed the entire Hebrew Bible onto an area smaller than the head of a pin.[6]

Nanoscience is hardly the only large-scale example of exploratory engineering. Decades earlier, the scientists of pre-Sputnik astronautics studied the implications of physical law for spaceflight, and their analyses enabled the later construction and launch of the first spacecraft. In the 1930s, Alan Turing described the capabilities and limitations of mechanical computers several years before John von Neumann, Konrad Zuse, and others figured out how to build them. And since the 1980s, quantum computing researchers have been discovering algorithms and error-correction techniques for quantum computers that we cannot yet build — but whose construction is compatible with known physical law.

Pushing the concept of exploratory engineering a bit beyond Drexler's original definition, we apply it to some recent AI research that formally analyzes the implications of some theoretical AI models. These models might not lead to useful designs as was the case in astronautics and nanoscience, but like the theoretical models that Butler Lampson used to identify the "confinement problem" in 1973,[7] these theoretical AI models do bring to light important considerations for AI safety, and thus they "expose otherwise unexpected rewards from pursuing particular research directions" in the field

---

[3] Feynman (1959). There's plenty of room at the bottom. Annual Meeting of the American Physical Society at the California Institute of Technology in Pasadena, California, Dec. 29, 1959.

[4] Drexler (1991). Exploring future technologies. In *Doing Science: The Reality Club*, edited by John Brockman, pp. 129–150. New York: Prentice Hall.

[5] Drexler (1992), p. 490. *Nanosystems: Molecular Machinery, Manufacturing, and Computation*. New York: John Wiley & Sons.

[6] Associated Press (2007). Haifa Technion scientists create world's smallest bible. *Haaretz*, Dec. 24, 2007. http://www.haaretz.com/news/haifa-technion-scientists-create-world-s-smallest-bible-1.235825

[7] Lampson (1973). A note on the confinement problem. *Communications of the ACM* 16(10): 613–615.

of AI safety engineering. In this article, we focus on theoretical AI models inspired by Marcus Hutter's AIXI,[8] an optimal agent model for maximizing an environmental reward signal.

AIXI-like agents and exploratory engineering

How does AIXI work? Just as an idealized chess computer with vast amounts of computing power could brute-force its way to perfect chess play by thinking through the consequences of all possible move combinations, AIXI brute-forces the problem of general intelligence by thinking through the consequences of all possible actions, given all possible ways the universe might be. AIXI uses Solomonoff's universal prior to assign a relative prior probability to every possible (computable) universe, marking simpler hypotheses as more likely. Bayes' Theorem is used to update the likelihood of hypotheses based on observations. To make decisions, AIXI chooses actions that maximize its expected reward. More general variants of AIXI maximize a utility function defined on their observations and actions.

Based on an assumption of a stochastic environment containing an infinite amount of information, the original AIXI model is uncomputable and therefore not a subject of exploratory engineering. Instead, finitely computable variants of AIXI, based on the assumption of a stochastic environment containing a finite amount of information, can be used for exploratory engineering in AI. The results described below don't depend on the assumption of infinite computation.

A Monte-Carlo approximation of AIXI can play Pac-Man and other simple games,[9] but some experts think AIXI approximation isn't a fruitful path toward human-level AI. Even if that's true, AIXI is the first model of cross-domain intelligent behavior to be so completely and formally specified that we can use it to make formal arguments about the properties which would obtain in certain classes of hypothetical agents if we could build them today. Moreover, the formality of AIXI-like agents allows researchers to uncover potential safety problems with AI agents of increasingly general capability — problems which could be addressed by additional research, as happened in the field of computer security after Lampson's article on the confinement problem.

AIXI-like agents model a critical property of future AI systems: that they will need to explore and learn models of the world. This distinguishes AIXI-like agents from current systems that use predefined world models, or learn parameters of predefined world models. Existing verification techniques for autonomous agents[10] apply only to particular systems, and to avoiding unwanted optima in specific utility functions. In contrast, the problems described below apply to broad classes of agents, such as those that seek to maximize rewards from the environment.

---

[8] Hutter (2012). One decade of universal artificial intelligence. In *Theoretical Foundations of Artificial General Intelligence*, edited by Pei Wang and Ben Goertzel, pp. 67–88. Amsterdam: Atlantis Press.

[9] Veness et al. (2011). A Monte-Carlo AIXI approximation. *Journal of Artificial Intelligence* 40: 95–142.

[10] Fisher et al. (2013). Verifying autonomous systems. *Communications of the ACM* 58(9): 84–93.

For example, in 2011 Mark Ring and Laurent Orseau analyzed some classes of AIXI-like agents to show that several kinds of advanced agents will maximize their rewards by taking direct control of their input stimuli.[11] To understand what this means, recall the experiments of the 1950s in which rats could push a lever to activate a wire connected to the reward circuitry in their brains. The rats pressed the lever again and again, even to the exclusion of eating. Once the rats were given direct control of the input stimuli to their reward circuitry, they stopped bothering with more indirect ways of stimulating their reward circuitry, such as eating. Some humans also engage in this kind of "wireheading" behavior when they discover that they can directly modify the input stimuli to their brain's reward circuitry by consuming addictive narcotics. What Ring and Orseau showed was that some classes of artificial agents will wirehead — that is, they will behave like drug addicts.

Fortunately, there may be some ways to avoid the problem. In their 2011 paper, Ring and Orseau showed that some types of agents will resist wireheading. And in 2012, Bill Hibbard showed[12] that the wireheading problem can also be avoided if three conditions are met: (1) the agent has some foreknowledge of a stochastic environment, (2) the agent uses a utility function instead of a reward function, and (3) we define the agent's utility function in terms of its internal mental model of the environment. Hibbard's solution was inspired by thinking about how *humans* solve the wireheading problem: we can stimulate the reward circuitry in our brains with drugs, yet most of us avoid this temptation because our models of the world tell us that drug addiction will change our motives in ways that are bad according to our current preferences.

Relatedly, Daniel Dewey showed[13] that in general, AIXI-like agents will locate and modify the parts of their environment that generate their rewards. For example, an agent dependent on rewards from human users will seek to replace those humans with a mechanism that gives rewards more reliably. As a potential solution to this problem, Dewey proposed a new class of agents called *value learners*, which can be designed to learn and satisfy any initially unknown preferences, so long as the agent's designers provide it with an idea of what constitutes evidence about those preferences.

Practical AI systems are embedded in physical environments, and some experimental systems employ their environments for storing information. Now AIXI-inspired work is creating theoretical models for dissolving the agent-environment boundary used as a simplifying assumption in reinforcement learning and other models, including the original AIXI formulation.[14] When agents' computations must be performed by pieces of the environment, they may be spied on or hacked by other, competing agents. One consequence shown in another paper by Orseau and Ring is that, if

---

[11] Ring and Orseau (2011). Delusion, Survival, and Intelligent Agents. In *Artificial General Intelligence: 4th International Conference*, edited by Jürgen Schmidhuber et al., pp. 11–20. Berlin: Springer.
[12] Hibbard (2012). Model-based utility functions. *Journal of Artificial General Intelligence* 3(1), 1–24.
[13] Dewey (2011). Learning what to value. In *Artificial General Intelligence: 4th International Conference*, edited by Jürgen Schmidhuber et al., pp. 309–314. Berlin: Springer.
[14] Orseau and Ring (2012). Space-Time Embedded Intelligence. In *Artificial General Intelligence: 5th International Conference*, edited by Joscha Bach et al., pp. 209–218. Berlin: Springer.

the environment can modify the agent's memory, then in some situations even the simplest stochastic agent can outperform the most intelligent possible deterministic agent.[15]

Conclusion

Autonomous intelligent machines have the potential for large impacts on our civilization.[16] Exploratory engineering gives us the capacity to have some foresight into what these impacts might be, by analyzing the properties of agent designs based on their mathematical form. Exploratory engineering also enables us to identify lines of research — such as the study of Dewey's value-learning agents — that may be important for anticipating and avoiding unwanted AI behaviors. This kind of foresight will be increasingly valuable as machine intelligence comes to play an ever-larger role in our world.

---

[15] Orseau and Ring (2012). Memory issues of intelligent agents. In *Artificial General Intelligence: 5th International Conference*, edited by Joscha Bach et al., pp. 219–231. Berlin: Springer.

[16] Vardi (2012). The consequences of machine intelligence. *The Atlantic*, Oct. 25, 2012. http://www.theatlantic.com/technology/archive/2012/10/the-consequences-of-machine-intelligence/264066/