

Omohundro's "Basic AI Drives" and Catastrophic Risks

1. Introduction

The idea of machine intelligences causing human extinction is a science-fictional cliché, but that status does not preclude a real risk (Bostrom, 2002; Chalmers, 2010; Friedman, 2008; Hall, 2007; Kurzweil, 2005; Moravec, 1999; Posner, 2004; Rees, 2003; Yudkowsky, 2008). At the 2008 Global Catastrophic Risk Conference at Oxford University, participants from diverse fields were surveyed on their estimates of catastrophic risks through the year 2100: among those responding, the median estimate of human extinction risk due to AI was 5%, with the median estimate of risk of human extinction for any reason at 19% (Sandberg & Bostrom, 2008). Such a large impact, even if unlikely to occur, and even if distant in time, deserves some attention (Bostrom, 2002; Bostrom, 2003; Matheny, 2007; Parfit, 1984; Posner, 2004; Sandel, 2006; Weitzman, 2009). Unfortunately, it is very difficult to make predictions about the characteristics and behavior of future AI systems before they are invented, without knowledge of their particular architectures or decision algorithms.

One approach in the face of such uncertainty is to focus on very 'generic' properties shared by diverse possible intelligent systems, such as evolutionary theory (Bostrom, 2005) or decision theory. Omohundro (2008) takes this route, using standard decision theory to argue for the likelihood of several "basic AI drives," behavioral tendencies that would advance diverse goals, other things being equal. Particularly relevant are the "survival drive" and "resource drive." The survival drive stems from the fact that an agent will tend to act to promote its own goals, and so self-preservation will be instrumentally useful, even without any "survival instinct" or non-instrumental desire to persist. Thus, we could expect diverse AIs to avoid destruction unless doing so would cause other outweighing losses in terms of their goals. The resource drive is a consequence of the existence of fungible resources, e.g. free energy, that can be applied to a wide variety of ends. Similarly, money is at least somewhat useful for almost any human goals today, from raising children to winning an Olympic medal to promoting vegetarianism. So, other things being equal, AI systems would tend to act to acquire fungible resources to use in pursuit of their goals.

For very weak AI systems, these drives are good news: even AIs quite indifferent to humans would tend to perform as demanded if humans controlled all the means to attain their goals. For very powerful AI systems, able to act without concern for human interference or retaliation, these drives would be quite threatening: free energy and other resources required for human survival could be diverted from humanity to the AI's purposes, e.g. to power computers or other machinery, resulting in human extinction unless the AI preferred sustaining human life to any alternative use of those resources (Yudkowsky, 2008). However, the analysis does not make such clear predictions for intermediate cases, in which AIs face risk of failure or retaliation in a conflict with humanity. In such cases, a decision to initiate conflict would depend on the relative valuation of the gains of victory, the status quo, and the costs of failure.

This paper applies Omohundro's (2008) general framework to those intermediate cases, in which AI systems are initially weak, but can pursue risky strategies to gain extreme power.

2. Considering conflict under uncertainty

For simplicity, we can crudely model an AI's decisionmaking about conflict with humans as a one-time choice to attempt to acquire power or to cooperate with human requests, including requests to help devise safeguards against AI subversion. In our simple model, an attempt to acquire power either wholly succeeds and gives the AI unilateral control over all Earth-accessible resources and immunity to subsequent human attack, or fails entirely.

If the AI initiates aggression, it expects to succeed with probability p , receiving expected utility $EU_{Success}$, and to fail with probability $(1-p)$, receiving expected utility $EU_{Failure}$. If it refrains from conflict, then it will receive $EU_{Cooperation}$. The AI will then attempt aggression if:

$$p EU_{Success} + (1-p) EU_{Failure} > EU_{Cooperation}.$$
¹

Note that these values are the probabilities and expected utilities assessed by the AI; whether the AI's assessments are right or wrong, it is necessarily its own estimates of the situation that will determine its choice. Thus, if we wished to alter the value of p , we might implement additional safeguards and make them known to the AI, or we might merely attempt to deceive it with false claims of such safeguards (but see Chalmers, 2010, for the difficulties of deceiving superhumanly intelligent systems).

This model could describe the decision of a unique early AI, or that of a single AI among many, e.g. one considering whether to attempt to instigate a collective AI effort to expropriate humans. Or it might describe the collective choice of a group of copies of an AI program with identical goals, which we might think of either as a group of AIs or a single superorganism.

In this stylized decision problem, AI aggression becomes more likely with higher values of $EU_{Success}$ and $EU_{Failure}$, and lower values of $EU_{Cooperation}$. The convergent instrumental 'drives' discussed by Omohundro have their impact by affecting these values.

3. Why consider 'generic' motivations?

The basic AI drives are generally useful subgoals of diverse aims, but human designers will attempt to produce AIs that reliably do what they are requested. Why not assume that engineers will produce the first powerful AIs with utility functions such that cooperation is always preferred to conflict, i.e. such that $EU_{Cooperation}$ always exceeds both $EU_{Success}$ and $EU_{Failure}$? AI systems might be constructed to assign substantial disutility to harm to existing humans (appropriately defined), to engaging in unsupervised self-modification,

¹ If $p EU_{Success} + (1-p) EU_{Failure} = EU_{Cooperation}$, the AI will be indifferent between aggression and cooperation.

to producing proxy systems to circumvent the previous restriction, to persuading humans to produce proxy systems to circumvent that restriction, and so on (Omohundro, 2008). Or it might be designed to share the collective aims of its creators, so that it would use even arbitrary power in human-desirable ways (Bostrom, 2006; Yudkowsky, 2008).

If done well, this approach would seem ideal, but it faces a major challenge in the specification of the relevant concepts, e.g. what constitutes harm to humans, or the construction of a proxy system? If the relevant representations are not hand-coded by human designers, a bootstrapping problem would arise: to learn the concepts required for robust safety, some simpler decision algorithm, which would not itself be robustly safe, would need to be used.

For those concerned about catastrophic risks, the worry would be that the relative ease of creating simple motivational systems and guiding them via control over resources and rewards would enable the construction of powerful AI systems before robustly safe decision algorithms could be created. The enormous competitive advantages of human-comparable AI systems could compel corporations or governments to trade safety for speed in a race to advanced AI. Chalmers (2010) reports a consensus among cadets and staff at the U.S. West Point military academy that the U.S. government would not restrain AI research even in the face of potential catastrophe, for fear that rival powers would gain decisive advantage.

Thus, while the first-best safety measure would be to construct AI utility functions that favor cooperation regardless of p , or which would attach strong disutility to aggression as such, doing so may be quite difficult, and it is worthwhile to examine the space of AIs with generic, human-indifferent preferences using the 'drives' framework.

4. How much risk aversion with respect to resources?

The resource competition argument for the danger of very powerful human-indifferent AIs depends only on an AI having some more preferred use for the resources required to support human survival, not the precise strength of that preference. One pithy summary is that "[t]he AI does not hate you, nor does it love you, but you are made out of atoms which it can use for something else" (Yudkowsky, 2008). Given arbitrary power, the AI has something to gain and nothing to lose by taking the resources humans require and reallocating them to some very slightly more preferred use.

However, when evaluating risky conflict, a chance at gaining increased resources through successful aggression must be weighed against a chance of losing existing resources in the case of failure. Thus the AI's expected utility as a function of resources becomes essential: how would it trade off probabilities of different quantities of resources? Over various ranges of resource levels, would an n -fold multiplication in resources lead to utility much less than n times as great, close to n , or greater than n ? Would its utility function be bounded or unbounded? Would that bound be small or large relative to the supply of accessible resources? In combination with the resources available in the cooperative scenario, these factors would determine the extent to which the "resource drive" favored aggression.

5. Resource-satiable preferences

Individual humans typically display diminishing marginal utility with resources: a human billionaire is not a thousand times as happy as a millionaire, and few millionaires would stake their entire fortunes on a 0.1% chance of multiplying their wealth by a thousandfold. Likewise, starting a fight with a 50% chance of death and a 50% chance of doubling one's wealth in victory (expropriating the loser) would be rejected in favor of the status quo by most humans. Further, most humans seem to behave as though they had not only diminishing but bounded utility, rejecting gambles with infinite expected returns such as in the St. Petersburg paradox (Martin, 2008).

A similar dynamic might be expected for reinforcement learning systems with reward functions based on the discounted sum (bounded from above) of certain sense inputs or reward signals. So long as the AI could keep its reward channel continually active, the only use for additional resources would be to maintain and protect its hardware. The system could attain near-maximal utility simply by being securely maintained in a warehouse for a period depending on its discounting function, in an analog to rats repetitively triggering intracranial stimulation or a caricature of a human drug addict. Call AIs with preferences like these, which could be near-maximally satisfied with a trivial (to humans) quantity of resources, "resource-satiable."

$EU_{Success}$ will be near-maximal utility, with superabundant resources, and $EU_{Failure}$ near minimal utility, with destruction preventing any further attainment of reward. If we take the AI's maximal utility to be 1, and its minimum 0, then the AI would initiate aggression when:

$$p > EU_{Cooperation}$$

$EU_{Cooperation}$ could be less than 1 for two main reasons. First, access to reward signal could be restricted by humans, varying reward levels in response to success and compliance. This might be done to reserve incentives of increased reward to incentivize the AI to undertake risky requests. If average reward levels in the cooperation scenario were low, there would be strong incentives for aggression to gain control of the reward signal. Second, even if the typical reward for cooperation was near-maximal, there would remain some risk that the AI would be destroyed, intentionally or accidentally, by human action, e.g. to make room for a replacement model or as a side effect of nuclear war. Peaceful cooperation would then reduce to a Hobbesian mutual assurance problem: the AI would tend to have reason for aggression only when this would actually reduce the risk to itself on balance, with minimal conflict of interest beyond that.

6. Resource-hungry and exotic preferences

The difference between a resource-satiable and a 'resource-hungry' system could then be a matter of degree. A system similar to the previous example's with a lower discount rate might require the resources to operate for aeons to approach its maximal utility. Systems that valued the production of replicable physical structures, e.g. offspring, might be able to consume arbitrary quantities of resources in doing so and thus could have slowly diminishing (or linear or increasing) utility in resources, approaching any particular bound. However, the implications would depend on the magnitude of the bound.

If a system's utility bound could be closely approached with a small fraction of the accessible resources (including extrasolar ones, etc), it might still receive near-maximal utility in a cooperative scenario where it received an absolutely large but proportionally small share of resources, with the remainder used for human purposes. However, as the necessary resources approached those available in nearby space, this would involve progressively greater allocations of resources. The potential for conflict would seem most severe with an AI with utility proportional to resources up to a bound near the apparent supply: such an AI would engage in aggression whenever the probability of success exceeded the share of resources it would attain in the cooperative outcome. However, this is a very narrow region in the space of possibilities.

If we consider systems that would value some apparently physically unattainable quantity of resources orders of magnitude more than the apparently accessible resources given standard physics (e.g. resources enough to produce 10^{1000} offspring), the potential for conflict again declines for entities with bounded utility functions. Such resources are only attainable given very unlikely novel physical discoveries, making the agent's position similar to that described in "Pascal's Mugging" (Bostrom, 2009), with the agent's decision-making dominated by extremely small probabilities of obtaining vast resources. Conditioning on the very unlikely event that physics allows the acquisition of ludicrous quantities of resources needed to approach the agent's upper bound on utility, it is unlikely that just that amount can be created, rather than vastly more, most of which could be enjoyed by humans. Thus, when we consider bounded functions from resources to utility, it seems that in only a narrow class of cases would $EU_{Success}$ exceed the utility of a proportionally modest (although not necessarily absolutely modest) share of resources.

One might also consider utility unboundedly increasing with resources, e.g. a system with utility proportional to offspring for any number of offspring. Unbounded utility functions imply vulnerability to Dutch books (Mcgee, 1999), and require nonstandard decision theory, e.g. use of hyperreals, if they are not to assign equal expected utility to all actions (Hajek, 2003; De Blanc, 2007). However, if those problems were dealt with, such a system might again focus on Pascalian scenarios (allowing infinite resources), but would also be concerned about its proportional share.

Finally, perhaps the most threatening sort of 'generic' utility function would be one with utility rapidly increasing with the *proportional share* of attainable resources used by the system for its purpose. However, while human fairness norms or status drives sometimes result in preferences of this form, it is less obvious how this might arise in early AIs, if not as an unfortunate side-effect of attempting to mimic human values.

7. Gains from cooperation, deals, and enforcement

The above analysis suggests a high probability that, conditioning on our producing a human-indifferent AI with some non-negligible, and non-overwhelming likelihood of successful aggression, there would be large potential gains from cooperation. Specifically, it seems likely that secure tenure over a modest proportion of total resources would deliver near-maximal utility to a wide range of AI systems. Thus, if these

resources could be reliably promised, $EU_{Cooperation}$ could come very near to $EU_{Success}$, so that even an AI fairly confident in its ability to win a conflict could prefer peaceful cooperation. Since human extinction following successful AI aggression would be very bad by human standards, such a deal would be mutually beneficial, with the primary barriers being transaction costs and difficulties of trust or enforcement.

AIs considering initiating conflict through escape from standard monitoring and control mechanisms might face a dilemma: by revealing their motives and methods to open negotiations, they might simply elicit destruction. Pre-established policies and resources to ensure that AIs coming forward with unexpected but easily satiable preferences could increase AIs' estimates of $EU_{Cooperation}$, and thus their willingness to e.g. reveal security bugs instead of attempting aggression. However, normal instabilities in human society, dangers of technology development that might cause widespread destruction affecting even compliant AI, and the oddity of keeping promises to computer programs could all limit the peak credibility of such a commitment. Still, even if an AI estimated that such events would erode half of the expected utility of cooperation, this would still give it reason to cooperate unless it estimated that it would be more likely to succeed at aggression than to fail.

The cases of systems with preferences such that they would be satiated only with very large shares of resources raise two additional problems. First, it would become more difficult for humanity to credibly promise large portions of all future accessible wealth. Second, some philosophers urge that we should choose as though we had such preferences ourselves. Specifically, while standard approaches to cost-benefit and risk analysis value outcomes based on the revealed preferences of current citizens, e.g. the rate at which they trade probabilities of death for financial compensation, (Posner, 2004; Nordhaus, 2009), and result in a bounded social welfare function, some argue instead for social welfare functions that could increase unboundedly and linearly with resources. This is, for instance, a result of any consequentialist account that holds that the value of fulfilling human lives should be combined additively (Bostrom, 2008). However, even if cooperation is more difficult in these cases, the overall prospects for cooperation may still be favorable.

8. Conclusions

We have attempted to extend the analysis of (Omohundro, 2008) to a limited region of future possibilities, in which error or competitive pressures lead to the development of AI systems that could plausibly but not confidently threaten humanity, and have not been successfully engineered to be benevolent regardless of their relative power.

We argue that the convergent instrumental drives discussed by Omohundro are not as threatening in such cases as the analysis of very powerful AIs suggests: reduced likelihood of success is further accompanied by reduced motivation for conflict as opposed to cooperation. If robustly safe AIs are infeasible, we might still reduce risks by producing systems with resource demands that could be cheaply satiated, and by credibly committing to reduce the utility differential between cooperation and successful aggression for potentially threatening systems.

References

- Bostrom, N. (2002). Existential risks: Analyzing human extinction scenarios and related hazards. *Journal of Evolution and Technology*, 9(1).
- Bostrom, N. (2003). Astronomical waste: The opportunity cost of delayed technological development. *Utilitas*, 15(3), 308–314.
- Bostrom, N. (2005). The future of human evolution. In C. Tandy (Ed.), *Death and Anti-Death: Two Hundred Years After Kant, Fifty Years After Turing* (pp. 339-371). Palo Alto: Ria University Press.
- Bostrom, N. (2006). “Ethical issues in advanced artificial intelligence”, *Review of Contemporary Philosophy*, 5, 66-73.
- Bostrom, N. (2008) The infinitarian challenge to aggregative ethics. Retrieved from <http://www.nickbostrom.com/ethics/infinite.pdf>
- Bostrom, N. (2009). Pascal's mugging. *Analysis*, 69(3):443-445.
- Chalmers, D. (2010). The Singularity: a philosophical analysis. Retrieved from <http://consc.net/papers/singularity.pdf>
- de Blanc, P. (2007). Convergence of expected utilities with algorithmic probability distributions. Retrieved from <http://arxiv.org/pdf/0712.4318>
- Friedman, D. (2008). *Future imperfect: Technology and freedom in an uncertain world*. Cambridge: Cambridge University Press.
- Hájek, A. (2003). Waging war on pascal's wager. *Philosophical Review*, 112(1), 27-56.
- Hall, J. (2007). *Beyond AI: creating the conscience of the machine*. USA: Prometheus Books.
- Kurzweil, R. (2005). *The Singularity is near: when humans transcend biology*. USA: Viking Adult.
- Martin, R. (2008). The St. Petersburg paradox. *The Stanford Encyclopedia of Philosophy*. Retrieved from <http://plato.stanford.edu/entries/paradox-stpetersburg/>
- Matheny, J. G. (2007). Reducing the risk of human extinction. *Risk Analysis*, 27(5):1335-1344.
- McGee, V. (1999). An airtight Dutch book. *Analysis*, 57(4), 257-265.
- Moravec, H. (1999). *Robot: Mere machine to transcendent mind*. Oxford: Oxford University Press.

Nordhaus, W D. (2009). An analysis of the dismal theorem. Cowles Foundation Discussion Paper, (1696), Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1330454

Omohundro, S.M. (2008). The basic AI drives. In P. Wang, B. Goertzel, & S. Franklin (Eds.), *The Proceedings of the First AGI Conference* (pp. 483-492). Amsterdam: IOS Press.

Parfit, D. (1984). *Reasons and persons*. Oxford, UK: Clarendon Press, Part 4.

Posner, R. (2004). *Catastrophe: Risk and response*. Oxford: Oxford University Press.

Rees, M. (2004). *Our final hour: A scientist's warning*. Basic Books.

Sandberg, A, & Bostrom, N. (2008). Global catastrophic risks survey. *Future of Humanity Institute Technical Report 2008/1*. Retrieved from http://www.philosophy.ox.ac.uk/data/assets/pdf_file/0020/3854/global-catastrophic-risks-report.pdf

Sandel, M. J. (2006). *Public philosophy: essays on morality and politics*. Harvard Univ Press.

Weitzman, M. L. (2009). On modeling and interpreting the economics of catastrophic climate change. *Review of Economics and Statistics*, 91(1):1-19.

Yudkowsky, E., (2008). Artificial intelligence as a positive and negative factor in global risk. In N. Bostrom, & M. M. Čirković (Eds.), *Global catastrophic risks* (pp. 308-345). Oxford: Oxford University Press.