Bootstrapping Safe AGI Goal Systems

CEV and variants thereof

Introduction

The field of machine ethics seeks methods to ensure that future intelligent machines will act in ways beneficial to human beings. Machine ethics is relevant to a wide range of possible artificial agents, but becomes especially difficult and especially important when the agents in question have at least human-level intelligence. This paper describes a solution, originally proposed by Yudkowsky (2004), to the problem of what goals to give such agents: rather than attempt to explicitly program in any specific normative theory (a project which would face numerous philosophical and immediate ethical difficulties), we should implement a system to discover what goals we would, upon reflection, want such agents to have. We discuss the motivations for and details of this approach, comparing it to other suggested methods for creating 'artificial moral agents' (Wallach & Collin 2007), and describe underspecified and uncertain areas for further research.

The difficulty of machine ethics

A superintelligence, as Nick Bostrom (2003) defines it, is *"an intellect that is much smarter than the best human brains in practically every field, including scientific creativity, general wisdom and social skills."* It is possible that humanity will eventually create such an entity; several authors have suggested that this may occur within the next few decades (e.g. Bostrom 2003, Hall 2007). We cannot confidently place an upper bound on the ability a superintelligence would have to manipulate the physical world; for purposes of safety (as opposed to futurism), absolute power is the *conservative* assumption, and great power appears likely (Yudkowsky 2008). Shulman et al. (2009a) elaborate on the challenges posed by superintelligence to machine ethics, including:

- The possibility of human-equivalent or lesser AI precipitating an "intelligence explosion" (Good 1965) rapidly culminating in a superintelligence, meaning that even infrahuman AI warrants great caution, and that attempting to test a human-level AI's morality in a supposedly 'controlled' environment could be hazardous. Given this possibility, any proposed system to guide the behavior of an AI that is at all likely to self-improve should continue to produce desirable results if implemented by a superintelligence.
- The wide range of options available to a superintelligence acting in the real world, meaning that testing in a constrained environment, even if safe, could not assure us of the safety of an AI once released (Yudkowsky 2008).
- The problem of 'value lock-in': as Omohundro (2007) argues, AIs will try to protect themselves and their goal systems against outside interference, even if not given explicit drives to do so, as these things are instrumental values for almost any set of top-level goals. A superintelligence would thus resist attempts to alter its values or shut it down, almost certainly successfully; whatever values it is created with, the world will indefinitely contain an agent trying to further

those values, and very capable of doing so.

- Winner takes all: if a superintelligent AI were created, its self-protection drive would encourage it to prevent any other superintelligence from being created, as a rival superintelligence would provide the greatest significant obstacle to it achieving whatever goals it had. Given that a superintelligence is likely to be able to quickly become the most powerful agent in the world, it would probably have the ability to overpower other merely human groups trying to construct rival superintelligences, becoming a singleton (Bostrom 2006). Nor would a diversity of superintelligences inherently produce better outcomes than a singleton. While they would check each others' power, nothing would prevent them from having as much collective power as a single superintelligence, and coordinating to oppose new entrants; multiple superintelligences would be beneficial to humans only insofar as one or more of them had human-favorable values.

It appears plausible that if we want AI to have beneficial consequences, we will have to get its goals right on the first try, without testing. The obvious difficulty of doing so is underscored by the great diversity of theories proposed in philosophical ethics, and the great divergence of actions prescribed for a superintelligence by these theories (Shulman et al. 2009b). Deciding what a superintelligence should do, let alone doing so in sufficient detail to guide the construction of a machine, requires much more insight into our preferences than we have.

Sources of normative fallibility

Why do we lack this insight? Simple models of rationality assume that agents have precisely specified goals and know them, though of course they may be greatly uncertain about the best means to achieve their goals. However, individual humans depart from this model in many ways:

- *Subjective normative uncertainty.* Individuals do not have full knowledge of their preferences and values, and are aware of this. On a personal level, we often devote substantial effort to discovering what we want; when it comes to interpersonal ethics, we can be persuaded to endorse a variety of incommensurable frameworks (consequentialism, deontology, virtue ethics) and a vast range of views within these frameworks.
- *Incoherent, context-dependent judgments.* People regularly display inconsistent personal decisions; often (e.g. in the case of drug addicts) it seems simple for others to determine their 'true' preferences, though enabling an AMA (artificial moral agent) to make such determinations is unlikely to be simple. Less obviously, and posing a greater problem, experiments in moral psychology have uncovered numerous ways in which individuals' judgments differ depending on seemingly irrelevant or even non-consciously-recognized features of a situation, so that they cannot be said to have coherent preferences about that type of situation (e.g. Wheatley and Haidt 2005). Defining individuals' actual preferences in such cases is a serious problem; using revealed preference to do so may not be possible.
- *Post-hoc justifications.* Investigations into the metaethical implications of human moral psychology and behavior (Greene 2002) indicate that much of explicit human moral reasoning is post-hoc justification for innate, automatic, non-conscious moral instincts. Experimental

results such as moral dumbfounding (Haidt 2001) support this view. This means that if the explicitly stated moral principles of an individual are transmitted to an AMA, the result might be highly undesirable. Instead, the underlying psychology itself must be transferred to the AMA.

- *Limited domains.* Technological expansion of human capacities has always required the gradual generalization of our preferences to new domains; constructing a superintelligence (conservatively assumed to be) capable of arbitrary actions would require the *complete* generalization of our preferences, *up front* with no opportunity to learn from experience. Even perfect knowledge of our preferred actions in 'normal' situations would underdetermine our preference ordering over all achievable outcomes, with different models that fit the normal range extrapolating in wildly different directions (Shulman et al. 2009b).

These are all obstacles to constructing an AI to fulfill the preferences of one individual, and in particular to doing so using purely object-level ethical reasoning.

Social choice and bootstrapping

Machine ethics goes beyond determining individual preferences. A superintelligence's actions would affect the entire world; its choices are social choices. The task of defining social preferences both inherits the ill-definedness of individual preferences, and creates the new requirement to choose a means of combining these preferences.

The choice of a method of aggregating and weighing individual preferences introduces numerous new degrees of freedom. Interpersonal utility comparisons are undefined for general agents; commonalities of psychology common-sensically allow comparisons between humans, but still do not give a clear canonical solution (Elster & Roemer 1991). Further, the desire to avoid tyrannies of the majority, or the standard repugnant implications of utilitarianism, likely requires a satisfactory theory of social choice to be more complex than linear aggregation — but in what ways?

For the designers of a superintelligence to decide these questions themselves would be inherently ethically and politically undesirable, even if (as is not the case) they had a reasonable chance of answering them all 'correctly'; these decisions about social choice are themselves social choices. Faced with the need for 'machine-grade' theories of both individual preference and social choice, we must somehow bootstrap our way to them from as neutral a position as possible.

Coherent extrapolated volition

Yudkowsky (2004) has suggested such a bootstrapping process for a superintelligence's value system:

> The initial dynamic should implement the *coherent extrapolated volition of humankind*. 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.

Yudkowsky's proposal, summarized in the above paragraph, is that the first superintelligence be given the goal of extrapolating human moral change under a process of idealization, letting the idealized humanity deliberate and reflect on and modify itself and this process. This reflection is analogous to Rawls' concept of reflective equilibrium (Rawls 1971), wherein concrete moral judgments and abstract principles are recursively used to revise each other. Since moral change is likely to depend partly on contingent conditions, like the order in which arguments are introduced or random emergent social phenomena, the result would be a range of possible idealizations; if these extrapolations converged sufficiently in some elements of preference, the AI would act on these convergent preferences.

While a full re-description of coherent extrapolated volition is beyond the scope of this paper, some points are worth noting:

- It addresses, in Yudkowsky's terms, the problem of *content* but not *structure*. It describes, at a very high level, goals to give to an AI, setting aside the profound challenges of describing these goals mathematically and creating a system that reliably implements them.
- CEV is an *"initial dynamic."* The intent is not that an AMA should act at all times according to humanity's volition, but that our volition be extrapolated *once* and acted on. In particular, the initial extrapolation could generate an object-level goal system we would be willing to endow a superintelligent AMA with.
- While an initial theory is required both to extrapolate and to combine preferences within and between extrapolations, intermediate results (and input from the designers) can be allowed to reflect on and modify the extrapolation and aggregation dynamics, and convergence can be tested between different choices of initial conditions as well as between possible outcomes of one initial condition.
- The success of CEV depends on sufficient coherence in the extrapolation: clear enough preferences, converged upon strongly enough within and between possibilities, to provide a guide for action. (This does not mean unanimity on all aspects of preference — radically different extrapolations may still agree on what immediate actions are desirable.) The existence of sufficient coherence is not certain; if it is not present, the system implementing CEV should execute a "controlled shutdown" rather than behaving unpredictably.

Comparative analysis

For the designers of an AMA to directly specify its object-level values is undesirable, due to their fallibility and the improbability of making all the required decisions correctly (Shulman et al. 2009b). Additionally, it is undesirable to give the designers a special position in deciding humanity's future: besides basic considerations of fairness, it could lead to serious conflict over the privilege (Shulman 2009c). Extrapolated volition eliminates the need to specify object-level values: it captures the (consciously inaccessible) processes which generate our ethical intuitions, which through conscious reflection lead in turn to our explicit moral theories (Haidt 2001). (It does not entirely eliminate the possibility of conflict. The initial population extrapolated is a potential area of dispute, as are free parameters in the extrapolation process which might affect the output (discussed below.) Also, other types of fair value systems are conceivable, though the absence of either extrapolation or detailed moral content imposes severe limitations.)

Compared to any proposal to directly design a superintelligent AMA's goals, CEV obviates most of the decisions that would need to be made, and allows some errors in the remaining dimensions (the initial model of social choice) to be corrected. It requires the additional specification of a theory of extrapolation (also error-corrected), and a set of constraints to ensure safety during the extrapolation process (though such constraints would be useful in the development stage of any AMA.) Clearly, the required components will still be extremely difficult to formalize. However, this complexity is likely to be a property of any satisfactory theory of machine ethics; to the extent that formalizing object-level values appears simpler, this is likely due to our blindness to the human-universal but specifically human complexity of moral intuitions (Yudkowsky 2008), in contrast to the non-intuitive concept of extrapolation.

CEV is not the only meta-level proposal in machine ethics. Others, such as Hibbard (2001) and Guarini (2005), have proposed inferring human preferences from behavior. This would surely be simpler than CEV. However, attempts to use revealed preference directly or alone suffer from all the ambiguities of individual preference: revealed preferences may not be consistent, may not be 'true' preferences, and underdetermine decisions about radically new situations. Furthermore, the question of social choice is left unanswered, or implicitly answered with total utilitarianism. Some of these problems may be solvable, at least in part, through relatively simple mechanisms such as prioritizing higher-order desires; however, the hope of satisfactorily capturing the complexity of moral deliberation through such heuristics may not be realistic.

Open questions

There is no unique meta-level, extrapolating candidate goal system for a superintelligent AMA. A number of algorithms share with CEV the following five desirable properties:

- *Meta-algorithm*: Most goals the AI has will be harvested at run-time from human minds, rather than explicitly programmed in before run-time. *Justification*: We want an AMA's actions to be grounded in our preferences, but those preferences are complex and opaque, making our reports unreliable. Also, we want an AMA to fairly take into account everyone's values, rather than privileging those of the designers.

- *Factually correct beliefs*: Using the AI's superhuman ability to ascertain the correct answer to any factual question in order to modify preferences or desires that are based upon false factual beliefs. *Justification*: Instrumental preferences over actions depend on relevant facts (unintended consequences), and preferences themselves must be defined in terms of a realistic ontology.
- *Singleton*: Only one superintelligent AMA is to be constructed, and it is to take control of the entire future light cone with whatever goal function is decided upon. *Justification*: a singleton is the likely default outcome for superintelligence, and stable co-existence of superintelligences, if achievable, would offer no inherent advantages for humans.
- *Reflection*: Individual or group preferences are reflected upon and revised, in the style of Rawls' reflective equilibrium. *Justification:* Helps to resolve moral fallibility and inconsistent preferences, to generalize preferences to new domains, and to bootstrap a theory of social choice.
- *Preference aggregation*: The set of preferences of a whole group are to be combined somehow. *Justification*: A group of humans may share goals enough to collaborate on building an AI and/or agree that they should all be beneficiaries of it, but not have identical goals.

The set of factually correcting, singleton, reflective, aggregative meta-algorithms is larger than just the CEV algorithm. For example, there is no reason to suppose that factual correction, reflection, and aggregation, performed in any order, will give the same result; therefore, there are at least 6 variants depending upon ordering of these various processes, and many variants if we allow small increments of these processes to be interleaved. CEV also stipulates that the algorithm should extrapolate ordinary human-human social interactions concurrently with the processes of reflection, factual correction and preference aggregation; this requirement could be dropped.

One variant that stands out is *Individual Extrapolated Volition followed by Negotiation*, where each individual human's preferences are extrapolated by factual correction and reflection; once that process is fully complete, the extrapolated humans negotiate a combined utility function for the resultant superintelligence. Another variant would be to weight the negotiating power of each human or extrapolated human in some way.

Given a particular structure of volition extrapolation, some tunable parameters still remain. Parameters in CEV include the threshold for coherence below which CEV shuts down, the extent to which the majority can overrule minorities, and many others. Which values should be chosen? How much sensitivity does the final outcome have to these choices?

Again, CEV is premised on the existence of a sufficient degree of coherence in humanity's extrapolated volition. How much coherence can realistically be expected is up for debate. On the one hand, Yudkowsky suggests that the human-universal nature of many moral norms (Brown 1991) is cause for optimism — that the process of idealized reflection will tend to wash out cultural and neurological differences in favor of shared, innate moral generating mechanisms. On the other hand, Greene (2002) emphasizes that *"human moral instinct is ... universal in form but local in content,"* and goes on to say that *"it is highly unlikely that there will be anything close to cross-cultural consensus on general principles for organizing the basic structure of society."* If humanity's volition does not cohere because of widely differing extrapolations, even under different choices of parameters (or if different parameters produce different coherent outcomes), using CEV would force on us an arbitrary but overwhelmingly influential choice. If it does not cohere because of irreconcilable differences within extrapolations, the additional alternative exists of narrowing the set of people initially extrapolated; however, such exclusion raises serious ethical questions and may entail political problems. Given our current uncertainty, the major features of the approach, and these open questions, seem well worth exploring further.

References

Bostrom, Nick. 2003. "Ethical Issues In Advanced Artificial Intelligence." In *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, Vol. 2, ed. I. Smit et al., Int. Institute of Advanced Studies in Systems Research and Cybernetics, pp. 12–17.

Bostrom, Nick. 2006. "What is a Singleton?" *Linguistic and Philosophical Investigations* 5(2): 48–54.

Brown, Donald E. 1991. *Human universals*. McGraw-Hill.

Elster, Jon, and John E. Roemer. 1991. *Interpersonal Comparisons of Well-Being*. Cambridge University Press.

Goertzel, Ben, and Cassio Pennachin. 2007. *Artificial General Intelligence*. Springer.

Good, Irving John. 1965. "Speculations Concerning the First Ultraintelligent Machine." *Advances in Computers* 31–88.

Greene, Joshua. 2002. "The Terrible, Horrible, No Good, Very Bad Truth About Morality and What to Do About It." Doctoral dissertation, Princeton University.

Guarini, Marcello. 2006. "Particularism and the Classification and Reclassification of Moral Cases." IEEE Intelligent Systems 21(4): 22–28.

Haidt, Jonathan. 2001. "The Emotional Dog and its Rational Tail: A social intuitionist approach to moral judgment", *Psychological Review* 108: 814–834.

Hall, John Storrs. 2007. *Beyond AI: Creating the Conscience of the Machine*. Prometheus Books.

Hibbard, Bill. 2001. "Super-intelligent machines." *ACM SIGGRAPH Computer Graphics* 35(1).

Omohundro, Stephen M. 2007. "The Nature of Self-Improving Artificial Intelligence." Presented and distributed at the 2007 Singularity Summit, revised version at http://selfawaresystems.com/2007/10/05/paper-on-the-nature-of-self-improving-artificial-intelligence/.

Rawls, John. 1971. *A Theory of Justice*. Harvard University Press.

Shulman, Carl, Henrik Jonsson and Nick Tarleton. 2009 (a). "Machine Ethics and Superintelligence." Presented at the 5th Asia-Pacific Computing and Philosophy Conference.

Shulman, Carl, Nick Tarleton and Henrik Jonsson. 2009 (b). "Which Consequentialism? Machine Ethics and Moral Divergence." Presented at the 5th Asia-Pacific Computing and Philosophy Conference.

Shulman, Carl. 2009 (c). "Arms Control and Intelligence Explosions." Presented at the 7th European Conference on Computing and Philosophy.

Wallach, Wendell and Allen Collin. 2005. "Android Ethics: Bottom-up and Top-down Approaches for Modeling Human Moral Faculties", in *Proceedings of the 2005 COGSCI workshop: Toward Social Mechanics of Android Science*, 149–159.

Wheatley, Thalia and Jonathan Haidt. 2005. "Hypnotic disgust makes moral judgments more severe." *Psychological Science* 16(10): 780–784.

Yudkowsky, Eliezer. 2004. "Coherent Extrapolated Volition." http://singinst.org/upload/CEV.html.

Yudkowsky, Eliezer. 2008. "Artificial Intelligence as a positive and negative factor in global risk." In *Global Catastrophic Risks*, eds. Nick Bostrom and Milan Cirkovic, Oxford University Press.