MIRI
MACHINE INTELLIGENCE
RESEARCH INSTITUTE

Ideal Advisor Theories and Personal CEV

Luke Muehlhauser
Machine Intelligence Research Institute

Chris Williamson

Abstract

Yudkowsky's "coherent extrapolated volition" (CEV) concept shares much in common with ideal advisor theories in moral philosophy. Does CEV fall prey to the same objections which are raised against ideal advisor theories? Because CEV is an epistemic rather than a metaphysical proposal, it seems that at least one family of CEV approaches (inspired by Bostrom's parliamentary model) may escape the objections raised against ideal advisor theories. This is not a particularly ambitious article; it mostly aims to place CEV in the context of mainstream moral philosophy.

Muehlhauser, Luke and Chris Williamson. 2013. "Ideal Advisor Theories and Personal CEV."
Machine Intelligence Research Institute, Berkeley, CA

1. Introduction

What is of value to an agent? Maybe it's just whatever they desire. Unfortunately, our desires are often the product of ignorance or confusion. I may desire to drink from the glass on the table because I think it is water when really it is bleach. So perhaps something is of value to an agent if they would desire that thing *when fully informed*. But here we crash into a different problem. It might be of value for an agent who wants to go to a movie to look up the session times, but the fully informed version of the agent will not desire to do so—they are fully informed and hence already know all the session times. The agent and its fully informed counterpart have different needs. Thus, several philosophers have suggested that something is of value to an agent if an ideal version of that agent (fully informed, perfectly rational, etc.) would *advise* the non-ideal version of the agent to pursue that thing.

This idea of idealizing or extrapolating an agent's preferences[1] goes back at least as far as Sidgwick (1907), who considered the idea that "a man's future good" consists in "what he would now desire . . . if all the consequences of all the different [actions] open to him were accurately foreseen." Similarly, Rawls (1971) suggested that a person's good is the plan "that would be decided upon as the outcome of careful reflection in which the agent reviewed, in the light of all the relevant facts, what it would be like to carry out [their] plans." More recently, in an article about rational agents and moral theory, Harsanyi (1982) defined what an agent's rational wants as "the preferences he *would* have if he had all the relevant factual information, always reasoned with the greatest possible care, and were in a state of mind most conducive to rational choice." A few years later, Railton (1986) identified a person's good with "what he would want himself to want . . . were he to contemplate his present situation from a standpoint fully and vividly informed about himself and his circumstances, and entirely free of cognitive error or lapses of instrumental rationality."

[1] Another clarification to make concerns the difference between idealization and extrapolation. An *idealized agent* is a version of the agent with certain idealizing characteristics (perhaps logical omniscience and infinite speed of thought). An *extrapolated agent* is a version of the agent that represents what they would be like if they underwent certain changes or experiences. Note two differences between these concepts. First, an extrapolated agent need not be ideal in any sense (though useful extrapolated agents often will be) and certainly need not be *perfectly* idealized. Second, extrapolated agents are determined by a specific type of process (extrapolation from the original agent) whereas no such restriction is placed on how the form of an idealized agent is determined. CEV utilizes extrapolation rather than idealization, as do some ideal advisor theories. In this post, we talk about "ideal" or "idealized" agents as a catch-all for both idealized agents and extrapolated agents.

Rosati (1995) calls these theories "Ideal Advisor" theories of value because they identify one's personal value with what an ideal version of oneself would advise the nonideal self to value.

Looking not for a metaphysical account of value but for a practical solution to machine ethics (Wallach and Allen 2009; Muehlhauser and Helm 2012), Yudkowsky (2004) described a similar concept which he calls "coherent extrapolated volition" (CEV):

> 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.

In other words, the CEV of humankind is about the preferences that we would have as a species if our preferences were extrapolated in certain ways. Armed with this concept, Yudkowsky then suggests that we implement CEV as an initial dynamic for Friendly AI. Tarleton (2010) explains that the intent of CEV is 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 [machine] with."

CEV theoretically avoids many problems with other approaches to machine ethics (Yudkowsky 2004; Tarleton 2010; Muehlhauser and Helm 2012). However, there are reasons it may not succeed. In this post, we examine one such reason: resolving CEV at the level of humanity (*global CEV*) might require at least partially resolving CEV at the level of individuals (*personal CEV*),[2] but personal CEV is similar to ideal advisor theories of value,[3] and such theories face well-explored difficulties. As such, these difficulties may undermine the possibility of determining the global CEV of humanity.

Before doing so, however, it's worth noting one key difference between ideal advisor theories of value and personal CEV. Ideal advisor theories typically are linguistic or metaphysical theories, while the role of personal CEV is epistemic. Ideal advisor theorists attempts to define *what it is* for something to be of value for an agent. Because of

[2] Standard objections to ideal advisor theories of value are also relevant to some proposed variants of CEV, for example Tarleton's (2010) suggestion of "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." Furthermore, some objections to ideal advisor theories seem relevant to global CEV even if they are not relevant to a particular approach to personal CEV, though that discussion is beyond the scope of this article. As a final clarification, see Dai (2010).

[3] Ideal advisor theories are not to be confused with "Ideal Observer theory" (Firth 1952). For more on ideal advisor theories of value, see Zimmerman (2003); Tanyi (2006); Enoch (2005); Miller (2013, chap. 9).

this, their accounts needs to give an unambiguous and plausible answer in all cases. On the other hand, personal CEV's role is an epistemic one: it isn't intended to define what is of value for an agent. Rather, personal CEV is offered as a technique that can help an AI to *come to know*, to some reasonable but not necessarily perfect level of accuracy, what is of value for the agent. To put it more precisely, personal CEV is intended to allow an initial AI to determine what sort of superintelligence to create such that we end up with what Yudkowsky (2004) calls a "Nice Place to Live." Given this, certain arguments are likely to threaten ideal advisor theories and not personal CEV, and vice versa.

With this point in mind, we now consider some objections to ideal advisor theories of value and examine whether they threaten personal CEV.

2. Sobel's First Objection: Too Many Voices

Four prominent objections to ideal advisor theories are due to Sobel (1994). The first of these, the "too many voices" objection, notes that the evaluative perspective of an agent changes over time and, as such, the views that would be held by the perfectly rational and fully informed version of the agent will also change. This implies that each agent will be associated not with one idealized version of themselves but with a set of such idealized versions (one at time t, one at time $t + 1$, etc.), some of which may offer conflicting advice. Given this "discordant chorus," it is unclear how the agent's nonmoral good should be determined.

Various responses to this objection run into their own challenges. First, privileging a single perspective (say, the idealized agent at time $t + 387$) seems ad hoc. Second, attempting to aggregate the views of multiple perspectives runs into the question of how trade-offs should be made. That is, if two of the idealized viewpoints disagree about what is to be preferred, it's unclear how an overall judgment should be reached.[4] Finally, the suggestion that the idealized versions of the agent at different times will have the same perspective seems unlikely, and surely it's a substantive claim requiring a substantive defense. So the obvious responses to Sobel's first objection introduce serious new challenges which then need to be resolved.

[4] This is basically an intrapersonal version of the standard worries about interpersonal comparisons of well-being. The basis of these worries is that even if we can specify an agent's preferences numerically, it's unclear how we should compare the numbers assigned by one agent with the numbers assigned by the other. In the intrapersonal case, the challenge is to determine how to compare the numbers assigned by the same agent at different times. See Gibbard (1986).

One final point is worth noting: it seems that this objection is equally problematic for personal CEV. The extrapolated volition of the agent is likely to vary at different times, so how ought we determine an overall account of the agent's extrapolated volition?

3. Sobel's Second and Third Objections: Amnesia

Sobel's second and third objections build on two other claims (see Sobel [1994] for a defense of these). First, some lives can only be evaluated if they are experienced. Second, experiencing one life can leave an agent incapable of experiencing another in an unbiased way. Given these claims, Sobel presents an *amnesia model* as the most plausible way for an idealized agent to gain the experiences necessary to evaluate all the relevant lives. According to this model, an agent experiences each life sequentially but undergoes an amnesia procedure after each one so that they may experience the next life uncolored by their previous experiences. After experiencing all lives, the amnesia is then removed.

Following on from this, Sobel's second objection is that the sudden recollection of a life from one evaluative perspective and living a life from a vastly different evaluative perspective may be strongly dissimilar experiences. So when the amnesia is removed, the agent has a particular evaluative perspective (informed by their memories of all the lives they've lived) that differs so much from the evaluative perspective they had when they lived the life independently of such memories that they might be incapable of adequately evaluating the lives they've experienced based on their current, more knowledgeable, evaluative perspective.

Sobel's third objection also relates to the amnesia model: he argues that the idealized agent might be driven insane by the entire amnesia process and hence might not be able to adequately evaluate what advice they ought to give the nonideal agent. In response to this, there is some temptation to simply demand that the agent be idealized not just in terms of rationality and knowledge but also in terms of their sanity. However, perhaps any idealized agent that is similar enough to the original to serve as a standard for their nonmoral good will be driven insane by the amnesia process and so the demand for a sane agent will simply mean that no adequate agent can be identified.

If we grant that an agent needs to experience some lives to evaluate them, and we grant that experiencing some lives leaves them incapable of experiencing others, then there seems to be a strong drive for personal CEV to rely on an amnesia model to adequately determine what an agent's volition would be if extrapolated. If so, however, personal CEV seems to face the challenges raised by Sobel.

4. Sobel's Fourth Objection: Better Off Dead

Sobel's final objection is that the idealized agent, having experienced such a level of perfection, might come to the conclusion that their nonideal counterpart is so limited as to be better off dead. Further, the ideal agent might make this judgment because of the relative level of well-being of the nonideal agent rather than the agent's absolute level of well-being. (That is, the ideal agent may look upon the well-being of the nonideal agent as we might look upon our own well-being after an accident that caused us severe mental damage. In such a case, we might be unable to objectively judge our life after the accident due to the relative difficulty of this life as compared with our life before the accident.) As such, this judgment may not capture what is actually in accordance with the agent's nonmoral good.

Again, this criticism seems to apply equally to personal CEV: when the volition of an agent is extrapolated, it may turn out that this volition endorses killing the nonextrapolated version of the agent. If so this seems to be a mark against the possibility that personal CEV can play a useful part in a process that should eventually terminate in a nice place to live.

5. A Model of Personal CEV

The seriousness of these challenges for personal CEV is likely to vary depending on the exact nature of the extrapolation process. To give a sense of the impact, we will consider one family of methods for carrying out this process: the *parliamentary model* (inspired by Bostrom [2009]). According to this model, we determine the personal CEV of an agent by simulating multiple versions of them, extrapolated from various starting times and along different developmental paths. Some of these versions are then assigned to a parliament where they vote on various choices and make trades with one another.

Clearly this approach allows our account of personal CEV to avoid the "too many voices" objection. After all, the parliamentary model provides us with an account of how we can aggregate the views of the agent at various times: we should simulate the various agents and allow them to vote and trade on the choices to be made. It is through this voting and trading that the various voices can be combined into a single viewpoint. While this process may not be adequate as a metaphysical account of value, it seems more plausible as an account of personal CEV as an epistemic notion. Certainly, your authors would deem themselves to be more informed about what they value if they knew the outcome of the parliamentary model for themselves.

This approach is also able to avoid Sobel's second and third objections. The objections were specifically targeted at the amnesia model where one agent experienced multiple lives. As the parliamentary model does not utilize amnesia, it is immune to these concerns.

What of Sobel's fourth objection? Sobel's concern here is not simply that the idealized agent might advise the agent to kill themselves. After all, sometimes death may, in fact, be of value for an agent. Rather, Sobel's concern is that the idealized agent, having experienced such heights of existence, will become biased against the limited lives of normal agents.

It's less clear how the parliamentary model deals with Sobel's fourth objection, which plausibly retains its initial force against this model of personal CEV. However, we're not intending to solve personal CEV entirely in this short post. Rather, we aim to demonstrate only that the force of Sobel's four objections will depend on the model of personal CEV selected. Reflection on the parliamentary model makes this point clear.

So the parliamentary model seems able to avoid at least three of the direct criticisms raised by Sobel. It is worth noting, however, that some concerns remain. For those who accept Sobel's claim that experience is necessary to evaluate some lives, it is clear that no member of the parliament will be capable of comparing their life to all other possible lives, as none will have all the required experience. As such, the agents may falsely judge a certain aspect of their life to be more or less valuable than it, in fact, is. For a metaphysical account of personal value, this problem might be fatal. Whether it is also fatal for the parliamentary model of personal CEV depends on whether the knowledge of the various members of the parliament is enough to produce a nice place to live regardless of its imperfection.

Two more issues might arise. First, the model might require careful selection of who to appoint to the parliament. For example, if most of the possible lives that an agent could live would drive them insane, then selecting at random which of these agents to appoint to the parliament might lead to a vote by the mad. Second, it might seem that this approach to determining personal CEV will require a reasonable level of accuracy in simulation. If so, there might be concerns about the creation of, and responsibility to, potential moral agents.

Given these points, a full evaluation of the parliamentary model will require more detailed specification and further reflection. However, two points are worth noting in conclusion. First, the parliamentary model does seem to avoid at least three of Sobel's direct criticisms. Second, even if this model eventually ends up being flawed on other grounds, the existence of one model of personal CEV that can avoid three of Sobel's objections gives us reason to expect that other promising models of personal CEV may be discovered.

References

Bostrom, Nick. 2009. "Moral Uncertainty – Towards a Solution?" *Overcoming Bias* (blog), January 1. http://www.overcomingbias.com/2009/01/moral-uncertainty-towards-a-solution.html.

Dai, Wei. 2010. "Complexity of Value $\neq$ Complexity of Outcome." *Less Wrong* (blog), January 30. http://lesswrong.com/lw/1oj/complexity_of_value_complexity_of_outcome/.

Enoch, David. 2005. "Why Idealize?" *Ethics* 115 (4): 759–787. doi:10.1086/430490.

Firth, Roderick. 1952. "Ethical Absolutism and the Ideal Observer." *Philosophy and Phenomenological Research* 12 (3): 317–345. doi:10.2307/2103988.

Gibbard, Allan. 1986. "Interpersonal Comparisons: Preference, Good, and the Intrinsic Reward of a Life." In *Foundations of Social Choice Theory,* edited by Jon Elster and Aanund Hylland, 165–193. New York: Cambridge University Press.

Harsanyi, John C. 1982. "Morality and the Theory of Rational Behavior." In *Utilitarianism and Beyond,* edited by Amartya Sen and Bernard Williams, 39–62. New York: Cambridge University Press. doi:10.1017/CBO9780511611964.004.

Miller, Alexander. 2013. *Contemporary Metaethics: An Introduction.* 2nd ed. Cambridge: Polity.

Muehlhauser, Luke, and Louie Helm. 2012. "The Singularity and Machine Ethics." In *Singularity Hypotheses: A Scientific and Philosophical Assessment,* edited by Amnon Eden, Johnny Søraker, James H. Moor, and Eric Steinhart. The Frontiers Collection. Berlin: Springer.

Railton, Peter. 1986. "Facts and Values." *Philosophical Topics* 14 (2): 5–31.

Rawls, John. 1971. *A Theory of Justice.* Cambridge, MA: Belknap.

Rosati, Connie S. 1995. "Persons, Perspectives, and Full Information Accounts of the Good." *Ethics* 105 (2): 296–325. doi:10.1086/293702.

Sidgwick, Henry. 1907. *Methods of Ethics.* 7th ed. Macmillan.

Sobel, David. 1994. "Full Information Accounts of Well-Being." *Ethics* 104 (4): 784–810. http://www.jstor.org/stable/2382218.

Tanyi, Attila. 2006. "An Essay on the Desire-Based Reasons Model." PhD diss., Central European University. http://web.ceu.hu/polsci/dissertations/Attila_Tanyi.pdf.

Tarleton, Nick. 2010. *Coherent Extrapolated Volition: A Meta-Level Approach to Machine Ethics.* The Singularity Institute, San Francisco, CA. http://intelligence.org/files/CEV-MachineEthics.pdf.

Wallach, Wendell, and Colin Allen. 2009. *Moral Machines: Teaching Robots Right from Wrong.* New York: Oxford University Press. doi:10.1093/acprof:oso/9780195374049.001.0001.

Yudkowsky, Eliezer. 2004. *Coherent Extrapolated Volition.* The Singularity Institute, San Francisco, CA, May. http://intelligence.org/files/CEV.pdf.

Zimmerman, David. 2003. "Why Richard Brandt Does Not Need Cognitive Psychotherapy, and Other Glad News about Idealized Preference Theories in Meta-Ethics." *Journal of Value Inquiry* 37 (3): 373–394. doi:10.1023/B:INQU.0000013348.62494.55.