MIRI
MACHINE INTELLIGENCE
RESEARCH INSTITUTE

Which Consequentialism? Machine Ethics and Moral Divergence

Carl Shulman, Henrik Jonsson
MIRI Visiting Fellows

Nick Tarleton
Carnegie Mellon University, MIRI Visiting Fellow

Abstract

Some researchers in the field of machine ethics have suggested consequentialist or utilitarian theories as organizing principles for Artificial Moral Agents (AMAs) (Wallach, Allen, and Smit 2008) that are 'full ethical agents' (Moor 2006), while acknowledging extensive variation among these theories as a serious challenge (Wallach, Allen, and Smit 2008). This paper develops that challenge, beginning with a partial taxonomy of consequentialisms proposed by philosophical ethics. We discuss numerous 'free variables' of consequentialism where intuitions conflict about optimal values, and then consider special problems of human-level AMAs designed to implement a particular ethical theory, by comparison to human proponents of the same explicit principles. In conclusion, we suggest that if machine ethics is to fully succeed, it must draw upon the developing field of moral psychology.

Shulman, Carl, Nick Tarleton, and Henrik Jonsson. 2009. "Which Consequentialism? Machine Ethics and Moral Divergence." In *AP-CAP 2009: The Fifth Asia-Pacific Computing and Philosophy Conference, October 1st-2nd, University of Tokyo, Japan, Proceedings*, edited by Carson Reynolds and Alvaro Cassinelli, 23–25. AP-CAP 2009. `http://ia-cap.org/ap-cap09/proceedings.pdf`.

This version contains minor changes.

1. Free Variables of Consequentialism

Suppose that the recommendations of a broadly utilitarian view depend on decisions about ten free binary variables, where we assign a probability of 80% to our favored option for each variable; in this case, if our probabilities are well-calibrated and our errors are not correlated across variables, then we will have only slightly more than a 10% chance of selecting the correct (in some meta-ethical framework) specification. As the number of variables increases, our options grow more plentiful, and as our uncertainty about each item grows, the likelihood of success continues to decline.

Unfortunately, the philosophical literature shows many highly contested dimensions along which consequentialisms vary. Even if we confine ourselves to broadly utilitarian views that consider only some combination of pain, pleasure, and preference satisfaction, we still face a vast bestiary of views.

Hedonistic utilitarianism requires an account of experiential states such as pleasure and pain, whose ontological status is disputed (Chalmers 1996; Place 1956; Searle 2007; Putnam 1960), with each construal bearing implications about what systems deserve moral consideration. Utilitarians disagree about the relative value of preventing suffering and promoting happiness, with views that strongly favor the former, such as negative utilitarianism (Smart 1958), suggesting that the painless extermination of all life capable of suffering would be desirable, while the opposite camp prizes galactic colonization (Bostrom 2003).

Preference utilitarians face other troubles. McCarthy (1979) attributes beliefs about temperature to a thermostat, which might imply that it has 'preferences' over the temperature as well. If simple definitions fail to map onto our 'emphatic preferences' (Binmore 2009; Harsanyi 1977), we will require more complex and error-prone ones. There is also dispute about whether to value the actual satisfaction of preferences or only the experience of such satisfaction, a distinction highlighted by Nozick's 'experience machine' scenario (Nozick 1974).

All utilitarianisms founder on interpersonal comparison of utility, which is not defined in standard expected utility theory (Binmore 2009; Elster and Roemer 1991). The selection of any comparison/aggregation rule introduces yet more free variables and problems: e.g., bounded utility functions can be compared by scaling them into the same range, but the result is strongly dominated by the relative strength of people's strongest preferences (Hausman 1995). Aggregation might involve the sum, average, or some more complicated function of individuals' well-being, with potentially counterintuitive results, such as the Repugnant Conclusion (Parfit 1986), attaching to many proposals. The offsetting Person-Affecting Restriction carries its own surprising implications, such as extreme indifference to the far future (Parfit 1986; Glover 1977).

Cases involving small probabilities of large utilities, as in Pascal's Wager or the St. Petersburg paradox (Bernoulli [1738] 1954), have motivated proposals for bounding the range of the utility function, or computing expected value in a nonstandard way. Standard decision theory, together with an unbounded and aggregative utility function, appears to favor extremely counterintuitive actions in some simple cases (Bostrom 2009); various proposals for bounding utility been made (Hardin 1982; Gustason 1994), but there is no universally accepted solution (Cowen and High 1988).

Combined with other factors, the picture that emerges is one of deep confusion: we have no idea what utility function to endow an explicit moral agent with.

2. Current Moral Theories Are Inadequate for Machine Ethics

One response to the difficulty of selecting a 'canonical' utility function is to note that actual human philosophers claim to hold a variety of consequentialist views and yet tend to show roughly similar and comparably moral behavior. Few human utilitarians rob banks to donate to the relief of famine and disease in Africa, or even donate much of their luxury budgets (Singer 1972). Total utilitarians rarely spend resources on increasing population growth among the global poor, even if they nominally endorse the Repugnant Conclusion. It might be imagined that the behaviors of AMAs with diverse consequentialist utility functions will tend to converge, so that errors in specifying the utility function will cause little harm. We hold that this view is probably mistaken.

Moral sentiments, including such specific sentiments as fairness and the prohibition of murder, are a universal property among human cultures (Brown 1991). Recent experimental work in moral psychology indicates that moral reasoning typically proceeds from unconscious intuitions, and that the verbal justifications presented for conclusions are largely post hoc rationalizations (Haidt 2001). AMAs developed using top-down approaches (Wallach, Allen, and Smit 2008) would actually be motivated by their explicit moral theories, rather than a hidden underlying intuitive system, and could carry ethical systems to what their creators would call extremes. A machine intelligence, unless carefully designed otherwise, will be indifferent to any informal intuitions or desires that its creators failed to specify for it.

The difficulty of formally specifying such intentions is easily underestimated: the history of AI has shown repeatedly that humans miss the complexity of tasks to which they are well-adapted, e.g. language understanding or vision (Russell and Norvig 2003). Judging that something is a 'pain' in one's own case seems simple, a direct perception, obscuring the lack of any physical definition.

3. Conclusions

How can machine ethics deal with the problem of specifying utility functions for AMAs that are compatible with our purposes in creating them? Creating a top-down (Wallach, Allen, and Smit 2008) AMA that shares our morals requires a full specification of those values—in other words, a complete moral philosophy with decisions on all the relevant points of contention. Expecting a reasoned consensus to emerge on this point through pure philosophical investigation seems overly optimistic. Bottom-up approaches might seem like a way to avoid that daunting requirement by inferring human values from training data, but this would involve a serious danger that the agent would learn the wrong values, generalizing inappropriately from the training set of moral situations with harmful results (Yudkowsky 2008).

Instead, we suggest that an external boost to direct ethical theorizing is required. Efforts in the fields of neuroscience, experimental philosophy, and moral psychology have recently provided powerful insights into the structure of our moral values and intuitions (Haidt and Graham 2007; Koenigs et al. 2007; Knobe 2003), and it is reasonable to expect further gains. In the full paper we will also discuss possible approaches for machine ethics to partially compensate for limitations in the state of moral psychology as the field develops.

References

Bernoulli, Daniel. (1738) 1954. "Exposition of a New Theory on the Measurement of Risk." *Econometrica* 22 (1): 23–36. doi:10.2307/1909829.

Binmore, Ken. 2009. "Interpersonal Comparison of Utility." In *The Oxford Handbook of Philosophy of Economics,* edited by Harold Kincaid and Don Ross, 540–559. New York: Oxford University Press. doi:10.1093/oxfordhb/9780195189254.003.0020.

Bostrom, Nick. 2003. "Astronomical Waste: The Opportunity Cost of Delayed Technological Development." *Utilitas* 15 (3): 308–314. doi:10.1017/S0953820800004076.

———. 2009. "Pascal's Mugging." *Analysis* 69 (3): 443–445. doi:10.1093/analys/anp062.

Brown, Donald E. 1991. *Human Universals.* New York: McGraw-Hill.

Chalmers, David John. 1996. *The Conscious Mind: In Search of a Fundamental Theory.* Philosophy of Mind Series. New York: Oxford University Press.

Cowen, Tyler, and Jack High. 1988. "Time, Bounded Utility, and the St. Petersburg Paradox." *Theory and Decision* 25 (3): 219–223. doi:10.1007/BF00133163.

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

Glover, Jonathan. 1977. *Causing Death and Saving Lives.* Pelican Books. New York: Penguin.

Gustason, William. 1994. *Reasoning from Evidence: Inductive Logic.* New York: Macmillan.

Haidt, Jonathan. 2001. "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment." *Psychological Review* 108 (4): 814–834. doi:10.1037/0033-295X.108.4.814.

Haidt, Jonathan, and Jesse Graham. 2007. "When Morality Opposes Justice: Conservatives Have Moral Intuitions That Liberals May Not Recognize." *Social Justice Research* 20 (1): 98–116. doi:10.1007/s11211-007-0034-z.

Hardin, Russell. 1982. *Collective Action.* Baltimore, MD: Resources for the Future.

Harsanyi, John C. 1977. *Rational Behaviour and Bargaining Equilibrium in Games and Social Situations.* New York: Cambridge University Press.

Hausman, Daniel M. 1995. "The Impossibility of Interpersonal Utility." *Mind,* n.s., 104 (415): 473–490. http://www.jstor.org/stable/2254638.

Knobe, Joshua. 2003. "Intentional Action and Side Effects in Ordinary Language." *Analysis* 63 (3): 190–194. doi:10.1093/analys/63.3.190.

Koenigs, Michael, Liane Young, Ralph Adolphs, Daniel Tranel, Fiery Cushman, Marc Hauser, and Antonio Damasio. 2007. "Damage to the Prefrontal Cortex Increases Utilitarian Moral Judgements." *Nature* 446 (7138): 908–911. doi:10.1038/nature05631.

McCarthy, John. 1979. "Ascribing Mental Qualities to Machines." In *Philosophical Perspectives in Artificial Intelligence,* edited by Martin Ringle. Atlantic Highlands, NJ: Humanities Press.

Moor, James H. 2006. "The Nature, Importance, and Difficulty of Machine Ethics." *IEEE Intelligent Systems* 21 (4): 18–21. doi:10.1109/MIS.2006.80.

Nozick, Robert. 1974. *Anarchy, State, and Utopia.* New York: Basic Books.

Parfit, Derek. 1986. *Reasons and Persons.* New York: Oxford University Press. doi:10.1093/019824908X.001.0001.

Place, Ullin T. 1956. "Is Consciousness a Brain Process?" *British Journal of Psychology* 47 (1): 44–50. doi:10.1111/j.2044-8295.1956.tb00560.x.

Putnam, Hilary. 1960. "Minds and Machines." In *Dimensions of Mind: A Symposium,* edited by Sidney Hook, 148–179. New York: New York University Press.

Russell, Stuart J., and Peter Norvig. 2003. *Artificial Intelligence: A Modern Approach.* 2nd ed. Upper Saddle River, NJ: Prentice-Hall.

Searle, John R. 2007. "Biological Naturalism." In *The Blackwell Companion to Consciousness,* edited by Max Velmans and Susan Schneide, 325–334. Malden, MA: Blackwell. doi:10.1002/9780470751466.ch26.

Singer, Peter. 1972. "Famine, Affluence, and Morality." *Philosophy & Public Affairs* 1 (3): 229–243. http://www.jstor.org/stable/2265052.

Smart, R. N. 1958. "Negative Utilitarianism." *Mind,* n.s., 67 (268): 542–543. http://www.jstor.org/stable/2251207.

Wallach, Wendell, Colin Allen, and Iva Smit. 2008. "Machine Morality: Bottom-Up and Top-Down Approaches for Modelling Human Moral Faculties." In "Ethics and Artificial Agents." Special issue, *AI & Society* 22 (4): 565–582. doi:10.1007/s00146-007-0099-0.

Yudkowsky, Eliezer. 2008. "Artificial Intelligence as a Positive and Negative Factor in Global Risk." In *Global Catastrophic Risks,* edited by Nick Bostrom and Milan M. Ćirković, 308–345. New York: Oxford University Press.