A conversation with Carl Shulman, August 19, 2016

Participants

- Carl Shulman – Research Associate, Future of Humanity Institute
- Luke Muehlhauser – Research Analyst, Open Philanthropy Project

Note: These notes were compiled by the Open Philanthropy Project and give an overview of the major points made by Mr. Shulman.

Summary

The Open Philanthropy Project spoke with Mr. Shulman of as part of its investigation into which types of beings should be of moral concern, and thus a potential target for the Open Philanthropy Project's grantmaking. This conversation focused on cognitive procedures that might be useful when considering one's moral intuitions and judgments (e.g. about which beings should be of moral concern or not). These included identifying the potential sources of one's current moral intuitions, considering various types of conditions and constraints under which one's current moral intuitions might have been different, and techniques for projecting what moral policies one would be likely to endorse in light of different or additional facts or experiences.

Potential sources of our moral intuitions

Mr. Shulman suggests that it can be helpful to identify (conscious or unconscious) cognitive processes that are potentially responsible for producing one's current moral policies, judgments, and intuitions, and then, based on one's (uncertain) models of those processes, consider how they would function, and what moral policies, judgments, and intuitions they would likely produce in a large variety of cases (e.g. in different evolutionary, social, market, or environmental circumstances). Based on those results, one can then apply some criteria to determine which processes one endorses upon reflection as reasonable sources of moral policies, judgments, and intuitions (hereafter, we'll abbreviate these three things as "moral intuitions").

Example involving evolutionary reasoning

For example, one could consider which moral intuitions evolution would be likely to produce under various constraints.

When proposing an evolutionary explanation for a trait, it is important to be able to point to some specific mechanism by which the trait might have at least in principle been encoded and selected for (or produced as a byproduct). Mr. Shulman thinks that people sometimes give generic accounts of evolution "favoring" a complex trait (e.g. an entire value system) without accounting for the complexity that would be

needed for a mechanism to directly specify that trait. Evolution is better understood as selecting for small adjustments to parameters that might end up contributing to a creature's observed values or preferences (e.g. "increased response to smiling"), rather than directly encoding value systems. Mr. Shulman suggests it is useful to identify parameters like this in human psychology that could be adjusted to different degrees along various axes and which observably influence our judgments.

Real-life exposure vs. abstract consideration

In thought experiments involving personal identity (such as classic "teletransporter" scenarios), a common intuition is that even an exact copy of oneself (i.e. an entity physically separate from oneself, but with the same memories, etc.) is not a continuation of one's identity and that the persistence of the copy would not count as "survival" if the original were destroyed. Mr. Shulman thinks that an agent that was actually exposed to such situations regularly would experience pressure to develop different intuitions about identity than those of an agent considering such situations only theoretically.

Similarly, consider a hypothetical world containing booths that, for some fixed monetary payout, produce pain in the person using the booth and then erase both their conscious memory and the conditioning from the experience. Mr. Shulman suggests that people in that world would tend to develop moral intuitions in favor of using the booths when needed or desired, and would produce theoretical arguments in favor of a value system that does not consider (at least some) unremembered experiences as morally relevant.

Unremembered pain and reinforcement learning

Mr. Shulman distinguishes between two types of "memory": episodic memory of particular events, and reinforcement learning (i.e. the adjustment of policies in response to positive or negative reinforcement).

A reinforcement learner would likely learn to treat situations producing (episodically) unremembered pain differently than situations producing no pain, because it will experience pressure to learn policies that confer advantages going forward (e.g. avoiding future pain), even if the past pain is unremembered.

Transition of instrumental values to terminal values

Many of the values encoded as "terminal" values (i.e. “intrinsic goods”) in human cognition may have originally been merely *instrumental* values (means to an end) that, over time, began to function like, and be perceived as, terminal values. Because what is instrumentally valuable is strongly dependent on circumstances, it could be beneficial to consider what different instrumental values one might have acquired under different circumstances, and evaluate those results by some criteria.

For example, Mr. Shulman thinks that many items commonly included in “objective

list theories" of well-being are instrumental values that are now instilled, by some combination of evolution and reinforcement learning, as terminal values.

Of course, one might choose, upon reflection, to endorse the terminal values one has, even if they originated as instrumental values useful in some particular environment.

Imitative learning

Humans exhibit a strong instinct for imitative learning (e.g., babies' behavior, the copying of speech patterns, the persistence of seemingly arbitrary customs — see Joseph Henrich's *The Secret of Our Success* for additional examples). In particular, people appear to learn taboos effectively by imitation. Imitative learning as a source of values seems particularly malleable and susceptible to random influences that we would not endorse upon reflection.

But again, it is possible that one would choose, upon reflection, to endorse many of the values originally acquired via imitation learning.

Virtue ethics as signaling

It may be useful to view certain kinds of moral intuitions, in particular those that endorse "virtuous" traits, in terms of the signaling value of those traits among partially transparent agents (i.e. agents that can only partially conceal their mental states from others). This is not to say that we necessarily consciously perceive these moral intuitions as signals to our community, but rather than these signaling dispositions, loyalties, and reliable behavioral features were selected for (socially, culturally, and perhaps evolutionarily) because they make us valuable participants in a community, and effective enemies of out-groups. These traits may then have become instilled as strong moral intuitions. If so, we might or might not endorse this process as a "valid" producer of some of our moral intuitions.

Associational learning

Many intuitions seem to be generated by simple, model-free associative learning (i.e. "X tends to accompany Y", for example the "affect heuristic"), and we might or might not endorse this intuition-generating process, upon reflection.

Potentially useful procedures for evaluating moral intuitions

Valuing processes despite not being able to consciously observe them

Mr. Shulman thinks it is reasonable to expect that our moral intuitions, by default, would not treat some kinds of cognitive processes as morally relevant — specifically, those cognitive processes of which "we" (our central, stream-of-consciousness decision-making center) have no conscious awareness, e.g. the enteric nervous system, the non-dominant brain hemisphere, and other cognitive processes that are "hidden" from "our" conscious awareness. Upon reflection, Mr.

Shulman does not endorse this intuitive discounting of the moral value of these hidden-to-“us” cognitive processes.

Luke suggests that the intuitive discounting of cognitive processes that are not "visible" to our decision-making center may be analogous to one's moral intuitions not naturally valuing people who are not directly observed (such as physically distant people or potential future people).

Simulating social feedback

Mr. Shulman suggests that the ability of individuals to generate reasonable beliefs or policies often depends on social inputs. Typically, considerations of whether a claim is true, or what a reasonable policy would be in a given situation, make use of a balance between explicit reasoning and social feedback (e.g. through heuristics like imitation). Mr. Shulman suggests that successful individual policies are often generated by "outsourcing" some pieces of our judgments to social consensus.

When considering unusual situations for which one does not have any social inputs, it might be reasonable to expect to end up biased, on net, away from optimal policies or intuitions. For this reason, Mr. Shulman finds it useful to imagine incorporating the “extended mind” of society when considering unusual hypothetical scenarios, such as by considering what his intuitions or policies about a scenario would likely be if the scenario were societally common (e.g. considering whether a practice would continue to be morally condemned if it had already been widely adopted).

Social feedback may reduce risk aversion via imitation

When presented with situations with the option to achieve a positive outcome at a high probability (e.g. 90%), or some much smaller positive outcome with 100% probability, people tend to display extreme risk aversion that does not align with the expected value of the options. This is a well-replicated result.

However, people who have been exposed to real-life feedback from experiencing or witnessing large payoffs from low-probability events tend to display higher risk tolerance. Mr. Shulman suggests that, perhaps in part because of hard-wired tendencies to imitate the policies of higher-status individuals, people are prone imitate "winners" in low-probability, large-payoff situations (e.g. types of payoffs that only affect one person in 1,000). This imitative behavior could be viewed as approximating a more proper sensitivity to expected value than the typical, risk-averse responses people give intuitively in risk-reward scenarios.

Experimental philosophy

Mr. Shulman thinks that experimental philosophy has produced some research that is helpful for considering how moral intuitions vary by culture and other environmental variables, but he warns that some of this research is unreliable, underpowered, and would likely fail to replicate (just as many studies in social

psychology have failed to replicate).

"Copy clan" forecasting technique

As a thought experiment to help evaluate what his moral intuitions would be upon ideal reflection, Luke has considered attempting to forecast what conclusions he would reach if a large population of copies of himself learned many facts, considered many arguments, and underwent a large variety of experiences (e.g. different environmental or societal circumstances, different histories, etc.) and then collectively advised him on what his values should be.

Mr. Shulman expects this technique to have some value. However, because it is not possible to get literal feedback on one's predictions of what hypothetical copies of oneself would come to endorse in different circumstances, he suggests it might be beneficial to also make predictions that do offer real-world feedback, e.g., predicting what one's position on a moral question will be one month from now, or what the results of an opinion survey on these topics will be. Part of the purpose of such an exercise would be to learn whether one's forecasts are, at least in the checkable case, adding any information value.

Testing the boundaries of the thought experiments

Mr. Shulman thinks there are limits to some of the methods described above: if one (in a thought experiment) adjusts the factors relevant to one's intuition-generating mechanisms radically enough, then none of one's existing values would be preserved, and in some cases no values at all would be preserved — after all, most possible mind designs do not have any coherent moral values. This calls for some pruning of which kinds of "reasonable" factors to consider, in pursuit of reflective equilibrium.