"Prospects for a Smithian Machine"
Thomas M. Powers
University of Delaware

ABSTRACT:

Building on recent empirical work on the psychology of moral judgment, I will investigate whether a machine could be programmed to make moral judgments according to a distributed, crowd-sourced model of empathy. The basic notion under consideration assumes that, while computers are themselves incapable of any emotional response, they might one day be capable of depicting formally the emotional responses of people in specific contexts and learning from them. Such a system might rely on a networked web similar to RoboEarth, which is a European research program to build "a worldwide, open-source platform that allows any robot with a network connection to generate, share, and reuse data." (IEEE Spectrum, 2011). I will argue that this crowd-sourced type of moral judgment will be possible for computers, and not (terribly) risky for the humans interacting with them because the judgments will conform to an inter-subjective and "massively multi-player" standard that is originally generated by human moral judgments.

INTRODUCTION

Moral philosophy, and especially work on moral judgment, has taken an empirical turn in the last two decades (Greene, 2003, 2007; Greene and Haidt, 2002; Greene, et al., 2001). With experimental results from fMRI, philosophically-oriented empirical psychologists have been able to study the phenomenon of moral judgment and associate that ability with other cognitive abilities by brain location. The general consensus of these empirical studies is that moral judgment in humans is more emotional than rational, that it functions (at least in part) by way of empathy, and that empathic ability is shared by higher primates possessing "mirror neurons." These results indicate evidence of an evolutionary basis for the ability of moral judgment, as well as a need for further empirical studies.

Building on this empirical work, I will investigate whether a machine could be programmed to

make moral judgments for its own actions according to a distributed, crowd-sourced model of an empathic agent. The basic notion under consideration assumes that, while computers are themselves incapable of any emotional response, they might one day be capable of depicting formally the emotional responses of people to particular situations or contexts. The computers would then conform their actions to produce expected approval in humans—much like the moral accounts of Smith and Hume suggest. Given a kind of software/hardware equivalent of the mirror neurons, machines could compute conformity by generating and communicating data, and also using data of many other machines that likewise generate, communicate, and use data from their interactions with humans.

The key to the appeal of this basic notion is that the computer itself need not experience emotions in order to learn from them. Specifically, it could learn what moral judgments to make relative to its own prospective actions—and this is the most important goal for machine ethics. I think that the computer need not learn to *evaluate* the actions of human beings; in this sense the empathic computer will develop only self-regarding moral judgments. It will not judge human actions.

In proposing an account of moral judgment in machines that uses data about emotion, I do not intend to side with "sentimentalist" as opposed to "rationalist" accounts of moral judgment more generally. Indeed, by comparing Smith and Kant, I will show how a Smithian machine would also share aspects of rationalist accounts of moral judgment, and argue that in the end a Smithian machine would actually be a hybrid "rational" and "emotional" agent. But first I will conceptualize a machine that behaves according to Adam Smith's "sentimentalist" or emotion-

centered theory of ethics (Smith, 1976) in a rather pure and abstract form. While a Smithian account may seem at first blush to be quite unlikely for a machine—AI is one thing, and Artificial Emotions quite another—I believe that a Smithian machine fairs at least as well as a Kantian machine, the prospects of which I have explored earlier (Powers, 2006).

The split between "rationalist and "sentimentalist" views of moral philosophy survives to this day in the recent empirical brain research that I mentioned earlier. Some researchers (Rizzololatti and Craighero, 2005; Talmi and Frith, 2007) believe that the experiments have ended the dispute in favor of the sentiments. After the "Parma Group" discovered the function of mirror neurons in a study of macaque monkeys, an evolutionary and neurophysiological basis was strongly indicated for the empathic ability in humans. Indeed, it appears that Smith's view, or something close to it, has been vindicated since one can now find neuroimaging studies which mention Smith as an influence.

Though there has been some confusion in various fields about the precise nature of empathy in primates (Preston and de Waal, 2002), I will assume the following definition: empathy is the ability to understand the feelings of others, and associate those feelings with their respective contexts. The first part of this definition is widely accepted; the latter associative clause will be important in order to model feelings with situations or contexts for a machine. With this definition in tow, let us see how a Smithian machine would have to sense, communicate, and compute in order to arrive at a moral judgment. After, we will see that there is a Smithian model for human moral judgment that is at least suggestive of the data structures for our prospective Smithian machine.

DATA STRUCTURES AND LEARNING

The Smithian machine would learn by the confluence of two data streams. First, from the data it generates of the emotional responses of human subjects in its actual field of agency (the humans that could be affected by some decision/action it makes), it could build a preliminary account of the emotions in humans that arise in various contexts in which it would be considered an agent. Second, those responses and contexts could then be compared to data sets generated by other computers. Such systems could communicate via a networked web similar to RoboEarth, which is a European research program to build "a worldwide, open-source platform that allows any robot with a network connection to generate, share, and reuse data." (IEEE Spectrum, 2011).

The relational database for the Smithian machine generates the following structure:

I. Each subject S_1-S_n exhibits emotions E_1-E_n in context C_1-C_n when faced with action A.

By sharing indexed information in the networked database, the machine could learn:

II. With frequency f, subjects S exhibit emotions E in contexts C when faced with A.

III. With frequency f, subjects S perform actions A in response to E in context C.

IV. With frequency f, Actions A in contexts C feed back to produce emotions E.

Given this information, the machine could compute i) how it typically ought to act in context C_n

to avoid untoward emotions in human subjects, and ii) whether the subjects in its field of agency are statistical outliers. After significant generation, use, and reuse of these structured data, the networked database would also yield information concerning iii) the defeating conditions under which statistically-generalizable emotional responses fail. Over time, the network might yield other sophisticated results about the interplay of emotion, action, and context in human societies.

The main conclusion for which I wish to argue is that such a crowd-sourced type of moral judgment is conceptually possible for computers, given some assumptions about the development of the networked database and the ability of machines to sense, identify, and evaluate emotional responses in humans. While establishing the empirical possibility of these assumed abilities is beyond my knowledge of the fields of psychology and computer science, I do not think that emotion is scientifically intractable. In a review of work on the measurement of emotion, Mauss and Robinson (2009) cite whole-body indicators of emotional states, as well as measurable component states such as body posture, vocal characteristics, and (with electromyography) facial behaviors. In the future, machines will likely be able to sense and aggregate these data, as well as make other measurements correlated with emotions such as facial flushing, body temperature and blood flow dynamics, and perhaps even neuronal activity localized to brain regions. I assume here that the computer will be able to make such measurements without making contact with or interfering with the human subjects in its field of agency.

If psychologists and computer scientists are able to establish correlations between measurable effects and emotion, the Smithian machine will be on its way to going "online" and active. I

believe--but am open to being convinced otherwise--that such moral abilities in machines will not be too risky for the humans interacting with them because the machine moral judgments will conform to an inter-subjective and "massively multi-player" standard that is originally generated by human emotional responses and moral judgments. In fact, this conception of machine moral judgment seems safer than "programming" a morality into a machine—the outcome of which would depend heavily on the programmers' views about morality. Wallach and Allen (2009) have already explored the problems with such a top-down approach. The safety latch with the Smithian machines is that they would merely follow the (generalized) moral judgments that humans make in response to human emotions in various contexts.

The appeal of an inter-subjective standard may allow the young field of machine ethics to avoid a roadblock. The main question of machine ethics—how to program a computer for ethical or moral behavior—faces a persistent difficulty related to a larger (and much older) philosophical debate. What is the correct ethical theory? My account of a Smithian machine would appear to circumvent that road block because it does not take sides on the issue of the correct ethical theory. But this aspect of the account may also be a weakness, for somehow the machine will have to distinguish proper from improper emotional responses in humans, independent of their frequency. Prinz (2011) has explored the ways in which empathy might lead humans astray in moral judgment, and I think there are valid reasons for concern with an empathic machine. For human morality, the rationalist moral accounts seem to have an advantage over purely sentimentalist ones, for rationalists would consider not just the emotional response to action in context, but whether the response is justified. In fact, some untoward emotions in humans seem to be justified—for instance, the distress in response to a procedurally administered punishment

when someone has broken the law. Disapproving or untoward emotions in humans are not an unequivocal guide to whether that human has been wronged.

Yet it may not be that the account of moral judgment that Smith had in mind was as "sentimentalist" as some contemporary empirical psychologists (especially Greene) believe (Greene, 2007). In this next section I explore the ways in which Smith's model of empathy relies on rational—and hopefully computable—processes that could help us to avoid the worry about properly untoward emotions.

SMTIH AND KANT

I do not think that the moral theories of Smith and Kant are so far apart, and consideration of the cognitive processes required for the Smithian view will help to show why. Smith proposed an account of the moral sentiments, primarily what he called sympathy, by means of which we make moral judgments of merit, propriety, and justice from the standpoint of an impartial spectator or "inner judge". The contemporary term 'empathy' is what Smith is now considered to have meant (Prinz, 2011). Kant, as a moral rationalist, believed that moral judgment should neglect feelings as data. He held that correct moral judgments are possible just when reason alone determines the will of the agent to act, without regard to personal inclination, sentiment, or consideration of effects. He did have an account of moral feelings (in the second *Critique*), but on this view the only emotions worth considering were those of guilt and awe at the magisterial moral law within us. The moral feelings are derivative, for Kant, and not constitutive of correct moral judgments.

To explore the difference between Smith's inner judge and Kant's moral law, consider Kant's Categorical Imperative of universalization. It does suggest a way in which the views of others could be aggregated so that an agent could reach a "crowd-sourced" decision about the propriety of a particular action (or, for Kant, maxim). In notes that predate his characterization of the Categorical Imperative by 20 years, Kant mentions our "ability to take moral positions as a heuristic means. For we are social beings by nature, and what we do not accept in others, we cannot sincerely accept in ourselves" (Rossvaer, 1979). Is this heuristic the prototype of our Smithian machine?

The key difference between Smith and Kant is that, for Kant, the appreciation of others' views does not accept their feelings as the raw data by which to determine a universalized moral judgment. For Smith, others' emotions are the raw data, and thus lead us to certain actions and away from others in order to produce social concord. The goal of social concord, as we will see in this next section, is the driver for moral judgments.

A SMITHIAN MODEL

Smith's *A Theory of the Moral Sentiments* (TMS) introduces an account of *extended inclination.* In TMS, Smith claims that "[a]s we have no immediate experience of what other men feel, we can form no idea of the manner in which they are affected, but by conceiving what we ourselves should feel in like situation" (TMS 9). When we engage in this limited act of "fellow feeling," we are able to pass judgment on the propriety of someone's actions, on his merit or demerit (to be paid in gratitude or resentment), and on the justice of the acts (reward and punishment). The

mechanism of this judging assumes a reciprocal feeling of being judged, for when the observed "views himself in the light in which he is conscious that others will view him… [he will] act so as that the impartial spectator may enter into the principles of his conduct... and bring it down to something which other men can go along with" (TMS 83).

The impartial spectator allows a transformation of perspective and provides the observer and the observed with an "inner judge." The ideal outcome for the person being judged (observed) is to modulate feelings (and attendant behaviors) of self-love so that the observer's feelings somehow match. In this way, Smith thought, the cause of social concord is advanced. Clearly, for Smith, the sentiments are the original data of moral judgments, but the impartial spectator tempers them with considerations of social concord.

There is no "inner judge" for Kant, in the sense in which Smith intends it. The perspective of the judge is "external"—what Nagel (1989) criticized as "the view from nowhere." Reason is the faculty that extends to all humans and provides their common basis for judgment. One wonders, though, if there is any point in this dispute—why must sentiment be internal and reason external?

One reason to follow Smith's internalist model comes from the results in empirical psychology—that whether we or not we believe we are rationalist moral theorizers, in practice we are all sentimentalists. When making moral judgments about the Trolley Problem and similar "moral dilemmas," the brains of most human test subjects do not show activity in areas associated with "higher" reasoning, but do show activity in areas of as "lower" empathizing (Greene, et al, 2001; Greene, 2003).

Taking these results at face value, we must still *suppose* that the test subjects are making moral judgments when they are asked to do so. Another explanation of the results would be that, when prompted by the experimental scenario, they may be lazy moral reasoners, or they may be distracted, bored, confused, etc. Or the experimenters may ask questions that they believe require moral judgment, but in fact do not. The test scenarios might call forth in the test subjects an ambiguous judgment--one that could be drawn from either rationalist or sentimentalist theories, and hence going from what the brain does to how the subjects morally judge would be difficult at best. The measurements required to justify the psychologists claims could be more precise than the technology affords; there is reason to think that this is true in some cases of fMRI (Illes and Racine, 2005; Hernandez, et al., 2002; Diedrichsen and Shadmehr, 2005). Finally, experimenters may be supplying scenarios and pictures to test subjects which bring forth a variety of cognitions—empathic, perceptual, associative, imaginative, and, in addition, genuine instances of moral reasoning. This would be the case if psychologists were to supply realistic examples of moral judgment, as opposed to (what I take to be) the quite fantastical and tragic scenarios related to the Trolley Problem.

What psychologists ought to observe, in cases of genuine moral reasoning, should indicate a complex moral psychology. Indeed, on my view a complex moral psychology is required by any plausible moral theory for primates. Such a complex moral psychology would have to have inputs from "internal" experience of the relations mentioned in the sketch of the Smithian machine: structured emotions over contexts, as a reaction to certain actions. It would also have "external" (or at least non-subjective) feedback about the consistency of emotional responses

over time, as well as the "crowd-sourced" external standard of what is socially accepted in terms of emotional response.

So Kantian moral judgments, on my view, can be complex, time consuming, dependent on probabilities, evocative of emotions, sensitive to past commissions and omissions of moral agents and patients, forward looking to concerns of social concord, protective of human values like dignity and autonomy, and still be (mostly) correct. If I am right, it would be a rare empirical psychologist who could pick the Kantian out of a crowd of test subjects just by looking at moral dilemma responses and brain activity.

If I am right, a rapprochement is plausible between sentimentalist and rationalist theories of moral judgment, and hence between Smith and Kant. Rationalism in moral theory is not restricted to the psychology required by "ethics as universalization." Rationalists can hold that moral reasoning is of a kind with means/ends deliberations (instrumental rationality), ranking of goals, explaining commissions and omissions, strategizing, and other reason-based processes. These reason-based processes will be subject to "internal" *and* social or "external" feedback, with all of the attendant guilt, stubbornness, and self-righteous indignation that we've come to experience in our species.

CONCLUSION

In the opening pages of TMS, Smith himself suggests the model for the Smithian machine when he writes:

> "When I endeavor to examine my own conduct, when I endeavor to pass sentence upon it, and either to approve or condemn it, it is evident that, in all such cases, I divide myself, as it were, into two persons; and that I, the examiner and judge, represent a different character from that other I, the person whose conduct is examined into and judged of."

Still, one wants to know the basis on which Smith's judge will judge himself. So it seems that we come back once again to the problem of finding the correct ethical theory.

The Smithian machine resolves that problem by allowing the "judge" to be whatever aggregated judgments have been justified by the knowledge of the network. Machine moral judgments circumvent the problem of the "correct theory" by means of an inductive method. They follow complex moral structures generated by human beings, instead of trying to introduce new structures themselves. In this way, the prospects for a Smithian machine seem good.

REFERENCES

Diedrichsen, J., and Shadmehr, R. (2005). "Detecting and adjusting for artifacts in fMRI time series data,". *Neuroimage*, 27 (3), 624-634.

Hernandez, L., et al. (2002). "Temporal Sensitivity of Event-Related fMRI," L., *NeuroImage* 17 (2), 1018-1026

IEEE Spectrum (2011). RoboEarth: a world-wide web for robots, http://spectrum.ieee.org/automaton/robotics/artificial-intelligence/roboearth-a-world-wide-web-for-robots, accessed 5/08/13.

Illes, J. and Racine, E. (2005). "Imaging or Imagining? A Neuroethics Challenge Informed by Genetics,", *American Journal of Bioethics* 5(2), 5-18

Greene, J. (2003). "From neural "is" to moral "ought": what are the moral implications of

neuroscientific moral psychology?" *Nature Reviews Neuroscience* 4, 847-850.

Greene, J. (2007) "Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains" *Trends in Cognitive Sciences* 11(8), 322-323 (2007)

Greene, J. and Haidt, J. (2002)."How (and where) does moral judgment work? *Trends in Cognitive Sciences* 6(12), 517-523.

Greene, J., et al. (2001). "An fMRI Investigation of Emotional Engagement in Moral Judgment" *Science* 293, 2105-08.

Mauss, I. and Robinson, M. (2009). "Measures of Emotion: A Review" *Cognition & Emotion* 23(2), 209-37.

Nagel, T. (1989). *The View from Nowhere*. New York: Oxford University Press.

Powers, T.M. (2006). Prospects for a Kantian Machine, *IEEE Intelligent Systems* 21 (4), 46-51.

Prinz, J. (2011). "Is Empathy Necessary for Morality?" in P. Goldie and A. Coplan (Eds.), Empathy: Philosophical and Psychological Perspectives. New York: Oxford University Press.

Rizzololatti, G. and Craighero, L. (2005)."Mirror neuron: a neurological approach to empathy," in *Neurobiology of Human Values.* Heidelberg: Springer Publishing.

Rossvaer, V. (1979). *Kant's Moral Philosophy*. Oslo: Universitetsforlaget.

Smith, A, (1976 [1759]). *The Theory of Moral Sentiments.* New York: Oxford University Press.

Talmi, D. & Frith, C. (2007). "Neurobiology: Feeling right about doing right" *Nature* 446, 865-866.

Waibel, M., et al (2011). RoboEarth – A World Wide Web for Robots. In *Robotics & Automation Magazine*, IEEE, 18 (2), 69-82.

Wallach, W. and Allen, C. (2009). *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press.