Metaethics in Context of Engineering Ethical and Moral Systems

Lily Frank

Philosophy and Ethics, Innovation Science,
Technological University of Eindhoven,
5612 AZ, Eindhoven, Netherlands

l.e.frank@tue.nl

Michał Klincewicz

Institute of Philosophy, Department of Cognitive Science,
Jagiellonian University,
Grodzka 52, Kraków, Poland
michal.klincewicz@uj.edu.pl

Abstract

It is not clear to what the projects of creating an artificial intelligence (AI) that does ethics, is moral, or makes moral judgments amounts. In this paper we discuss some of the extant metaethical theories and debates in moral philosophy by which such projects should be informed, specifically focusing on the project of creating an AI that makes moral judgments. We argue that the scope and aims of that project depend a great deal on antecedent metaethical commitments. Metaethics, therefore, plays the role of an Archimedean fulcrum in this context, very much like the Archimedean role that it is often taken to take in context of normative ethics (Dworkin 1996; Dreier 2002; Fantl 2006; Ehrenberg 2008).

Realism and antirealism

One of the divisions in metaethics is between realism and antirealism. Realist views claim that the universe features objective and mind-independent moral properties. Moral theories aim to "get it right" with respect to these properties in a way similar to how science aims to "get it right" about the properties of relevance in its respective disciplines. Moral realism is the view that moral judgments either succeed or fail in corresponding to a mind independent moral reality. Antirealist views reject realism and typically also preclude moral theories that aim to "get it right" about moral properties. Antirealists typically deny that moral properties or facts are objective or mind independent in the same way the realist takes them to be.

It is important to note that this binary division by no means exhausts the possible metaethical positions. Views like Christine Korsgaard's Kantian constructivism, Simon Blackburn's quasi-realism, and Alan Gibbard's norm-expressivism provide alternatives to the strict realist/anti-realist divide (Korsgaard 1996; Blackburn 1993; Gibbard 1990). These diverse views, as well as others, aim to capture the realist and objectivist phenomenology of moral language and thought, while making minimal ontological commitments. However, for the purposes of this paper we will focus on a simplified division between realism and antirealist positions.

Our expectations for engineering an AI that does ethics in any sense must, at least to some extent, be conditional upon one of these views. Realists can aim at the ambitious goal of engineering a machine that has the potential to detect moral properties or give new insight into moral facts. A realist can also entertain the hope that a machine can help bridge the troublesome epistemic gap between our limited cognitive faculties, the biases we are vulnerable to, our incomplete understanding of causal relations etc., and the moral facts that exist independently of what we happen to believe or think of them. Of course, just the possibility of such a machine leaves open a multitude of other ontological questions about what it is we are at-

tempting to detect: the properties of events, things, persons or perhaps something altogether different.

Using a simplifying analogy, we can think about an AI capable of making moral judgments by detecting moral properties themselves as akin to a microscope. The microscope allows us to observe the existence and behavior of microscopic organisms and better understand the relations between their presence and a disease or to confirm or disconfirm various hypotheses about infection. With the help of a moral microscope in the guise of an AI, in a similar way our hypotheses about what is right and wrong could be confirmed or disconfirmed.

A fuller discussion of the possibilities of such a device would involve not only a defense of the metaethical ontological assumptions about the existence of moral properties, but also of assumptions about how we acquire moral knowledge. All this should ideally be settled before we would start speculating about what kind of tool would be able to detect moral properties or be useful in improving on our ability to acquire moral knowledge;a lot of philosophical work would have to get done first. But assume this work is done or we simply take a realist metaethical position for granted. At that point we still have to decide whether the moral microscope AI is going to be a machine that helps us to settle first order normative questions or whether it advises us on the permissibility of a particular action. Alternatively, perhaps its role is to help us adjudicate between competing normative theories,such as Kantianism or utilitarianism, and provide us with the foundational elements of right and wrong. Speculation about such devices invites criticism from several angles. First, the task of creating models of these devices and then implementing them is daunting. Second, we face the additional problem of verification of whether the machine is working at all.

Whatever result the moral microscope AI produces, it would presumably be in the form of an answer to some first order moral question, such as "should X do Y now?" The problem is that verifying whether the machine is giving us a morally acceptable answer can only be made in light of our assessment of its consistency with other values, principles, and cases, or how well it fares when put through a process of wide reflective equilibrium (Rawls 1975). This raises the philosophical question of whether we should have more confidence in the judgments of the AI than we have in our own well considered deliberative conclusions regarding difficult moral questions.

None of this implies that the idea of a moral microscope AI is a non-starter. We can envision several ways in which an AI that issues moral judgments could be useful, given background realist metaethical assumptions. If we assume, for example, that moral properties are causally efficacious (as prominent versions of naturalistic realism do), perhaps the machine could detect the effects of moral properties to which we are not sufficiently sensitive (Sturgeon 1985, 1986). Or the moral judgments that the AI issues could be free of cognitive distortions such as scope insensitivity, which is the phenomenon of our feelings of empathy and willingness to aid each individual decreasing with the increase in numbers of individuals (Persson and Savulescu 2012, p. 30).

For an antirealist the idea that a machine could play the role of a microscope is completely misguided. Moral facts or properties are not out there in the world to be found. They are instead features of our beliefs, emotive reactions, or attitudes towards what is otherwise a morally neutral world. In Hume's words we gild or stain the world with morality through the projection of our sentiments (Hume 1975/1751 p. 294). The antirealist can only expect an AI to do whatever we do when we form moral judgments. For example, moral error theory, a type of antirealism, claims that although our moral judgments and moral language have the form assertions or propositions. This means that moral judgments are consistently false when they are used in ways to make claims about particular things being right or wrong, virtuous or vicious, etc (Mackie 1977; Joyce 2001). An AI created with this metaethical stance informing it would be radically different from the one created with a realist theory in mind.

In contrast to the analogy of the microscope, an AI that forms moral judgments in a way that simulates what we do when we form moral judgments (replicating our limitations, flaws, etc.) could be likened to weather forecasting model. Given certain inputs the model would make predictions about the likely outputs. It could presumably also engage in the expression of attitudes (some bad weather ahead), imperatives (do not go outside!) or prescriptives (use caution while driving).

The simulation approach is compatible with all metaethical positions, not just antirealism. This is because the goal of making such a device is to replicate some features of human psychology, irrespective of whether this psychology is responsible for tracking actual moral properties in the world or just projecting them. Consequently, the set of metaethical questions that are most relevant to this project are those that have to do with the kind of mental state that a moral judgment is and what kinds of capacities for motivation, emotion, etc. are required to be able to form that mental state.

Cognitivism, non-cognitivism, internalism and externalism

The project of simulating moral judgments in an AI can have strong and weak ambitions. The strong project will take for granted that the project of so-called strong artificial intelligence is viable, meaning, that we can create human level mental states, such as beliefs and desires, in a machine. The weaker project would aim at replicating some subset of abilities or capacities, without the presumption that these are actually mental states in any sense. Both of these strategies have to address the same basic metaethical questions.

The most important question is about the ontology of moral judgments themselves. An answer to it conditions what it would take to implement human level moral judgements in a machine. One part of the ontological problem of moral judgments is that it is highly controversial what people do, exactly, when they make them. Metaethicists disagree whether moral judgments are beliefs, desires, expressions of emotions, or some combinations of these. The second part of the ontological problem is that we need to at least provisionally settle on the relationship between moral judgments and moral motivation. So, before we begin engineering or modeling human moral judgments, a host of positions in metaethics needs to be either settled or assumed.

One critical distinction in the ontology of moral judgments is between cognitivism and non-cognitivism. As a view about moral psychology, cognitivism is the position that moral judgments are primarily beliefs and thus not conative states, such as pro-attitudes or desires. As a view about language, cognitivism is the position that moral language is truth apt, meaning, that moral sentences express propositions that can be evaluated as true or false or that moral terms refer to moral properties.

Non-cognitivism is the view that moral judgments are conative states, such as desires, emotions, or pro-attitudes, and not beliefs. As a view about language, non-cognitivism typically implies that moral language is not truth apt, that is, moral language does not assert that something is the case.

There have recently been a number of arguments for views that reject the cognitive-non-cognitive dichotomy altogether. In place of that dichotomy some theorists offer hybrid views, which characterize moral judgments as mental states that have features of both beliefs and desires or views that characterize moral judgments as consisting of more than one type of mental state. For example, the "besire" theory, which posits the existence of one mental state. On this view, moral judgments have characteristics of a belief in that they purport to represent some state of affairs *and* the motivational component of desires in that the moral judgment itself is sufficient to move one to act on it (Altham 1986; Bedke 2009). Other hybrid views include non-descriptivist cognitivism (Horgan and Timmons 2000) and there are others (Ridge 2006).

These views each have their advantages and disadvantages. Cognitivism, for example, is thought to better capture some important surface features of moral discourse, such as its appearing to be descriptive. Another purported advantage of cognitivism is that it is consistent with moral language that suggests objectivity in cases when people hold conflicting moral judgments on a particular topic and assume that they cannot all be correct. Furthermore, people seem to recognize the possibility that they can be mistaken in their own moral judgments and that they can alter them in light of evidence—an assumption that fits best with the view that moral language is truth apt. Finally, cognitivism has the advantage of providing a straightforward picture of moral propositions. On this view, moral propositions have the same meaning whether they are in asserted or unasserted context (the Frege-Geach problem) (Geach 1964).

For non-cognitivists moral assertions such as 'Stealing is wrong' are understood as expressions of non-cognitive states like the speaker's negative attitude toward stealing. If non-cognitivism is correct, then it is hard to understand how moral terms can function as antecedents of conditionals, such as "if stealing is wrong, then stealing bread is wrong" or when they are used in valid arguments. In valid moral arguments, the moral terms are assumed to mean the same thing in all of the premises. But if they are merely expressions of attitudes or other non-cognitive states, then prima facia they do not mean the same thing in all contexts.

Non-cognitivism, on the other hand, has the advantage of explaining persistent moral disagreement that seemingly cannot be settled even when people agree on all of the relevant nonmoral facts. Non-cognitivists can argue that in such situations people who disagree are merely expressing two distinct affective states which need not be responsive to argument or evidence. Non-cognitivism also seems to be able to better explain the close relationship that appears to exist between making a moral judgment and acting on that moral judgment. Emotions, attitudes, desires, and prescriptions motivate us to act, propositions like beliefs do not.

Hybrid views face the criticism that it is impossible for one mental state to have both a world to mind and a mind to world direction of fit. which would Such a mental state would have to be a state that is both responsive to evidence and not responsive to evidence (Smith 1994 p. 118). However, hybrid views have the advantage of capturing some of the unique properties that moral judgments have, especially their motivational force, the authority that they claim in our mental lives and decision making, as well as their aiming to represent some state of affairs

all at once. The AI engineer that aims to create a simulation machine for moral judgments will have to navigate this technical debate and choose sides.

Things get even more complicated. The most technical aspect of the philosophical debate about the psychological nature of moral judgments concerns the relationship that moral judgments have to action and motivation. The discussion often begins with the observation that it is odd for someone to judge that something is morally wrong and subsequently claim that this gives them no motivation whatsoever to refrain from doing it. Hence, one of the unique features of moral judgments, as compared to other types of judgments, is assumed to be its special connection to motivation.

Many philosophers argue that to make a moral judgment is necessarily to have a (defeasible) motivation to act on it (Garrard and McNaughton.1998). This family of views, which are often called motivational internalism, claims that in order for a judgment to qualify as a moral judgment it has to be motivational itself, or necessarily bring with it a motivational state (like a desire). Internalist views come in different forms and can range from characterizations of moral judgments as expressions or pro or con attitudes towards some object, all the way to rationalist positions that claim that moral judgments include reasons and considerations of reasons are in themselves motivational states. If the task is to build an AI that makes moral judgments and internalism is the option the engineer opts for, then the AI would also have to be capable of being motivated by its moral judgments—whatever form they take.

The opposing view, externalism, holds that moral judgments do not necessarily or inherently motivate, nor can they motivate by themselves (Brink 1986, 1989). On this view, moral judgments are only contingently connected with motivation. This allows a situation in which a person forms a moral judgment (usually as a belief) and lacks the corresponding motivation to act on that judgment. When people are motivated to act morally, it is because they have a moral belief that connects up in the relevant way to a desire, for example, the desire to be a good person or the desire to help a friend.

Realists and antirealists alike have to take sides on the question about the relationship between moral judgment and motivation. This task is challenging for many reasons, not the least of which being that the very nature of motivation and the psychological state it is a proxy for is obscure. An additional difficulty is that an adequate account of this relationship has to contend with empirical data on the neurobiology of individuals with abnormal moral thinking or behavior (psychopaths, people affected by brain lesions, autistic individuals, etc.). An adequate account must also be able to explain the wide variability in the extent to which people are motivated by their moral judgments and the many ways in which people fail to be motivated by their moral judgments.

This brief review of some fairly coarse-grained metaethical distinctions suggests that the task of making an AI that does ethics in the sense of being able to make moral judgments, requires settling on many thorny metaethical questions about moral psychology and metaphysics of moral properties.

Does this matter to the engineer?

The metaphysics of moral properties and the nature of moral judgments and in general what it means to "do ethics" are subjects of philosophical dispute. This is why the task of creating an AI that does any of these things seems to be hostage to metaethics, specifically when it comes to questions about what moral judgments are, what moral properties are (and whether they exist at all), and whether motivation plays a central role in constituting moral judgments. What may be particularly disheartening about this situation is that there are currently no obvious scientifically based paths forward in settling these issues. But this does not mean that progress in this domain of AI research should halt and patiently wait for the philosophers to stop arguing. On the contrary, research in this area could lead to breakthroughs that may help the philosophers. Furthermore, the task of creating an artificial system that does any of the things we discussed here is in itself interesting and challenging. So, at the least, the engineer could simply take from metaethics whatever may be useful to that task and perhaps use it as a guide to their work.

Arguably, this situation is not very different from that which exists with the research program into artificial consciousness. Attempts to make an AI that has the capacity to be conscious or self-aware are in their infancy and there exists significant disagreement in philosophy on the nature of consciousness. Nonetheless, research in neuroscience and psychology, as well as in artificial intelligence that concerns consciousness, does not halt in anticipation of philosophical consensus. On the contrary, these disciples now inform contemporary philosophical debates (Block et al. 2014). The same pragmatic interdisciplinary approach may be useful in the task of creating an AI that makes moral judgments.

A possible strategy for continuing the project of creating an AI that can make moral judgments would be to identify the least demanding metaethical position and theory on the role that motivation plays in moral judgments. Philosophers might find this strategy frustrating. This is because metaethical positions that are easier to implement in an AI will never make up for them getting it

wrong about morality.[1] In general, one may be skeptical about the project of engineering an AI that, for all we know, may not be getting it right regarding moral properties. Nonetheless, we will allow ourselves some speculation about what the relatively easy to implement position may be. While a full defense of the view we end up with is outside the scope of this paper, some general remarks can already be made given what has been said so far.

First of all, it seems that the engineering demands of a fully realist project, which involves a moral microscope AI, is far beyond what we can presently achieve. The more plausible strategy involves simulating some aspect of our moral psychology. Which one, of course, depends on other considerations in metaethics.

Another difficult engineering problem comes from the discussion of the relationship between moral judgments and action. Take, for example, internalism, according to which there is a distinction to be made between agents who make genuine moral judgments and agents who appear to make moral judgments but do not because they lack any motivation to act on those judgments. A telling example of the distinction is that of psychopaths (or "amoralists"). While psychopaths can discriminate between right and wrong in a way that a normal person can, they lack any corresponding motivation to act based on those discriminations. Psychopaths do not care about morality in the same way that other people do (Roskies 2003). This is why the moral claims made by psychopaths are often characterized by internalists as moral judgments in only an inverted comma sense (Hare 1952; Prinz 2007).

If the internalist is right, then the engineering project of creating an AI that engages in moral judgments in any sense involves not only the challenge of replicating the ability or capacity to discriminate right from wrong, but also the arguably more difficult task of creating an internal mental economy that involves motivation to act on those discriminations. What compounds the already difficult situation in metaethics is the additional mystery of what motivation could be for an AI. On its own, an AI lacks the capacity to act, even though it can behave. These diffulties make the task of an internalist engineer an order-of-magnitude more difficult than the task that faces an externalist. An externalist denies that we need some motivational "oomph" to get genuine moral judgments. The externalist thinks that psychopaths make moral judgments—they just do not care about them in the same way that we do (Cima, Tonnaer, and Hauser 2010). From an engineering perspective the externalist position is easier to implement.

On the distinction between cognitivism and non-cognitivism the engineer should prefer the more elegant and less demanding cognitivism. Cognitivism, remember, is the view that moral judgments are types of thoughts and beliefs. So their content is propositional, in the same way in which the content of a thought or belief is. Propositional content can be expressed by tokens of sentences or symbols. Artificial intelligence techniques currently available, such as programming languages PROLOG or LISP, or cognitive architectures SOAR or ACT-R, largely depend on processing symbols and sentence tokens.

Non-cognitivism, on the other hand, presents the extra difficulty of simulating conative states, such as emotions or desires. While some of these states clearly have an aspect that involves contents other aspects of these states are qualitative. It is difficult to say what it would take to put qualitative character of fear or happiness into an AI.

We discussed three metaethical dimensions along which there is significant disagreement and made the further observation that depending on the assumptions and choices made in each of these dimensions we end up with distinct engineering challenges. These dimensions are: realism/antirealism, cognitivism/non-cognitivism, and externalism/internalism. The least demanding combination of views for the least demanding project of creating an AI that makes moral judgments is one that embraces antirealism, cognitivism, and externalism about moral motivation. Such a system issues moral judgments, which have a structure akin to sentences (cognitivism). This is irrespective of whether or not tokens of these sentences are hooking up with an external moral reality and whether such a reality even exists (antirealism). Because of the extraordinary demands of internalism, these moral judgments will also not be inherently motivating or hook up with action in a robust way (externalism).

Although it is not obvious what to conclude from this, it is a somewhat surprising combination of positions that is relatively rare in the metaethical literature. Error theory potentially comes close to be consistent with all of these positions, yet some of the classic statements of error theory include an element of internalism (Mackie 1977). A specific version of error theory, fictionalism, may come closest to fitting the description of the kind of minimal assumptions to be built into the AI (Joyce 2007). Fictionalism is the view that although we are systematically mistaken in the way we form moral judgments and use moral language, we should go on doing so for various practical reasons. This is what we should expect the first AI that "does ethics" to do as well.

References

Altham, J.E.J. 1986. The Legacy of Emotivism. in Macdonald & Wright, eds. *Fact, Science, and Morality*. Oxford University Press.

1 We thank an anonymous reviewer for bringing our attention to this issue.

Block, N., et al. 2014. Consciousness science: real progress and lingering misconceptions. *Trends in cognitive sciences* 18.11:556-557.

Brink, D. O. 1986. Externalist moral realism. *The Southern Journal of Philosophy* 24 (S1):23-41.

Brink, D. O. 1989. *Moral Realism and the Foundations of Ethics*. Cambridge: Cambridge University Press.

Cima, M., Franca T., and Hauser, M. D. 2010. Psychopaths know right from wrong but don't care. *Social cognitive and affective neuroscience* 5.1:59-67.

Dreier, J. 2002. Meta–Ethics and Normative Commitment. *Noûs* 36.s1: 241-263.

Dworkin, R. 1996. Objectivity and truth: You'd better believe it. *Philosophy & Public Affairs* 25.2:87-139.

Ehrenberg, K. M. 2008. Archimedean metaethics defended. *Metaphilosophy* 39.4-5:508-529.

Fantl, J. 2006. Is metaethics morally neutral?. *Pacific Philosophical Quarterly* 87.1:24-44.

Garrard, E., and McNaughton, D.1998. "Mapping moral motivation." *Ethical Theory and Moral Practice* 1.1:45-59.

Geach, Peter T. 1965. "Assertion," *Philosophical Review* 74: 449-465.

Gibbard, A. 1992. *Wise Choices, Apt Feelings: A theory of normative judgment*. Boston: Harvard University Press.

Hare, R.M. 1952. *The Language of Morals*. London:Oxford University Press, London.

Horgan, T., and Timmons. M. 2000. Nondescriptivist cognitivism: Framework for a new metaethic. *Philosophical Papers* 29.2:121-153.

Hume, D. 1975/1751. *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*, edited by L. A. Selby-Bigge, revised by P. H. Nidditch. Oxford: Oxford University Press.

Joyce, R. 2007. *The Myth of Morality*. Cabridge: Cambridge University Press.

Korsgaard, C. 1996. *The Sources of Normativity*. New York: Cambridge University Press.

Mackie, J. 1977. *Ethics: Inventing Right and Wrong*. London: Penguin.

Persson, I. and Savulescu, J. 2012. *Unfit for the future: The need for moral enhancement.* Oxford: Oxford University Press.

Prinz, J. 2007. *The Emotional Construction of Morals*. USA: Oxford University Press.

Rawls, J. 1975. *A Theory of Justice*. Belknap Press.

Ridge, M. 2006. Ecumenical Expressivism: Finessing Frege*. *Ethics* 116.2:302-336.

Roskies, A. 2003. Are ethical judgments intrinsically motivational? Lessons from acquired sociopathy. *Philosophical Psychology* 16.1:51-66.

Sturgeon, N.1985. Moral explanations. In *Morality, Reason, and Truth*, edited by D. Copp and D. Zimmerman. Totowa, New Jersey: Rowman & Littlefield:49-78.

Sturgeon, N. 1986. Harman on Moral Explanations of Natural Facts. *Southern Journal of Philosophy* 24 (Supplement):69–78.

Smith, M. 1995. *The Moral Problem.* Oxford: Blackwell.