Philosophical Writing

Lavanya Singh

October 12, 2021

Contents

1 Choice to Formalize the FUL

In *Groundwork of the Metaphysics of Morals*, Kant presents three formulations, or versions, of what he calls the "supreme law of morality." I will focus on the first of these three formulations, and below I explain the formulations and defend my choice.

Kant argues that if morality exists, it must take the form of a categorical imperative or a law that holds unconditionally. Categorical imperatives are contrasted with hypothetical imperatives, which take the form of conditionals as in, "If I want to get good grades, I must study hard." Hypothetical imperatives only have force so long as the antecedent holds, but the categorical imperative is unconditionally binding [Kant, 1785, 28]. In the first half of *Groundwork*, Kant examines what the categorical imperative, if such a thing exists and has force, must be. He concludes that there are three "formulations" of the categorical imperative, or three ways of articulating the supreme law of morality.

The first formulation of the categorical imperative is the formula of universal law (FUL), which reads, "act only according to that maxim through which you can at the same time will that it become a universal law." [Kant, 1785, 34] This formulation generates the universalizability test, which "tests" the moral value of a maxim by imagining a world in which it becomes a universal law and attempting to will the maxim in that world. The second formulation of the categorical imperative is the formula of humanity (FUH): "So act that you use humanity, in your own person, as well as in the person of any other, always at the same time as an end, never merely as a means." [Kant, 1785, 41]. This formulation is often understood as requiring us to acknowledge and respect the dignity of every other person. The third formulation of the categorical imperative is the formula of autonomy (FOA), which Korsgaard summarizes in her introduction to the Groundwork as, "we should so act that we may think of ourselves as legislating universal laws through our maxims." [Korsgaard, 2012, 28] While closely related to the FUL, the FOA presents morality as the activity of perfectly rational agents in an ideal "kingdom of ends," guided by what Kant calls the "laws of freedom."

I choose to focus on formalizations of Kant's first formulation of the categorical imperative, the formula of universal law (FUL), because it is the most formal and thus the easiest to formalize and implement. Onora O'Neill explains that the formalism of the FUL allows for greater precision in philosophical arguments analyzing its implications and power [O'Neill, 2013, 33]. This precision is particularly useful in a computational context because any formalism necessarily makes its content precise. The FUL's existing precision reduces ambiguity, allowing me to remain faithful to Kant's writing and philosophical interpretations of it. Precision reduces the need to make choices to resolve debates and ambiguities. Some of these choices may be well-studied and grounded in literature, but some may be unique to formalizing the FUL and thus understudied. Minimizing these choices minimizes arbitrariness in my formalization and puts it on solid philosophical footing. Given

that this thesis is a proof-of-concept, the formalism of the FUL is attractive because it reduces both the computational and philosophical complexity of my work.

While some criticize the FUL for its formalism and percieved "sterility" [O'Neill, 2013, 33], Kantian constructivists embrace it [Ebels-Duggan, 2012, 173]. My project is not committed to Kantian constructivism. I believe that computational ethics is likely a valuable tool for any ethicist, and I make the case for Kantian ethics specifically. Nonetheless, Kantian constructivists may find the focus on the FUL particularly appealing.

Though Kantians study all formulations of the categorical imperative, Kant argues in Groundwork that the three formulations of the categorical imperative are equivalent [Kant, 1785]. While this argument is disputed [Johnson and Cureton, 2021], for those who believe it, the stakes for my choice of the FUL are greatly reduced. If all formulations are equivalent, then a formalization of the FUL lends the exact same power as a formalization of the second or third formulation of the categorical imperative. In fact, future work could formalize the other formulas and try to prove that they are identical. Kant believes that his argument for the equality of the formulas is analytical, and if he is correct, it should be possible to recreate the argument in logic.

2 Definition of a Maxim

The central unit of evaluation for the universalizability test is a "maxim," which Kant defines in a footnote in *Groundworkd* as "the subjective principle of willing," or the principle that the agent acts on [Kant, 1785, 16]. Modern Kantians differ in their interpretations of this definition. The naive view is that a maxim is an act, but Korsgaard adopts the more sophisticated view that a maxim is composed of an act and the agent's purpose for acting [Korsgaard, 2005]. She also compares a maxim to Aristotle's logos, which includes these components and information about the circumstances and methods of the act. O'Neill concludes that Kant's examples imply that a maxim must also include circumstances [O'Neill, 2013], and Kitcher [Kitcher, 2003] uses textual evidence from the Groundwork to argue for the inclusion of a maxim's purpose or motivation. In order to formalize the notion of a maxim, I must adopt a specific definition and defend my choice.

I define a maxim as a circumstance, act, goal tuple (C, A, G), read as "In circumstances C, act A for goal G." Isabelle's strict typing rules mean that the choice of the type of each member of this tuple is significant. A circumstance is represented as a set of worlds t where that circumstance holds. A goal is also a term because it can be true or false at a world if it is realized or not. An act is an open sentence because an act itself is not the kind of thing that can be true or false (as in, an act is not truth-apt), but the combination of a subject performing an act can be true or false at a world depending on whether or not the act is indeed performed by that subject. For example, "running" is not truth-apt, but "Sara runs" is truth-apt.

My definition of a maxim is inspired by O'Neill's work on maxims. I will defend my representation below and consider an additional component that Kitcher argues for.

2.1 O'Niell's Original Schematic and The Role of Practical Judgement

O'Neill [O'Neill, 2013, 37] presents what Kitcher [Kitcher, 2003] calls the widely accepted view that a maxim is a circumstance, act, goal tuple. A maxim is an action-guiding rule and thus naturally includes an act and the circumstances under which it should be performed, which are often referred to as "morally relevant circumstances."

She also includes a purpose, end, or goal in the maxim because Kant includes this in many of his example maxims and because Kant argues that human activity, because it is guided by a rational will, is inherently purposive [Kant, 1785, 4 : 428]. A rational will does not act randomly (else it would not be rational), but instead in the pursuit of ends which it deems valuable. This inclusion is also essential for the version of the universalizability test that I will implement, explained in Section ??.

O'Neill's inclusion of circumstances is potentially controversial because it leaves open the question of what qualifies as a relevant circumstance for a particular maxim. This is gives rise to "the tailoring objection" [Kitcher, 2003, 217][1], under which maxims are arbitrarily specified to pass the FUL. For example, the maxim "When my name is Lavanya Singh, I will lie to get some easy money," is universalizable, but is clearly a false positive. One solution to this problem is to argue that the circumstance "When my name is Lavanya Singh" is not morally relevant to the act and goal. This solution requires some discussion of what qualifies as a relevant circumstance.

O'Neill seems to acknowledge the difficulty of determining relevant circumstances when she concedes that a maxim cannot include all of the infinitely many circumstances in which the agent may perform the action[O'Neill, 2013, 4 : 428]. She argues that this is an artifact of the fact that maxims are rules of practical reason, the kind of reason that helps us decide what to do and how to do it [Bok, 1998]. Like any practical rule, maxims require the exercise of practical judgement to determine in which circumstances they should be applied. This judgement, applied in both choosing when to exercise the maxim and in the formulation of the maxim itself, is what determines the "morally relevant circumstances."

The upshot for computational ethics is that the computer cannot perform all ethical activity alone. Human judgement and the exercise of practical reason are essential to both formulate maxims and determine when the actual conditions of life coincide with the circumstances in which the maxim is relevant. Choosing when to exercise a maxim is less relevant to my project because analyzing a formal representation

[1] Kitcher cites [Wood, 1999] as offering an example of a false positive due to this objection.

of the FUL requires making the circumstances in a given scenario precise, but will be important for applications of computational ethics to guiding AI agents. The difficulty in formulating a maxim, on the other hand, demonstrates the important fact that ethics, as presented here, is not a solely computational activity. A human being must create a representation for the dilemma they wish to test, effectively translating a complex, real situation into a flat logical structure. This parallels the challenge that programmers face when translating the complexity of reality to a programming langauge or computational representation. Not only will some of the situation's complexity inevitably be lost, the outcome of the universalizability test will depend on how the human formulates the maxim and whether or not this formulation does indeed include morally relevant circumstances. If the human puts garbage into the test, the test will return garbage out.

While this may appear to be a weakness of my system, I believe that it actually allows my system to retain some of the human complexity that many philosophers agree cannot be automated away.[2] Ethics is a fundamentally human activity. Kant argues that the categorical imperative is a statement about the properties of rational wills. In fact, Korsgaard argues that morality derives its authority over us, or normativity, only because is it a property of a rational will, and we, as human beings, are rational wills. If ethics is meant to guide human behavior, the role of the computer becomes clear as not a replacement for our will, but instead as a tool to help guide our wills and reason more efficiently and more effectively. Just as calculators don't render mathematicians obsolete, computational ethics does not render human judgement or philosophy obsolete. Chapter 4 Section ?? will be devoted to a more complete discussion of this issue.

2.2 Exclusion of Motive

Kitcher begins with O'Niell's circumstance, act, goal view and expands it to include the motive behind performing the maxim [Kitcher, 2003]. This additional component is read as "In circumstance C, I will do A in order to G because of M," where M may be "duty" or "self-love." Kitcher argues that the inclusion of motive is necessary for the fullest, most general form of a maxim in order to capture Kant's idea that an action derives its moral worth from being done for the sake of duty itself. Under this view, the FUL would obligate maxims of the form "In circumstance C, I will do A in order to G because I can will that I and everyone else simultaneously will do A in order to G in circumstance C." In other words, if Kant is correct in arguing that moral actions must be done from the motive of duty, the affirmative result of the FUL becomes the motive for a moral action.

While Kitcher's conception of a maxim captures Kant's idea of acting for duty's own sake, I will not implement it because it is not necessary for putting maxims through the FUL. Indeed, Kitcher acknowledges that O'Neill's formulation suf-

[2] Powers presents the determination of morally relevant circumstances as an obstacle to the automation of Kantian ethics [Powers, 2006].

fices for the universalizability test, but is not the general notion of a maxim. In order to pass the maxim through the FUL, it suffices to know the circumstance, act, and goal. The FUL derives the motive that Kitcher bundles into the maxim, so automating the FUL does not require including a motive. The "input" to the FUL is the circumstance, act, goal tuple. My project takes this input and returns the motivation that the dutiful, moral agent would adopt. Additionally, doing justice to the rich notion of motive requires modelling the operation of practical reason itself, which is outside the scope of this project. My work focuses on the universalizability test, but future work that models the process of practical reason may use my implementation of the FUL as a "library." Combined with a logic of practical reason, an implementation of the FUL can move from evaluating a maxim to evaluating an agent's behavior, since that's when "acting from duty" starts to matter.

3 Practical Contradiction Interpretation

Kantians debate the correct interpretation of the formula of universal law because Kant appears to interpret the universalizability test in different ways. My project uses Korsgaard's practical contradiction interpretation, broadly accepted as correct within the philosophical community [Ebels-Duggan, 2012, 177]. Below, I briefly reconstruct Korsgaard's argument for the practical contradiction interpretation. While she believes that the text partially supports this interpretation, her argument is philosophical and derives its strength from the plausibility of the practical contradiction interpretation.

Recall that the formula of universal law is "act only in accordance with that maxim through which you can at the same time will that it become a universal law" [Kant, 1785, 4 : 421]. To determine if a maxim can be willed as a universal law, one must use the "universalizability test," which requires imagining a world in which everyone for all of time has willed the maxim. If willing the maxim in such a world generates a contradiction, then the action is prohibited. There are three interpretations of what sort of contradiction is necessary: (1) the teleological view, prohibiting actions that conflict with some assumed teleological end when universalized, (2) the logical contradiction view, prohibiting maxims that are logically impossible when universalized, and (3) the practical contradiction view, prohibiting maxims that are self-defeating when universalized.

Under the logical contradiction interpretation, falsely promising to repay a loan to get some quick cash fails the universalizability test because, in such a world, the practice of promising would die out so making a false promise would be impossible. Korsgaard appeals to Dietrichson [Dietrichson, 1964] to construct the example of a mother killing her children that tend to cry more than average so that she can get some sleep at night. Universalizing this maxim does not generate a logical contradiction, but it is clearly morally wrong. The problem here is that killing is a natural action, which Korsgaard distinguishes from a practice, like promising.

Natural actions will never be logically impossible, so the logical contradiction view fails to prohibit them.

Under the teleological contradiction interpretation, a maxim is prohibited if it undercuts some natural or assigned purpose for some practice, act, or object. For example, the purpose of promising is to create a system of mutual trust and false promising undercuts this purpose and is thus prohibited. The problem with this view is that it assumes that the agent is committed, either because of their own goals or because of some property of a rational will, to some teleological system. Acton formulates Hegel's argument that [Ewing, 1972], an agent doesn't have to be committed to promising as a system of mutual trust. Korsgaard concludes that assigning teleological purposes to actions is difficult because "such purposes may have nothing to do with what the agent wants or ought rationally to want, or even with what any human being wants." If the agent is not committed to the purpose, then will not see a contradiction in willing an act that violates this purpose.

This difficulty with the teleological contradiction interpretation drives Korsgaard to look for purposes that an agent must necessarily be committed to, and she concludes that this must be the purpose of the maxim itself. By willing a maxim, an agent commits themselves to the goal of the maxim, and thus cannot rationally will a system in which this goal is undercut. This system satisfactorily handles natural actions like those of the sleep-deprived mother: in willing the end of sleeping through the night, she is implicitly willing that she be alive in order to secure and enjoy her sleep. If any mother is allowed to kill any loud child, then she cannot be secure in the possession of her life, because her own mother may have grown frustrated with her crying. Her willing this maxim thwarts the end that she sought to secure.

The practical contradiction interpretation not only addresses the problems with the first two interpretations, it also offers a much more satisfying explanation of why certain maxims are immoral. The problem is not the existence of a contradiction itself, but instead the fact that these maxims involve parasitic behavior on social conditions that the agent seeks to benefit from. The false promiser simultaneously wants to abuse the system of promising and benefit from it, and is thus making an exception of themselves. It is this kind of free-riding that the universalizability test seeks to draw out. The test raises the same kinds of objections that the question "What if everyone did that?" seeks to draw out.

4 Philosophical Contributions

I argue that computational ethics should be useful for and interesting to philosophers for two reasons. First, it serves as the basis for AI agents with the capacity for philosophically sophisticated ethical reasoning. For example, my project contributes an implementation of the Formula of Universal Law that an AI agent could use to reason about the world using the categorical imperative. Second, computational ethics helps philosophers think about ethics in the same way that calculators or theorem provers help mathematicians think about math. I am not arguing that the computer can replace human reasoning or prove things that humans theoretically couldn't do. Instead, I argue that the computer bolsters human reasoning, firstly by forcing precision due to the rigid syntax of a computer program and secondly by making formal proofs less tedious. Below, I explore each of these contributions in greater detail.

4.1 AI Agents

As artifical intelligence becomes more powerful, science fiction looms closer to reality and the danger of "unethical" AI becomes more urgent. Ethically intelligent artificial agents will need to be able to reason about sophisticated ethical theories in order to navigate the world. My project could serve as one component of a complete ethical reasoner. Specifically, my project could easily be repurposed into a library or "engine" that takes as input the logical representation of a maxim and determines if it is obligatory, prohibited, or permissible.

As it stands, my project can evaluate the moral status of maxims represented in my logic and potentially serves as one component of an "ethics engine," which an AI agent could use to make ethical decisions. For example, my system could be combined with an input parser to translate moral dilemmas or situations into maxims in my logic and an output to translate the moral status output of my system into a prescription for what action should be taken. The diagram below depicts the workflow of this example of an ethics engine.

In this workflow, an AI agent is faced with a moral dilemma in some internal representation and passes this representation to a parser, which converts it to the appropriate logical representation for my system to evaluate. For example, if an AI agent represents the moral dilemma in natural language, the parser would convert the natural language representation to a representation in my logic.

Once the input is a sentence in my logic, my project can evaluate its moral status using my implementation of the FUL. Concretely, my project would return a value indicating if the maxim is obligatory, permissible, or prohibited. The maxim would be prohibited if it fails the universalizability test, permissible if it passes, and obligatory if its negation fails the universalizability test. All three of these properties amount to testing if a certain theorem holds or not in my logic, a calculation that I demonstrate repeatedly in my tests.

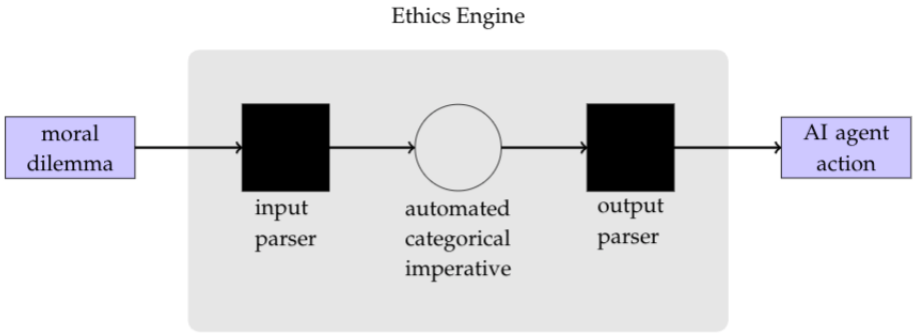

Figure 1: An example of an ethics engine for an artificial agent

This output could then be converted into some actionable, useful response for the AI agent with another output parser. For example, if the AI agent is equipped to evaluate natural language prescriptions, the status of the maxim could be parsed into a natural language sentence. This output will be passed back to the AI agent, which will make a decision using it.

The ethics engine depicted above is merely one way to use my project to guide an artifical agent. The upshot is that an automated version of the categorical imperative could function as the ethical engine for an AI agent, with some work to parse the input and the output. Effectively, some version of the kind of automated ethics I implement could be exposed as an "ethics API" that AI developers could use to easily give AI agent the capacity for sophisticated ethical reasoning faithful to philosophical literature. This represents an improvement over existing ethics engines, which, as I examine in Section ??, rarely attempt to capture the complexity of any ethical theory that philosophers plausibly defend.

4.2 Computational Philosophy

4.2.1 Example of a Philosophical Insight

As I tested prior formulations of the categorical imperative and implemented my own, the process of implementing a formalization and testing it using an interactive theorem prover resulted in philosophical insights that were novel to me. The process resulted in surprising logical results that provoked interesting philosophical conversations as I tried to understand their implications for the ethical theory I am formalizing.

For example, as I was implementing my formalization of the FUL, I realized that my formalization was inconsistent unless I specified that the FUL only held for

"well-formed maxims," such that neither the act nor goal were already achieved in the given circumstances. Below I document how I came to this conclusion. First, I used Sledgehammer to show that my formalization of the FUL[3] resulted in a contradiction. Sledgehammer was able to tell me which axioms it used to complete this proof, showing me that my formalization contradicted the axiom O_diamond, which states that an obligated term cannot contradict its context. I hypothesized that there was some tension between the antecedent of the FUL, which states that a maxim is acted on by all agents, and the consequent, which states that a maxim is prohibited. If the maxim has already been acted on, then not acting on it becomes impossible so the prohibition is impossible to obey!

To solve this problem, I returned to Korsgaard's practical contradiction interpretation and focused on the imaginatory component of the FUL. Specifically, the universalizability test requires that we IMAGINE a world where the maxim is universalized, not that the maxim is actually universalized at this world. I implemented another version of the FUL under which, if a maxim is universalized at any world and rendered ineffective at that world, it is prohibited at the current world. I hypothesized that this would remove the contradiction found above. However, Nitpick was still timing out when looking for a model that satisfied this new version of the FUL. Nitpick is a model checker that generates models that satisfy some axioms and theorems, and usually it is able to find small satisfying models in a matter of seconds. The fact that Nitpick was timing out was a serious indication that my formalization was still not consistent.

I suspected that Nitpick could possibly be timing out due to checking large models that exhausted its time limit. To avoid this, I decided to help the system out and specify the exact number of maxims in the system by passing as an argument to Nitpick the cardinality of my desired model. This did not fix the problem. I next defined a particular (circumstance, act, goal) tuple as a constant and, instead of stating that the FUL held for all maxims, I stated that the FUL held for the specific maxim formed by this tuple. To my surprise, Nitpick was now able to show that the FUL was consistent!

This initially appeared counterintuitive—after all, what is the difference between a model of cardinality 1 and a model with one constant object? Professor Amin pointed out that as constants, the circumstances, act, and goal for which the FUL held were all distinct, but when quantified over they were not. One line of code later, I tested this hypothesis and found that, indeed, specifying that the circumstances could not entail the act or goal fixed the inconsistency! This logical inconsistency showed me that the FUL could not hold for maxims in which the act or goal have already been achieved in the circumstances.

To understand this property as a philoosphical insight, I returned to Korsgaard and O'Neill's interpretations of maxims as practical, action-guiding principles, and concluding that a maxim in which the circumstances already entail the act or goal

[3]The full logical representation is $\mathit{FUL0} \equiv \forall c\ a\ g\ s.\ \mathit{not\text{-}universalizable}\ (c,\ a,\ g)\ s \longrightarrow \models \mathit{prohibited}\ (c, a, g)\ s$.

is vacuous and not a useful practical principle to evaluate. This is a non-trivial philosophical insight that the computer helped me discover!

4.2.2 Two Uses of Computational Ethics

Forcing precision; making proofs fast

4.2.3 Looking Forward

My project does not demonstrate that computational ethics can discover new philosophical insights that humans are incapable of reaching. After all, my understanding of a well-formed maxim could very well exist in the literature and certainly could be reached by a philosopher working without any computational tools. Instead, the upshot is that working with a computer, today, is another way for philosophers to arrive at philosophical insights. Just as theorem provers do not replace mathematicians entirely, computational ethics does not outdo the philosopher but instead helps them do their work faster. The computer offers a different perspective and prompts new questions that lead to insights.

Moreover, computational ethics is at its infancy. The use of theorem provers in mathematics is just now beginning to make headway (cite Kevin Buzzard), even though theorem provers were first invented in the 1960's (cite https://www.cl.cam.ac.uk/ jrh13/papers/joerg.pdf In contrast, the first attempts to use theorem provers for ethics occurred in the last decade. The fact that this nascent technology is already helping humans reach non-trivial philosophical conclusions is reason to, at the very least, entertain the possibility of a future where computational ethics becomes as normal as using a calculator for arithmetic.

To the skeptic, the ethical insights uncovered by the computer are not necessarily trendy or impressive philosophy. Indeed, the fact that a theorem prover requires specialized knowledge outside of the field of philosopher indicates that the technology is nowhere near ready for universal use in philosophy departments. However, history indicates that as computing power increases and computer scientists make progress, computational ethics will become more usable. Theorem provers in mathematics began as toys incapable of proving that the real number 2 is not equal to the real number 1, but Buzzard showed that moving from such a primitive system to a tool for Fields medal winning mathematics is possible in a matter of years (cite Buzzard). Countless examples from the history of computer science, from the Turing Test to AI game playing to protein folding, demonstrate that progress in computer science can make seemingly obscure computer programs useful and usable in ways that exceed our wildest imaginations. Indeed, programmable computers themselves initially began as unwieldy punch card readers, but their current ubiquity need not be stated. If computer scientists and philosophers invest in computational ethics, it can become as much a tool for philosophy as a calculator is for

for arithmetic.[4]

References

[Bok, 1998] Bok, H. (1998). *Freedom and Responsibility*. Princeton University Press.

[Dietrichson, 1964] Dietrichson, P. (1964). When is a maxim fully universalizable ? 55(1-4):143–170.

[Ebels-Duggan, 2012] Ebels-Duggan, K. (2012). *Kantian Ethics*, chapter Kantian Ethics. Continuum.

[Ewing, 1972] Ewing, A. C. (1972). *Philosophy*, 47(180):173–175.

[Johnson and Cureton, 2021] Johnson, R. and Cureton, A. (2021). Kant's Moral Philosophy. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2021 edition.

[Kant, 1785] Kant, I. (1785). *Groundwork of the Metaphysics of Morals*. Cambridge University Press, Cambridge.

[Kitcher, 2003] Kitcher, P. (2003). What is a maxim? *Philosophical Topics*, 31(1/2):215–243.

[Korsgaard, 2012] Korsgaard, C. (2012). *Groundwork of the Metaphysics of Morals*, chapter Introduction. Cambridge University Press, Cambridge.

[Korsgaard, 2005] Korsgaard, C. M. (2005). Acting for a reason. *Danish Yearbook of Philosophy*, 40(1):11–35.

[O'Neill, 2013] O'Neill, O. (2013). *Acting on Principle: An Essay on Kantian Ethics*. Cambridge University Press.

[Powers, 2006] Powers, T. M. (2006). Prospects for a kantian machine. *IEEE Intelligent Systems*, 21(4):46–51.

[Wood, 1999] Wood, A. W. (1999). *Kant's Ethical Thought*. Cambridge University Press.

[4] Is this too like, lalalala fantasy of computational philosophy?