Paper 3

25306749
Philosophy 104
UC Berkeley

Fall 2017

> There cannot be any very intersting, tidy or self-contained theory of what morality is ... [This] has never succeeded, and could not succeed, in answering the question, *by what right* does it legislate to the moral sentiments?
>
> Bernard Williams[Wil81]

Abstract

A central theme in moral philosophy is the relation between rightness and goodness.[1] Consequentialism, contractualism, and natural goodness theory may be seen as different ways of establishing the relation between rightness and goodness. Many moral philosophers agree that the three approaches are irreconcilable. I argue that this is a mistake. In what follows, I show how the three approaches are intertwined to yield a self-contained theory of what morality is, which has the right to legislate to the moral sentiments.

0 Preliminaries

In 3,000 words I squeeze in a unification of three unruly major ethical theories. Necessarily I value economy of style, at cost of clarity and sometimes modesty. To combat this, I make liberal use of footnotes, plus an appendix.[2] I do not consider the footnotes or appendix as part of the word count, for most are mathematical, not philosophical, in nature, and they should be omitted during grading. This paper proceeds as follows. First, I briefly overview consequentialism, natural goodness theory, and contractualism. Second, I describe a partial reconciliation through Korsgaard's approach. Finally, I formulate a complete reconciliation.

[1] As John Rawls says, "The two main concepts of ethics are those of the right and the good. The structure of an ethical theory is largely determined by how it defines and connectes these two basic notions". [Raw71]. We may define the terms, as Korsgaard (roughly) puts it: "'Good' names the problem of what we are to strive for, aim at, and care about in our lives. 'Right' names the more specific problem of which actions we may perform". [KO96, p. 114] Korsgaard is quick to point out that the relation between good and right is more subtle than that, and in fact "difficult to articulate".

[2] It is a sort of division of labor: for a mathematically inclined reader, this explanation will not be needed, and for a reader less inclined as such, the explanations are there to provide psychological assurance just in case he finds my assertions suspicious.

1 View Overview

1.1 Consequentialism

In consequentialism, the right action is the action that leads to the best states of affairs. To borrow Williams's words,

> In any form of direct consequentialism ... the right action is that which out of the actions available to the agent brings about or represents the highest degree of whatever it is the system in question regards as intrinsically valuable. [Wil88, p. 23]

There are two issues with consequentialism: (1) what, exactly, will the system regard as intrinsically valuable?, and (2) how will the system calculate the values?

Issue (1) fractures consequentialism into different kinds. But this is a relatively mild issue, because it keeps the basic structure of consequentialism intact, and disagreements simply swap in something else as the intrinsically valuable thing.

Issue (2) is more severe, as it questions consequentialism's basic structure. As Sidgwick lays out in *The Methods of Ethics*, the calculation of values is done from "the point of view of the universe". [Sid11] That is, the evaluation is done impartially. For many, this is the most unsettling feature of consequentialism. As Railton writes, it leads to scenarios like the following, which cause "alienation":

> When a friend remarks upon the extraordinary quality of John's concern for his wife, John responds ... "I've always thought that people should help each other when they're in a specially good position to do so. I know Anne better than anyone else does ... I get a lot of satisfaction out of it." [Rai84, p. 135]

Something is strange: John impartially justifies his love, but it seems that John should love his wife, not for any other reason, but "in some ultimate sense, in part of *her* sake"[Rai84, p. 136]. Furthermore, the consequentialist, in his incessant calculation of values, is plauged with the sentiment in that famous phrase, "one thought too many". [Wil81, p. 18] And because consequentialism obligates the agent to always perform what would lead to the best state of affairs, it takes away his free will.

There is one more problem with issue (2). When the consequentialist needs to calculate the value of a state, how, exactly, will he *do the calculation*? Mental arithmetic? Using a pocket calculator? Under reasonable assumptions, it is an inconvenient mathematical fact that this calculation may take billions of years, or be logically impossible, even with supercomputers.[3] I will return to this idea.

Despite its faults, consequentialism remains a dominant idea in moral philosophy. Scheffler writes that this is because of its "deeply plausible-sounding feature that one may always do what would lead to the best available outcome overall." [Sch82]

1.2 Natural Goodness Theory

Natural goodness theory holds natural normativity assertions, declares individuals good or bad according to these assertions, and defines actions performed by good individuals as right actions.

[3] Consider the problem of choosing the action that leads to the greatest happiness of the greatest number. It is reasonable that an action may make some happy, and others not happy. Now let us suppose the world has just three hundred people, and we want to make as many of them happy as possible. A person p_i may have some relation with another person p_j such that p_i and p_j can either both be happy or both be miserable; think star-crossed lovers. On the other hand, a person p_m may have some relation with another preson p_k such that p_m's happiness has no bearing on p_k. This can be formulated as a Maximum Boolean Satisfiability Problem, [Sun06] a formal mathematical problem which, with the aforementioned three hundred variables, requires time proportional to 2^{300}. Even if each atom in the universe did one computation per nanosecond, the universe would have to start over several times before we can find a solution, or reasonable approximation, to this problem.

Foot's main thesis is that human good can be evaluated in the same grammar under which animal or plant good is evaluated, that is, under a natural normativity assertion. When we say that a rabbit, for instance, is good, we mean that it accords with the natural norms of the rabbit species. When we say a human is good, we mean something similar:

> There is no change in the meaning of 'good' between the word as it appears in 'good roots' and as it appears in 'good dispositions of the human will'. [Foo01, p. 39]

What this theory ultimately yields is a sort of Aristotelian virtue ethics: as Foot endorses it, "men need virtues as bees need stings." [Foo01, p. 35] Therefore virtues like promise-keeping, diligence, loyalty, and so on are valued under Foot's system.

One problem in Foot's theory is that a "natural normativity assertion" is ill-defined. Foot asserts that they are not quantifiable[Foo01, p. 28], not statistical[Foo01, p. 33], and not about evolution[Foo01, p. 32], but she never provides an exact positive definition. So it is difficult to argue whether such assertions are true or not. This does not mean they are meaningless, for there is a clear sense in the natural normativity assertion "humans keep promises." Still, the ill-defined nature of such assertions is a weakness in Foot's theory, for it makes the theory unfalsifiable.

1.3 Contractualism

Contractualism says, in Scanlon's words, "judgments of right and wrong ... are judgments about what would be permitted by principles that could not reasonbly be rejected. [Sca98, p. 4]."

This idea has "an obvious similarity to Kant's Categorical Imperative." [Sca98, p. 5] However, there are significant differences also. The most glaring is that for Kant, the authority of moral judgment is grounded on rational agency, whereas for Scanlon, it is grounded on "justfiability to others" [Sca98, p. 5]

Therefore, in contractualism, reason is primal: that we have reason to do something counts as the bedrock justification for doing it. Scanlon is aware that this idea runs counter to Hume's age-long worry that "reason is, and ought only to be the slave of the passions." [HM85, p. 415] Scanlon therefore explains away this worry. But his explanation is opaque, concealing at least one leap of logic. It is worth examining Scanlon's argument in detail, with the help of formal language, to lay bare his mistake.

To start things off, Scanlon writes that "judgment-sensitive attitudes" are that "class of things *for which reasons in the standard normative sense can sensibly be asked for or offered* [italics mine]." [Sca98, p. 21] For convenience, I will call this italicized requirement, requirement R, and call judgment-sensitive attitudes, J. Then we can say:

$$\forall x \in Things : J = \{x | R(x)\}$$

Scanlon writes that attitudes, such as belief, belong to this class. Then he makes a more daring move: he wants to bring actions into this class, writing "actions are the kind of things for which normative reasons can be given only insofar as they are intentional, that is, are the expression of judgment-sensitive attitudes." [Sca98, p. 21] Formalizing this, we have

$$Actions \subset Things \tag{1}$$
$$a_i \in Actions \tag{2}$$
$$j_i \in J \tag{3}$$
$$intentional(a_i) = expression(j_i) \tag{4}$$
$$intentional(a_i) \implies R(a_i) \tag{5}$$
$$intentional(a_i) \implies a_i \in J \tag{6}$$

(1) formalizes "Actions are [a] kind of things."[Sca98, p. 21] (2) just says, "let a_i be some action." Similarly, (3) says, "let j_i be some judgement-sensitive attitude." (4) formalizes Scanlon's definition of "intentional" as "expression of a judgment-sensitive attitude", which is implicit when

he writes that some actions "are intentional, that is, are the expression of judgment-sensitive attitudes." (5) formalizes "actions are the kind of things *for which normative reasons can be given* only insofar as they are intentional", where the italicized requirement, again, is requirement R. (6) follows from the definition of judgment-sensitive attitudes: any *thing* that satisfies requirement R is a judgment-sensitive attitude, an a_i is a *thing* because actions are things.

From this exposition, it is clear that what is missing is the definition of *expression*. If it were plausible that an *expression of an attitude* belongs to the same class as an *attitude*, then Scanlon's argument would work.

But I do not believe this is plausible. In theoretical computer science, there is a distinction between a description of an algorithm and an execution of an algorithm.[4] Roughly, it is the distinction between a sentence with quotes, and a sentence without quotes. A sentence with quotes is *referred to*; a sentence without quotes is *meant*. If we can say that an attitude is *referred to*, while an expression of an attitude is *meant*, then the distinction between an attitude, and the expression of an attitude, is isomorphic to the distinction between an algorithm, and the execution of the algorithm. This is not a trivial distinction, for it is at the core of the most infamous idea in theoretical computer science: the unsolvability of the *Entscheidungsproblem*.[5] So I have argued that Scanlon made a mistake.

In practical terms, Scanlon's view may lead to morally uneasy scenarios. Consider P, a person living under a brutal regime that routinely massacres innocents. It seems that the proper moral philosophy would obligate P to rebel against the regime. But Scanlon's ethical theory would not obligate P to do that. For Scanlon, it is enough for P to have an *attitude* of rebellion. P is not obligated to *act on* that attitude, because acts and attitudes are in the same class anyway. Later, when the regime has been deposed of, and someone asks P, *why did you not resist against that regime?*, it is enough, according to Scanlon, for P to say *well, I had the attitude of resistance.* But of course that is not enough.

2 A Partial Reconciliation

I now start the synthesis. Immediately the critic might object: the difference between the three systems runs very deep. How could they ever be reconciled? Just listen to Foot:

> [U]tilitarianism, like any other form of consequentialism, has as its foundation a proposition linking goodness of action in one way or another to the goodness of *states of affairs. And there is no room for such a foundational proposition in the theory of natural normativity.* [Foo01, p. 49]

Or just listen to Scanlon, responding to Moore's claim that friendship is valuable because it makes the universe better:

> It seems overblown to say that what is important about friendship is that it increases the value of the state of the universe in which it occurs. [Sca98, p. 88]

So what gives? As a first order approximation, I believe Korsgaard has done much of the synthesis already.

[4] A description of an algorithm is just some information. An execution of an algorithm is the actual carrying out of orders which the description of the algorithm specifies.

[5] To briefly explain, the *Entscheidungsproblem* is an unsolvable, or *uncomputable*, mathematical problem. The problem is to find an algorithm that takes as input some arbitrary algorithm and outputs whether that inputted algorithm ever stops running, or not. Suppose this algorithm exists; let's call it A, such that $A(X) = 1$ if X halts and $A(X) = 0$ if X runs forever. Then we can devise F, another algorithm, that takes in an algorithm, let's call it I, such that if $A(I) = 1$ (i.e. if I halts), F runs forever, and if $A(I) = 0$ (i.e. if I runs forever), F halts. Now does $F(F)$ halt or does it run forever? Both (or neither), which is a contradiction, so A can't exist. The core knot, the reason, is the self-referential nature of the suspicious-looking function call $F(F)$. At a deeper level, the core knot is what makes self-referential function calls possible at all. When we execute $F(F)$, notice there are two F's. The first F is *executed*; the second F is *referred to*. It is as if the first F were not in quotes and the second F were in quotes. See [Tur37] for a further exposition.

2.1 Reconciliation through Korsgaard's Approach

For Korsgaard, the right action is the action that follows from a conception of a practical identity, which in turn springs from valuing humanity as an intrinsic good. So the right action, for Korsgaard, is the action that values humanity: we may say, the action that makes possible valuable relations of mutual recognition. So her view may subsume contractualism. This is the easiest relation to make, no doubt because both owe a great debt to Kant and the categorical imperative.

But for Korsgaard, the right action is the action performed by an autonomous, rational agent, which is a good thing to be. So the right action is the action that would be performed by the person who is good as a human being. To see how Korsgaard's method may relate to the natural goodness theory, we can formulate it in natural goodness terms. We start with a natural normativity assertion, "humans are autonomous". So an autonomous individual is a good individual, and an autonomous individual makes free actions, which are right actions.

But then again for Korsgaard, there is at least a sense in which we can say the right action is the action that produces the best state of affairs. She writes,

> A good maxim is good in virtue of its internal structure ... [which] makes it fit to be willled as a law. A good maxim is therefore an intrinsically normative entity. [KO96, p. 87]

So a good maxim is an *entity*, and furthermore a *good* entity. It is plausible to suggest that a state of affairs is composed of entities. So we can interpret Korsgaard as saying that the right action is the action that leads to a state of affairs with good maxims. Or can we?

2.2 Limitations of Korsgaard's Approach

A synthesis of Korsgaard with consequentialism seems like an especially fanciful stretch of the imagination. It seems implausible that a maxim is an element of a state of affairs. In particular, are maxims really things whose values can be aggregated by simple summation? I would say, roughly, that not a good maxim itself, but the carrying out of a good maxim, yields some value.[6] But how exactly?

So Korsgaard's approach is not enough. Spiritually, the main difference between my approach and Korsgaard's lies in their attitudes towards the Modern Scientific World View. Korsgaard considers the MSWV and its claims on modern moral philosophy as a sort of disease that has no real merit. With contempt she describes it as "supposed to be somehow inimical to ethics"[KO96, p. 18], that it is the "philosopher's bugbear"[KO96, p. 94], that it led Mackie to the clearly wrong conclusion that there are no normative entities. [KO96, p. 108] She thinks that the MSWV's scary claims on morality lead many a moral philosopher to, so to speak, bury their heads in the sand. Her move is to pull out her head. My move is to bury my head as far as it goes, so that it digs a looping tunnel through the sand, comes out the other side, and looks at the world upside down. Instead of dismissing the MSWV, I take it to its limits: my metaphysical position is an outrageous flavor of computationalism, that mind and thought can be described in computational terms, indeed that everything can be described in computational terms.[7]

3 The Full Reconciliation

We start with the ethical theory with an existing mathematicization: consequentialism. Consequentialism has proven to be very germane to mathematicization, so much so that almost all arti-

[6] In exactly the same manner as I argued in section 1.3 that an attitude and an expression of an attitude belong in different classes, so I argue a maxim and the execution of a maxim belong in different classes.

[7] Putnam and Searle's famous arguments against the theory, I believe, have been successfully debunked. For one such debunking, see [MD17].

ficially intelligent systems today operate under some formalization of consequentialism. [RN09, p. 483]

Formally, a consequentialist value system of some agent g is a 3-tuple $\{v, S, A\}$ such that

$v : S \to \mathbb{Q}$ is a *value function*;
$S = \{s_1, s_2, ..., s_n\}$ is a set of *states of affairs* (just *states* from here on);and
$A = \{a_1, a_2, ..., a_n\}$ is a set of *actions*.

The value function v "evaluates" a state, taking as input a state and yielding a number as how "good" that state is. S is the set of states. A defines a set of actions available to the agent g.

As I explained in section 1.1, there are two issues with consequentialism. The first issue is the choice of the value function. The second, and more substantial, issue is consequentialism's insistence on evaluating moral decisions under the "point of view of the universe."

There are in turn two worries with the second issue: (1) that it gives a positive decision procedure for moral evaluation and thus takes away free will, and (2) that it causes alienation. But what if there were a special kind of value function that sidesteps both worries? As I hinted at in section 1.1, we may define a value function that is meaningful, yet whose calculation is logically impossible.

Because of this impossibility, such a value function would not yield any positive decision procedure, and it would leave plenty of room for free will, negating worry (1).

At the same time, this value function would avoid alienation. When one looks closely at the structure of alienation, what is repulsive is the consequentialist's willingness to bend over backwards to get to the best state of affairs. This is a consequence of the consequentialist's incessant calculation (the "one thought too many") of the values of states. But if our value function is incalculable, the consequentialist no longer has reason to engage in such calculations. Thus worry (2) is negated.

One such value function is Kolmogorov complexity, which measures the amount of uncomputability in a state. We write:

$$\forall s \in S : v(s) = K(s)$$

where K is Kolmogorov complexity. K is an uncomputable, or incalculable, function, therefore it negates both worries (1) and (2). It also addresses the first issue, because it is a very reasonable choice as a value function: uncomputability can be naturally interpreted as the result of autonomy, in Korsgaard's sense.

Here is why. Korsgaard writes that humans are reflective agents: "Our capacity to turn our attention on to our own mental activities is also a capacity to distance ourselves from them, and to call them into question."[KO96, p. 93] This leads to reflective endorsement and eventually autonomy. Under a computationalist lens, this may be interpreted as: humans, confronted with a program, do not necessarily execute the program; instead, humans can see a program (mental activity) *as a program* (*mental activity from a distance*), which gives them the ability to be uncomputable. [8] Therefore, to say a human is autonomous is to say that he is executing an

[8]For example: a cat, upon seeing a rat, is "programmed" to chase after the rat. We can view the rat as invoking a program that is inside the cat, and the cat, being non-reflective, cannot deliberate on this program, so it simply executes the program, i.e. gives chase. A human, on the other hand, can, upon seeing a doughnut, "see" the "program" inside her or him, which program's execution *is* her or him chasing the doughnut. What I am saying is roughly that some grounds for action is a program, and acting on those grounds is an execution of the program. Korsgaard writes that human reflection as such, and reflective endorsement, yields a rationally free and autonomous human being. The process by which a human becomes free has an isomorphic structure with the process by which the *Entscheidungsproblem*[Tur37] becomes uncomputable. That problem is uncomputable because a program can take as input some information that describes another program. Similarly, when a human sees a mental activity *as a program*, that is, the information that describes the program, *so* the human need not necessarily execute the program, may be *free* from that mental activity.

uncomputable program.[9] It is to say that he is executing an *arbitrary* program.[10]

Each time a person is autonomous, he is a good person, and so he does the right action, which is to execute an arbitrary program, which yields one bit of uncomputability, which is good. On the other hand, when a person is not autonomous, he executes a specific program, which yields no bit of uncomputability. A person would do this if he treated humans as a mere means, and by extension did not value relations of mutual recognition. [11]

4 Conclusion

I described an ethical theory which takes the mathematical structure of consequentialism, but which has a Kantian value function. The theory covers some weaknesses of consequentialism, natural goodness theory, and contractualism, while keeping their strengths intact. Consequentialism's weaknesses were resolved in the manner just described. Natural goodness theory's one weakness lies in its impreciseness and its inability to pin down exactly what is meant by "natural normativity assertions". Our system addresses this issue with mathematical precision. Contractualism's one weakness lies in its conflation of a program and the execution of a program. Our system uses that distinction as the very core of its thrust.

When Kant said we must not treat humans as mere means, but ends in themselves, he was saying that humans are arbitrary programs, not specific programs. Therefore, we *cannot* use them as mere means, and *must* treat them as ends in themselves. And the sense just used in *cannot* and *must* is not that of mere moral indoctrination. It is the authority of mathematics. So Williams has been answered: we legislate to the moral sentiments *by right of mathematical fact*.

[9] *Externally* uncomputable, that is; it is *internally* computable, because we already assumed that all thought is computable at the end of section 2.

[10] The use of the word "arbitrary" is a technical one. The unsolvability of the *Entscheidungsproblem* tells us that no program exists which tells us whether some arbitrary program ever stops running. The "arbitrary" is necessary here; if the theorem read instead, "no program exists which tells us whether some program ever stops runnning", I can easily say, "yes, it does: look at this program here, $ADDITION$, which adds two numbers together. I can write a program that tells us that $ADDITION$ always stops running, and I would be correct." The term "arbitrary" blocks us from pointing to a specific program, and thus prohibits me from saying that previous sentence.

[11] See the appendix for a discussion on when, exactly, a person would be executing a specific program, and how we would know.

5 Appendix

Proof of proposition: A human H ought not to be computing an arbitrary human H'.

Assumption 1. *The Church-Turing Thesis is true: everything that is physically computable is computable by some Turing machine.*

Definition 1. *A human H is a thing that does computation and is in the physical world.*

Explanation: In other words, a human H is an automaton to which the Church-Turing Thesis applies; for each thought process of human H, there exists a Turing machine.

Definition 2. *A human H is free if and only if H is uncomputable.*

Corollary: A human H is not free if and only if H is computable.

Definition 3. *We say a human H "ought not to be" executing some Turing machine M in the case that H is not free if H is executing M.*

Proposition 1. *A human H is at most Turing-complete.*

Proof: This follows from Assumption 1, that the Church-Turing Thesis is true.

Proposition 2. *There exists no Turing machine M that computes the output of an arbitrary Turing machine A.*

Proof: This follows from the undecidability of the halting problem.

Proposition 3. *A human H cannot compute an uncomputable function.*

Proof: By Proposition 1, H cannot compute any function no Turing machine can compute. No Turing machine can compute an uncomputable function. Therefore H cannot compute an uncomputable function.

Definition 4. *An automaton S is said to be "stronger" than an automaton W if and only if the functions W can compute is a strict subset of the functions M can compute. Conversely, W is "weaker" than S if and only if S is said to be "stronger" than W.*

Explanation: This definition exists purely for the sake of linguistic convenience. In each subsequent proposition, replace "stronger" or "weaker" with the formal definition here.

Proposition 4. *For some automaton M, if M is computing the output of an arbitrary Turing machine A, M is either stronger than or weaker than a universal Turing machine.*

Proof: By Proposition 2, no Turing machine M computes the output of an arbitrary Turing machine A. Therefore, if M computes the output of an arbitrary Turing machine, M is not a Turing machine. In particular, M is not a universal Turing machine. There are two possibilities for M. (1) M is Turing-complete and has extra computing capabilities. For example, M may be a universal Turing machine with a halting problem oracle. (2) M is sub-Turing-complete, that is, there are Turing machines which M cannot simulate. Therefore, in this case, M is either stronger than or weaker than a universal Turing machine.

Proposition 5. *If a human H is computing the output of an arbitrary Turing machine A, H is weaker than a universal Turing machine.*

Proof: By Definition 1, a human H is an automaton. By Proposition 4, if an automaton H computes the output of an arbitrary Turing machine A, H is either stronger or weaker than a universal Turing machine. By Proposition 1, a human H is no stronger than a Turing-complete machine. Therefore, H is weaker than a universal Turing machine.

Proposition 6. *If a human H is computing the output of an arbitrary Turing machine A, H is computable by some Turing machine.*

Proof: By Proposition 5, if a human H is computing the output of an arbitrary Turing machine A, H is weaker than a universal Turing machine. Therefore, H is a sub-Turing-complete machine.

Lemma 1: There exists a Turing machine that can compute the outcome of any sub-Turing-complete machine. *Proof is left as an exercise for the reader.*

By Lemma 1, if H is a sub-Turing complete machine, H is computable by some Turing machine.

Proposition 7. *If a human H is computing the output of an arbitrary Turing machine A, H is not free.*

Proof: By Proposition 6, if a human H is computing the output of an arbitrary Turing machine A, H is computable by some Turing machine. By the corollary to Definition 2, if H is computable, H is not free.

Proposition 8. *If a human H is computing a free human H', H is not free.*

Proof: By Definition 2, a human H' is free if and only if H' is uncomputable. By Proposition 2, H' is at most Turing-complete. Because H' is uncomputable, H' must be at least Turing-complete. Therefore, H' is exactly Turing-complete. To compute the output of a Turing-complete machine is tantamount to computing the output of an arbitrary Turing machine. By Proposition 7, if a human H computes the output of an arbitrary Turing machine A, H is not free. Therefore, if a human H computes the output of a Turing-complete machine H', H is not free. Therefore, if a human H computes a free human H', H is not free.

So far, so good. However, at this point, one problem remains. A human H may compute some H' and simply claim that H' is not free, therefore H is free. But how should H know if H' is free or not? We show that there is no Turing machine to do just that. This lets us squeeze out a stronger result: *If a human H is computing an arbitrary human H', H is not free.* This formalizes the intuitive dictum, "all persons are innocent until proven guilty."

Proposition 9. *There is no Turing machine M that takes as input an arbitrary human H and outputs whether H is free or not.*

Proof: Suppose such a Turing machine M exists. Then M takes as input an automaton H and outputs whether H is an arbitrary Turing machine or not. If H were an arbitrary Turing machine, M could not know if H halts or not. If H were a sub-Turing-complete machine, then M can run H until it halts. Any Turing machine that halts can be simulated by a sub-Turing-complete machine. If H were to halt, H can be simulated by a sub-Turing-complete machine. Therefore M is equivalent to the solution to the halting problem. Therefore no M exists.

Remark: Clearly, there exists a Turing machine M that takes as input a human H with a specific semantic description – namely, that H is computing a free human H' – and outputs whether H is free or not: that Turing machine is described by Propositions 1-8. We may gain such a semantic description about H through, for example, something H has said or done. However, we are talking here about an *arbitrary* human H that may or may not possess this semantic description.[12] We have shown that, in this general case, there exists no such M.

Proposition 10. *If a human H is computing an arbitrary human H', H is not free.*

Proof: By Proposition 10, no Turing machine M exists that takes as input an arbitrary human H and outputs whether H is free or not. The rest of the proof mirrors the structure of the proof to Proposition 8.

Proposition 11. *A human H ought not to be computing an arbitrary human H'.*

Proof: This follows from Definition 3 and Proposition 10.[13]

[12] Interestingly, by Rice's Theorem, there is no Turing machine that gives us any such semantic description.

[13] A first version of this proof appeared in [Bae17a], then a revised version in [Bae17b].

References

[Tur37] A. M. Turing. "On Computable Numbers, with an Application to the Entscheidungsproblem". In: *Proceedings of the London Mathematical Society* s2-42.1 (1937), pp. 230–265. DOI: 10.1112/plms/s2-42.1.230. eprint: /oup/backfile/content_public/journal/plms/s2-42/1/10.1112/plms/s2-42.1.230/2/s2-42-1-230.pdf. URL: +%20http://dx.doi.org/10.1112/plms/s2-42.1.230.

[Raw71] John Rawls. *A Theory of Justice*. Harvard University Press, 1971.

[Wil81] Bernard Williams. *Moral Luck: Philosophical Papers 1973–1980*. Cambridge University Press, 1981. DOI: 10.1017/CBO9781139165860.

[Sch82] Samuel Scheffler. "The Rejection of Consequentialism". In: *Philosophical Review* 93.3 (1982), pp. 489–492.

[Rai84] Peter Railton. "Alienation, Consequentialism, and the Demands of Morality". In: *Philosophy & Public Affairs* 13.2 (1984), pp. 134–171. ISSN: 00483915, 10884963. URL: http://www.jstor.org/stable/2265273.

[HM85] D. Hume and E.C. Mossner. *A Treatise of Human Nature*. Classics Series. Penguin Books Limited, 1985. ISBN: 9780140432442. URL: https://books.google.com/books?id=JPCgWYSZCUwC.

[Wil88] Bernard Williams. "Consequentialism and Integrity". In: *Consequentialism and its Critics*. Ed. by Samuel Scheffler. Oxford University Press, 1988, pp. 20–50.

[KO96] Christine M. Korsgaard and Onora O'Neill. "The Sources of Normativity". In: (1996). DOI: 10.1017/CBO9780511554476.

[Sca98] Thomas Scanlon. *What We Owe to Each Other*. Belknap Press of Harvard University Press, 1998.

[Foo01] Philippa Foot. *Natural Goodness*. Oxford University Press, 2001.

[Sun06] Phil Sung. "Maximum Satisfiability". In: (2006). URL: http://math.mit.edu/~goemans/18434S06/max-sat-phil.pdf.

[RN09] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. 3rd. Upper Saddle River, NJ, USA: Prentice Hall Press, 2009. ISBN: 0136042597, 9780136042594.

[Sid11] Henry Sidgwick. *The Methods of Ethics*. Cambridge Library Collection - Philosophy. Cambridge University Press, 2011. DOI: 10.1017/CBO9781139136617.

[Bae17a] Jongmin Jerome Baek. *Culture, Computation, Morality*. arXiv, 2017. URL: https://arxiv.org/abs/1705.08502.

[Bae17b] Jongmin Jerome Baek. *Culture, Computation, Morality, or: The Poetry of Computer Science, the Computer Science of Poetry*. Philosophy of Computation at Berkeley Press, 2017. URL: https://books.google.com/books?id=6jFBDwAAQBAJ.

[MD17] Robert J. Matthews and Eli Dresner. "Measurement and Computational Skepticism". In: *Noûs* 51.4 (2017), pp. 832–854. ISSN: 1468-0068. DOI: 10.1111/nous.12142. URL: http://dx.doi.org/10.1111/nous.12142.