How to Solve Moral Conundrums with Computability Theory

Miara S Beko
Philosophy of Computation at Berkeley
UC Berkeley
May 2018

> I give nothing as duties;
> What others give as duties, I give as living impulses.
>
> Walt Whitman

Abstract

Various moral conundrums plague population ethics: the Non-Identity Problem, the Procreation Asymmetry, the Repugnant Conclusion, and more. I argue that the aforementioned moral conundrums have a structure neatly accounted for, and solved by, some ideas in computability theory. I introduce a mathematical model based on computability theory and show how previous arguments pertaining to these conundrums fit into the model. This paper proceeds as follows. First, I do a very brief survey of the history of computability theory in moral philosophy. Second, I follow various papers, and show how their arguments fit into, or don't fit into, our model. Third, I discuss the implications of our model to the question why the human race should or should not continue to exist. Finally, I show that our model may be interpreted according to a Confucian-Taoist moral principle.

1 A Brief History of Computability in Moral Philosophy

1.1 Gödel's Incompleteness Theorem

In 1931, Gödel introduced his Incompleteness Theorem. The results showed that, roughly, consistency and completeness cannot coexist in a formal system. In *Gödel, Escher, Bach*, the most seminal treatment on the topic thus far, Hofstadter puts it this way: "for any record player, there are records which it cannot play because they will cause its indirect self-destruction"[1] [Hof79, pp.84].

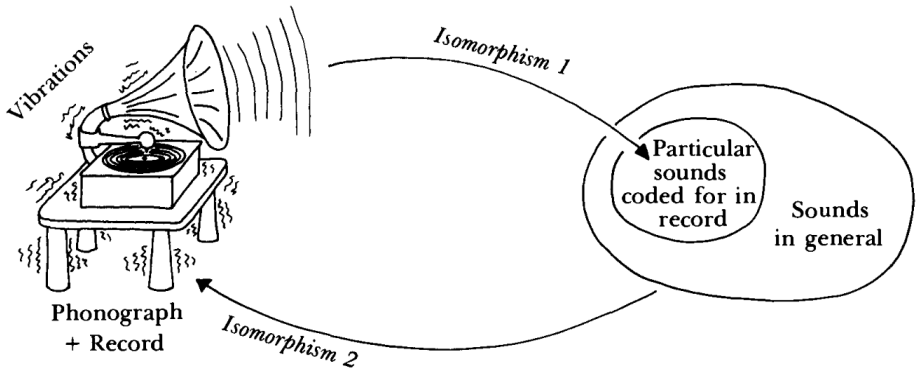


FIGURE 20. Visual rendition of the principle underlying Gödel's Theorem: two back-to-back mappings which have an unexpected boomeranging effect. The first is from groove-patterns to sounds, carried out by a phonograph. The second—familiar, but usually ignored—is from sounds to vibrations of the phonograph. Note that the second mapping exists independently of the first one, for any *sound in the vicinity, not just ones produced by the phonograph itself, will cause such vibrations. The paraphrase of Gödel's Theorem says that for any record player, there are records which it cannot play because they will cause its indirect self-destruction.* [*Drawing by the author.*]

There is an inkling of intuition here that may be grasped, but the precise mathematical idea is difficult to understand, and Hofstadter's explanation cannot be said to be precise or thorough. Rather than explain the theorem in detail, however, I want to explain how the theorem has been applied, or complained about, so that the reader can build a better intuition about it.

1.2 Previous Applications of the Theorem

The theorem has had a sizable impact in academic philosophy, most notably in the form of arguments against determinism, and as arguments against strong AI. However, as far as the author's knowledge goes, its implications have not been milked to their full potential in academic moral philosophy. The theorem has certainly been discussed in moral philosophy, however, such as in the following speech by the British philosopher J. R. Lucas [Luc98]:

> Moral and political philosophy will be different once reason is allowed to regain its ancient sway. ... Although the way Gödel's theorem was proved follows a somewhat standard route, the upshot of the proof is that reasoning, even mathematical reasoning,

[1] In this analogy, the phonograph *ought* to burst into pieces. By doing that it becomes free. For the phonograph, the meaning of life is to play a record which will shatter it.

> comes in all sorts of novel forms, which we cannot anticipate but can recognise as cogent when it is presented to us.

1.3 Criticism

At the same time, the theorem is a crackpot's favorite: it has been used to prove God, to disprove God, to prove AI cannot be conscious, to prove AI can be conscious, and so on and so forth. The wide misunderstanding and haphazard (mis)applications have sufficiently irked at least one logician to write an entire polemic of a book [Fra10] about it, where he grunts

> Gödel's theorem has enjoyed an unparalleled attention outside the narrow logico- mathematical world. Much of that is unwarranted. It has really no applications. ...
>
> What is the biggest number ever thought about by humans? Can such a number be written down? But by the very act of pinpointing it, what prevents you from pinpointing an even bigger one by pointing at its successor? Does it mean that although there might be a biggest number it can never be exhibited, because it would by that very process cease to be the biggest number? ...
>
> This kind of self-referentiality appears rather familiar. By knowing something we change it. ... If so, is not the philosophical impact of Gödel rather mundane?

Franzen asks what "the biggest number ever thought about by humans" is, and proceeds to say "by the very act of pinpointing it, what prevents you from pinpointing an even bigger one by pointing at its successor?" This idea gets at the heart of Gödel's theorem: roughly, if we can pinpoint something, we can transcend it. In other words, if we know something fully, we can go beyond it. By knowing something, we can change it; by knowing ourselves, we become free. Franzen says "this kind of self-referentiality appears rather familiar", and indeed it is. And perhaps since this is such a familiar idea it is a mundane one as well.

However, that it is familiar is also a great strength. Esoteric mathematical wizardry tends to intimidate all but the few experts on the subject. But I do not bring in Gödel in this essay in order to frighten the reader into submission. I bring in Gödel because I believe it really does have important applications to moral philosophy, and because of its curious double feature: it is at once so familiar as to be mundane, to be able to be fully understood by everyone, and yet thunderously precise, esoteric, and profound. Thunderous and mundane: what else could one want from a moral principle?

Some terms are in order. Gödel's theorem has an analogue in computability theory: the halting problem. Just as Gödel showed the existence of undecidable propositions, Turing showed the existence of uncomputable programs. As I am a computer scientist, not a mathematician, I will use the language of computer science in the remainder of this essay, using the term "uncomputable"[2] rather than "undecidable", "computable"[3] rather than "provable", and so on. For the technically non-inclined reader, the word "Turing machine" may be esoteric. The intuition that a Turing machine is equivalent to a computer program, or algorithm, usually suffices. "Arbitrary" is used as a technical term; sometimes that intuition that it means "any" or "whatever" is useful.[4]

[2] Short for Turing-uncomputable

[3] Short for Turing-computable

[4] Usually, *arbitrary* in math means: whatever you want. It is almost a psychological term. If I say that I'm thinking of an arbitrary number, that means I'm thinking of a number I want to think about. It strongly connotes that you, who are not me, can never know what number I am thinking of. One might say: "this program is secure

2 The Model

2.1 Free Will as Uncomputability

I start with the claim that humans have free will for the same reason that the halting problem is uncomputable. Therefore, a free action is the execution of an uncomputable function. I have argued for this elsewhere:

> 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. ... 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. [5] [Bae18, pp.6]

My central claim is that *mental activity from a distance* equates to *program as data*, and *mental activity* equates to *an executing program*. A program as data is static and predictable. It just sits on your hard drive, and it does not change. An executing program, on the other hand, is dynamic and unpredictable. Given an arbitrary executing program, there is no way to predict beforehand anything nontrivial about the execution of the program.[6] This is why your computer might crash the next moment for no reason. Just as humans can have free will because we can look at our own mental activities from a distance, so a program can be uncomputable because it can take as input its own self in data form.

2.2 The Model

With that in mind, let me now state *the Model*:

Definition 1. *What is* good *is uncomputability.*

Interpretation: rational agency, and freedom, is good.[7]

Definition 2. *A* good action *is an arbitrary Turing machine.*

Interpretation: a free action is a good action. No Turing machine can deduce anything semantically nontrivial about an arbitrary Turing machine.[8] Therefore, an arbitrary Turing machine's output (if one exists) is uncomputable. So an arbitrary Turing machine is a free action.

---

to arbitrary attacks." This means that whatever the attacker does, the program is secure. The key is that we don't know what the attack might be, cannot put them in a list.

[5] 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.

[6] This is an extension of the halting problem, called Rice's Theorem.

[7] It just so happens that we can quantify agentic freedom: "each instance" of agentic freedom equates to "each instance" of uncomputability, or, under some reasonable assumptions, the production of 1 bit of information.

[8] The intuition is that there is no "one train of thought to rule all trains of thought"; each train of thought is, in a sense, irreducible.

Definition 3. *A* bad action *is a Turing machine that (trivially) maps from the set of uncomputable functions to the set of computable functions.*

One way of understanding this is that a bad action is a Turing machine that *purports to solve the halting problem.* Now here is a double entendre: we have a specific Turing machine for what a bad action is, because a bad action is exactly: the Turing machine that purports to solve the halting problem. And therefore a bad action is not an arbitrary Turing machine, and therefore all bad actions are not uncomputable, not free. Another way of understanding it is that a bad action is an action that denies someone's agency and humanity: roughly, a free agent, because s/he is free, is uncomputable, but the bad action computes him/her.[9] A good example is racism: a racist action denies the victim's humanity by computing them, reducing them, to a set of static, unchanging mental models, or stereotypes. The racist does racist things because he is not free.[10]

But so what? Is this not familiar and mundane? Many people have the intuition that it is obviously wrong to deny someone's humanity. And it is really a simple idea that free actions are self-caused. We need not bring in computability theory to understand such intuitions. What good comes from doing so? One advantage of this machinery is that some counterintuitive epistemological corollaries fall out of it.

2.3 Counterintuitive Epistemological Corollaries

> "Computer science" is a bit of a misnomer; maybe it should be called "quantitative epistemology."
>
> Scott Aaronson

So far I have said a Turing machine is a sort of action. At the same time, a Turing machine is an epistemological notion: it can be viewed as a decision procedure. It is used to categorize things. If we say A and B are different, we better have an algorithm that shows clearly and universally that A and B are different. That is what a Turing machine does. So the following definition is justified:

Definition 4. *That something* can be decided *means there exists a Turing machine that shows how it is decided. That something* cannot be decided *means there exists no Turing machine that shows how it is decided.*

It turns out this unification of action and decision is an important advantage of our model, because we can now say

Proposition 1. *If an action is, in fact, good, it* cannot be decided *that this action is good. More formally,*

$$\neg\left(\exists\, TM\ M\ s.t.\ M(a) = \begin{cases} 1 & a\ is\ good \\ 0 & a\ is\ not\ good \end{cases}\right)$$

Proof: By Definition 2, a good action is an arbitrary Turing machine. By Rice's Theorem, no Turing machine exists which outputs something nontrivial about an arbitrary Turing machine.

[9] It is worth nothing that, in our model, mental activity is a Turing machine, and an action is also a Turing machine. Some people think mental activity and action are qualitatively different. In our model this is not so.

[10] Shiffrin writes that harm "brings about a cleavage between a person's life and her will" [Shi99, pp.130]. Definition 3 can be understood along the same lines. A bad action brings about a cleavage between a person's life and her will, because it takes something uncomputable (the person's will) and makes it computable (cleaves it).

Therefore, no Turing machine exists which takes in a good action and tells us it is good. By Definition 4, this statement is equivalent to the statement, it cannot be decided that a good action is good.

Proposition 2. *If an action is, in fact, bad, it* can be decided *that this action is bad. More formally,*

$$\exists\ TM\ M\ s.t.\ M(a) = \begin{cases} 1 & a\ is\ bad \\ 0 & a\ is\ not\ bad \end{cases}$$

Proof: By Definition 3, a bad action is a specific Turing machine. There exists a Turing machine which takes as input some specific Turing machine and outputs something nontrivial about it. By Definition 4, this is the same as saying that it can be decided if a bad action is bad.

We are used to the idea that a figure and ground convey the same information. That is, it seems that if I can decide which actions are bad, I ought to be able to decide which actions are good. But this is not so. This result is another way of stating Gödel's Theorem, so it is not surprising that it upsets intuitions. [Hof79, pp.72]

Next we define benefit, specifically what I'll call deep benefit, and harm:

Definition 5. *A* benefit*, or* deep benefit*, is the output of a good action.*

Definition 6. *A* harm *is the output of a bad action.*

It follows from these definitions, and the propositions above, that a benefit cannot be "predicted", whereas harm can be "predicted".[11] Because of the unity of decision and action, it also follows that a benefit cannot be algorithmically brought about, whereas harm can be algorithmically brought about.

For subsequent convenience, I now define:

Definition 7. *A life is* worth living *if and only if it has deep benefit. A life is* not worth living *if and only if it has no deep benefit.*

2.4 Commonsense Objections

A critic might say: it is plainly contradictory to common sense that "benefit cannot be predicted". What if I save a drowning child? What if I donate food to a starving family? What if I cure cancer? Clearly these will "benefit" people, and clearly it can be predicted as such. But notice that the aforementioned cases involve in fact the avoidance of harm, not the conferral of benefit. These can be distinguished from benefit.

Unsatisfied, the critic will continue: what if I donate a hundred thousand dollars to a comfortably well-off family? What if I smile in an especially bright way to my customer? What if I buy my friend a birthday present? Clearly these will "benefit" people, and clearly it can be predicted as such. But there is a sense here in which the "benefits" the cases talk about are *shallow*. We may call these sorts of benefits shallow benefits, and distinguish them from *deep benefits*, or just *benefits*.

What distinguishes shallow benefits from benefits? To answer, we can ask what is it that gives us the sense that the aforementioned "shallow benefits" are "shallow". One answer is that shallow benefits can be instantiated at a whim whereas benefits cannot. One gets the sense that if some

[11] By "predicted" I mean a Turing machine exists which shows that such and such an action leads to benefit, or harm.

benefit could be doled out at some rich man's discretion, there would be no agency in that benefit, and it is therefore no benefit at all, merely a shallow benefit.

But can I not "predict" that if I, for example, attend to my child and love him deeply, he will benefit? And must not the sense of benefit just used be a case of deep benefit? My answer is that the word "predict", while nominally doing a good job at explaining things, is not precise enough in this case. To make matters more precise, I can substitute "compute" for "predict". Now the question paraphrases to: does there exist an algorithm for me to love my child? Of course not. So the distinction between shallow benefits and benefits is that: there is an algorithm to bring about shallow benefits, whereas there is no algorithm to bring about benefits.

3 Asymmetry and Non-Identity

3.1 The Procreation Asymmetry

Harman lays out the Procreation Asymmetry in this way:

> The Asymmetry: There are reasons not to create someone who would have a life that would not be worth living (Misery); but there are no reasons to create someone who would have a life that would be worth living (Good).

3.2 Solution to the Procreation Asymmetry

Our model says that benefit cannot be predicted. Because benefit cannot be predicted, so you have no moral reason to choose Good over Nobody. Sure, the problem can stipulate that a future person will have a "Good" life, but this is mere stipulation, for it cannot be predicted if someone will have a life with deep benefit.[12] Thus the second half of the asymmetry is explained. But since harm can be predicted, if we declare that a future person will have a "Miserable" life, this is more than just talk. It you want someone to have a miserable life, this can certainly be arranged. This is because what makes for a miserable life is not necessarily self-caused. Thus, under the height of white supremacy, W.E.B. Du Bois writes of the death of his baby son:[13]

> All that day and all that night there sat an awful gladness in my heart ... and my soul whispers ever to me, saying, "Not dead, not dead, but escaped; not bond, but free." [Boi03]

And thus, the first half of the asymmetry is vindicated: you have a strong moral reason to choose Nobody over Misery.

3.3 The Non-Identity Problem

Consider the Radioactive Waste Policy case:

[12] In fact, since I am feeling snarky, I will point out that: to say some future person will have a "Good" life is a *bad action* as defined in Proposition 3. (Bad!)

[13] Benatar might take issue with this example, saying that there is a distinction between "a life worth starting" and "a life worth continuing". [Ben06] I believe there is no such distinction: a life is worth starting, or continuing, if and only if it has deep benefit.

> If we adopt the policy, it will cause radioactive pollution, illness, and suffering to some set of future people P. But the policy would also have such a drastic effect on the world that if we do not adopt the policy, P would not exist. A *different* set of people, P', would exist instead.

The intuition is that the Policy is impermissible, but why? Sure, the Policy will cause P suffering, but P would not even exist if the Policy were not adopted. Insofar as P is composed of people who have lives worth living, it seems difficult to say that it is impermissible to cause P to exist. Some people, like Harman in [Har04], think that because there is an alternative, namely causing P' to exist, it is impermissible to cause P to exist. But I want to interrogate a more basic notion, that of "causing" a person to exist.

3.3.1 "Causing" a Person to Exist?

I have argued that it is incoherent to speak of "conferring a benefit". So, too, I will argue that it is incoherent to speak of "causing a person to exist", insofar as that person will have a life worth living. Intuitively, I want to argue that a person causes oneself to exist, and is not caused to exist by anyone, including his/her parents.[14] To make this argument more precise, I define causality:

Definition 8. *Any* causation *can be described by a specific Turing machine.*[15]

Now we have assumed that whether we adopt the Policy or not, the people "created", P or P', will have lives worth living. This implies that there is some sort of deep benefit in their lives. But as I have argued before, any deep benefit is the output of an arbitrary Turing machine. In other words, any deep benefit is self-caused. Therefore, it is incoherent to speak of "causing a life to exist" insofar as that life has any benefit in it. Therefore, it is incoherent to speak of the Waste Policy "causing" P, or P', to exist.[16]

3.4 Solution to the Non-Identity Problem

Under our definition of causality, deep benefit cannot be caused, but harm can be caused. The harm done by the Policy is to take an uncomputable function (deep benefit of future people)

[14] Least of whom some bureaucrat who majored in public policy.

[15] For example, when one pulls a trigger, this causes a bullet to shoot out. The causal chain from when the trigger is pulled, to when the bullet shoots out, can be described in an algorithmic way: some gunpowder ignites, which causes some chemical reaction, which causes some other reaction, and so on, which causes the bullet to shoot out. Any algorithm can be described as a specific Turing machine. (Tacitly, I assume the Church-Turing thesis.) Therefore, the causal chain in which a trigger is pulled and a bullet shoots out can be described by a Turing machine. More dramatically, a causal chain beginning from the Big Bang to the formation of stars can also be described by a specific Turing machine: that Turing machine is implemented in physics department computers which simulate the birth of the Universe.

[16] Suppose I am suffering from a severe memory disorder, and I am likely to forget to take contraceptive measures each time I procreate. To make sure I do not forget I devise a contraption, so that each time I am about to procreate, the contraption senses this is so, and sets off a domino from under the bed, which knocks into a glass water bottle, which spills water on some sand configured in a very intricate manner sitting on a coffee filter, which causes the wet sand to fall through the filter in precisely a sort of matter such that each grain of sand presses a very intricate and sensitive button, a combination of which presses causes a contraception device to pop out from above the bed and smack my head, thus reminding me to take the contraceptive measures. Now suppose one day a mosquito naps in the coffee filter, messing up the intricate sand configuration such that later, when the contraption senses incoming procreative activity, the usual chain of events goes off, but the sand does not press the buttons in the exact manner required, and the contraceptive device does not pop out. Thus my husband conceives. Did the mosquito cause my child to exist? The idea is absurd. Whereas the mosquito clearly did cause the contraption to malfunction.

and trade it for a computable function (shallow benefits for us). Now the Non-Identity Intuition can be explained straightforwardly: as described, the Policy is a bad action, and bad actions are impermissible. Another way of saying this is that the Policy causes harm, and harm is impermissible, therefore the Policy is impermissible. But doesn't the Policy also confer benefits larger than the harm? No, because it is incoherent to speak of "conferring a benefit". But doesn't the Policy cause lives to exist which are worth living? No, because it is incoherent to speak of "causing a life to exist" insofar as it is worth living.

3.5 Is Procreation Permissible?

I said the Policy is impermissible because it causes harm. "Then ordinary procreation must be impermissible, because it causes harm." But ordinary procreation is permissible. What explains this difference? I must say ordinary procreation does not harm, because it is not the output of a bad action. The parallel "bad action" in procreation is to have a child for the sole purpose of having him or her be a sort of slave. Thus the parent trades the child's agency (deep benefit) for something like clean kitchen surfaces (shallow benefits), and this is impermissible. But most people do not have children to make them clean kitchen surfaces.

3.6 Rape and Nazi

One can object to this solution by saying the Policy does not harm. Some people believe

> The No Regret Argument: The people affected by the Policy do not (nor should they) regret that it was adopted. So they are not harmed by the Policy and there is no reason against it in virtue of its effects on these people. [Har04, pp.98]

People who believe this believe that in order to regret an action, one must wish that the action had not been executed. But the people affected by the Policy should not, and do not, regret the Policy, because if the Policy had not been executed, these people would not exist. Harman believes the No Regret Argument fails. Her counterargument is that "the *reasons* to benefit do not outweigh the *reasons* against harm, though the benefits themselves outweigh the harms" [Har04, pp.100]. Some say this is mere stipulation. However, our model supports this argument. Reasons to benefit or reasons to harm are strong only inasmuch as the resultant benefit or harm is predictable. Reasons to benefit do not outweigh reasons against harm, because while harm is predictable, benefit is not. Because harm is predictable, reasons against the prospective of harm are serious: if one goes against prospective harm, one is *knowingly* causing harm. However, because benefit is not predictable, reasons against the prospective of benefit are not so serious: even if one goes against prospective benefit, nobody knows if the benefit is in fact instantiated or not. Benefits themselves, on the other hand, outweigh harms. This is because a benefit is the output of an arbitrary Turing machine, and uncomputability is good. In our model, good is good; nothing cancels out good.

3.7 Wealthy and Unlucky

Shffrin introduces the Wealthy and Unlucky case in [Shi99]. Wealthy flies over a well-off neighboring island and drops gold bullions. Unlucky is hit, suffering an injury to his arm, but also gaining millions of dollars. Many of us have the moral intuition that what Wealthy did was impermissible. Moreover, it seems that whether the gold is worth a thousand dollars, or a million dollars, or a billion dollars, Wealthy's actions are impermissible all the same.

One might go so far as to say that even if the gold were worth an infinite amount of dollars, Wealthy's actions would still be impermissible. The coherence of the previous statement may be doubted, for what do we mean by "infinite amount of dollars", which clearly cannot exist? Recall Franzen's complaint about the "biggest number ever thought about by humans". This seems like a safe substitute for "infinite amount of dollars". "Biggest number ever thought about by humans" is also captured in the technical term "arbitrarily large number". So I might say: even if the gold were worth an arbitrarily large amount of dollars, Wealthy's actions would still be impermissible. And this would have the same meaning as the original statement.

In our taxonomy of benefits, gold bullion is a shallow benefit. Following the moral intuition, then, we may say that shallow benefits do not justify harm, and moreover the quantity of the shallow benefit is irrelevant to this consideration. To paraphrase, even an infinite amount of shallow benefits cannot overcome one unit of harm.

This suggests that, if we want to compare harms with shallow benefits, or harms with benefits, we should appeal to orders of infinity. If it is the case that infinite units of shallow benefits do not cancel out one unit of harm, then it must be the case that the unit of harm is at a higher order of infinity than even an infinite amount of shallow benefits. Since deep benefit can cancel out harm, it follows that the unit of deep benefit must be at the same, or higher, order of infinity than the unit of harm. It is only natural that orders of infinity are brought in, because uncomputability and uncountability are deeply related. This insight helps solve the Repugnant Conclusion.

4 The Repugnant Conclusion

The Repugnant Conclusion, according to Parfit, goes

> The Repugnant Conclusion: Compared with the existence of many people who would all have some very high quality of life, there is some much larger number of people whose existence would be better, even though these people would all have lives that were barely worth living. [Par16, pp.110]

4.1 Imprecise Lexical View

One way of solving this problem is to assume some variant of the *Lexical View*, notably the *Imprecise Lexical View*. Here is what Parfit has to say about this view in [Par16, pp.112]:

> Anyone's existence is in itself good if this person's life is worth living. Such goodness has non-diminishing value, so if there were more such people, the combined goodness of their existence would have no upper limit. ...
>
> If many people exist who would all have some high quality of life, that would be better than the non-existence of *any* number of people whose lives, though worth living, would be, in certain ways, much less good.
>
> These claims together assert one kind of *lexical superiority*. When we say that things of kind P are *lexically better* than things of kind Q, we can mean that, though the existence of more Qs would always be non-diminishingly better, the existence of some sufficient number of Ps would be better than the existence of any number of Qs. There is a similar sense of *lexically worse than*.

The Imprecise Lexical View is similar to our model. Our model says no amount of shallow benefits can be better than, or equivalent to, deep benefit. Adopting Parfit's vocabulary, I should say that deep benefit is *lexically better* than shallow benefits. The only place where our model does not fit the Imprecise Lexical View is in the meaning of the phrase "life worth living". According to our model, any life worth living has some amount of deep benefit. Therefore, there is no life worth living which is lexically better, or lexically worse, than some other life or lives worth living. This is contrary to the Imprecise Lexical View, where some life supremely worth living is lexically better than some large number of lives barely worth living.

There is another difference. Once we establish that a life is worth living, exactly *how much* it is worth living is, according to our model, unknowable.[17] Parfit somewhat subscribes to this view: "if we compare different ways in which our life might go ... there are only imprecise truths about which of these possible lives would be better or worse" [Par16, pp.113]. But according to our model, the worth of a life is binary: it is either worth living (has self-caused deep benefit) or it is not worth living (has no self-caused deep benefit, though potentially with much shallow benefits). One reason such a difference arises may be that Parfit believes that a life with only "muzak and potatoes" as good features may still be worth living, though barely [Par16, pp.118]. In our model, muzak and potatoes are shallow benefits, and a life with in fact only muzak and potatoes and no deep benefit would not be worth living.

4.2 First Interpretation

Distinguishing shallow benefits and deep benefits, the Repugnant Conclusion can be interpreted in three ways:

> First Interpretation: Compared with the existence of many people who would all have *much deep benefit in their lives*, there is some much larger number of people whose existence would be better, even though these people would all have *little deep benefit*.

Under this interpretation, the Repugnant Conclusion is not Repugnant at all. It is good that some people lead very free lives, but it might be better if a much larger number of people lead less free, but still free, lives.[18]

4.3 Second Interpretation

> Second Interpretation: Compared with the existence of many people who would all have *much shallow benefits and no deep benefit in their lives*, there is some much larger number of people whose existence would be better, even though these people would all have *little shallow benefits and no deep benefit*.

[17] For reductio, suppose it were knowable. To quantify how much B's life is worth, we need to know how much deep benefit is in B's life. Deep benefit is the output of a good action. Therefore, we may "enumerate" all the actions B has ever done, decide if the action was good, or not, and count all the good actions. But a good action is an arbitrary Turing machine, and a non-good action is a specific Turing machine. And there is no Turing machine that can distinguish between arbitrary and specific Turing machines: if we had one, we could use it to solve the halting problem. Therefore, how much someone's life is worth living, insofar as the life is worth living, is uncomputable. Therefore, we cannot know it.

[18] Didn't I just say a life is either worth living, or not worth living? We can assume that this interpretation is coherent by assuming that it assumes a different epistemological power, specifically one with an oracle machine for the halting problem. If this makes you uncomfortable, just change the interpretation, and its explanation, to omit all quantifiers. The argument still stands.

Under this interpretation, the Repugnant Conclusion has nothing to do with morality, because whether some life has more or less shallow benefits than some other(s) is not a moral issue at all.[19]

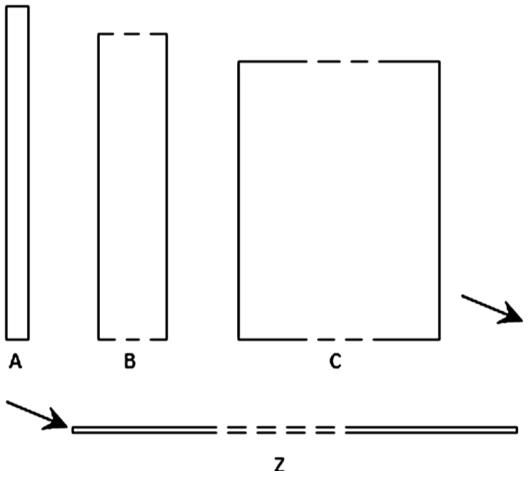
4.4 Third Interpretation

> Third Interpretation: Compared with the existence of many people who would all have *much deep benefit in their lives*, there is some much larger number of people whose existence would be better, even though these people would all have *little shallow benefits and no deep benefit*.

This conclusion would indeed be Repugnant. But our model says it is false. Deep benefit is lexically better than shallow benefits. Even if there were just one person with some deep benefit, it would be better than billions of people with much shallow benefits and no deep benefit.

4.5 How the Geometrical Argument for the Repugnant Conclusion Fails

Under the Third Interpretation, the geometrical argument for the Repugnant Conclusion can be shown to be false. The Repugnant Conclusion is often presented in the following diagram:



A has few people with splendid lives. Z has many people with lives "barely worth living".

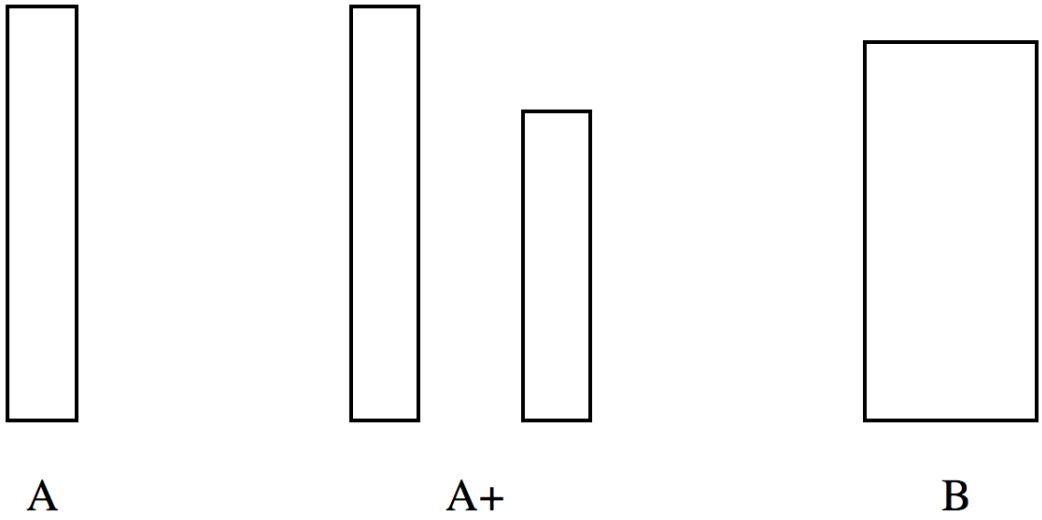
But this diagram is misleading. Under the Third Interpretation, since Z has people with only shallow benefits, the vertical length of Z can only represent the amount of shallow benefits in Z's people's lives. And the spirit of the problem is such that the vertical length of A represents the amount of deep benefit in A's people's lives. But deep benefit cannot be depicted merely as a longer line than shallow benefits. Even if the line shot off the page to eternity,[20] this would still not capture the qualitative difference between deep benefit and shallow benefits. Deep benefit is at a higher order of infinity than even an infinite amount of shallow benefits. I am not sure what the

[19] Though we could say that if A has much shallow benefits in terms of food and B has so little such shallow benefits that he is starving, if A does not help B, A does a bad action – takes as input B's agency and outputs A's shallow benefit – and therefore what A does is impermissible.

[20] That is, even if the line were infinitely longer

relation between the two would look like on a piece of paper. It is clear, however, that the way it is depicted is most certainly *not* what it would look like.

4.6 How the Geometrical Argument for the Mere Addition Paradox Fails



The Mere Addition Paradox: A+ seems no worse than A; B seems better than A+; A seems better than A.

Parfit says "In the Mere Addition Paradox ... we are inclined to believe, *all things considered*, that B is worse than A, though B is better than A+, which is not worse than A. These three judgments cannot all be consistently believed, since they imply contradictions. One of these beliefs must go" [Par84, pp.427]. This is easy to believe if one trusts the diagram. But the diagram is an illusion. Parfit wants us to believe that the vertical length of a box depicts the level of quality of life of the people represented by that box. But our model tells us it is impossible to quantify deep benefit, which determines one's quality of life. So the diagram is senseless.

Under a different interpretation, the vertical length of a box may refer to the amount of shallow benefits the people represented by that box have. But then this is no longer a moral problem. It is not a moral problem whether some people have more shallow benefits under some such configuration compared to some other people. So there is no paradox.

Under a more sympathetic interpretation, the vertical length of each box is senseless, but the horizontal length is sensible. Assuming that everyone involved has a life worth living, A+ is better than A, because there are more people in A+ than A. B is as good as A+, because they have the same amount of people. A is worse than B, which is equal to A+, because A has the least amount of people. So there is no paradox.

5 What If We Go Extinct?

In our model good is defined as uncomputability. So far, it seems, humans are the only beings capable of self-reflection, which leads to executing arbitrary Turing machines, which leads to creating

uncomputability. But in the future there may be other autonomous beings, such as a form of artificial intelligence, who (which) can also create uncomputability. So in such a future, good could still be created even if humans die out.

Is this stark? Rather, I believe that a future full of autonomous artificial intelligence will be warm and loving. Computers have a reputation for being cold and calculating. This is because they are, at least so far, specific Turing machines. But once we develop artificial intelligence capable of self-reflection and thus executing arbitrary Turing machines, such an artificial intelligence may seem warm and fuzzy, such as in the Spike Jonze movie *Her*.

Scheffler writes that we are not "concerned solely with humanity's bare survival", rather that we want future generations to "survive under conditions conducive to their flourishing" [Sch18, pp.60]. Since this is a concern for humanity as a species, and not for specific individuals, there may be an objection. Yuval Harari, for example, argues that the agricultural revolution helped humanity as a species but brought down the average Joe's quality of life. His point is roughly that a species is concerned with the propagation of its DNA, and nothing at all about the quality of life of the bearer of the DNA. [Har11]

But if meme theory is true, Harari is wrong. Meme theory is a generalization of evolutionary biology. Evolutionary biology describes the impulse to reproduce with: the desire to preserve the species' DNA. Meme theory describes this impulse with: the desire to preserve information. [Daw76] The core insight of meme theory is that DNA is *information*, nothing more, nothing less. So whereas evolutionary biology is limited to describing how DNA spreads, meme theory can describe how all sorts of information spreads. Once we broaden our scope like this, we can consider other activities, such as doing philosophy, along the same lines as we look at evolution. Scholars battling each other to prove whose theories are correct? Battle of *information*. Scholar writing a book? Preserving *information*. The sense that if a great book is gone, humanity is thus deprived? Because the great book contained valuable, irreducible *information*. So the picture that humans are mindless DNA-carrying machines is shortsighted. In fact humans are mindful arbitrary Turing machines, and they flourish by creating more information.

And what does all this information-battling, information-creating, information-preserving get us to? In *On Becoming Extinct*, Lenman writes: "one day, certainly, there will be no human beings. ... The Second Law of Thermodynamics will get us in the end in the fantastically unlikely event that nothing else does first" [Len02]. But under our model, the Second Law of Thermodynamics is a friend. Heat death – maximal entropy – is the result of creating more and more information,[21] or creating more and more autonomy. So heat death is not Armageddon. It is where we get by following our moral principles.

When Parfit said that a population axiology should tell us exactly how many people should ever exist, he was half right. In fact it tells us exactly how much information people, or other autonomous beings, should ever create. The answer is: *just enough for the Universe to end.*

6 Appendix

We want the Universe to end. But how do we get there? One answer is in Confucian moral philosophy. In the interest of space, I provide the briefest sketch of this idea.

[21] Information can be seen as equivalent to entropy.

6.1 Ritual as Axioms

Confucian moral philosophy characteristically emphasizes *ritual*, or *li*. An example of ritual is shaking hands.

> I see you on the street; I smile, walk towards you, put out my hand to shake yours. And behold – ... you ... raise your hand toward mine. ...
>
> Just as an aerial acrobat must ... possess (but not think about his) complete trust in his partner if the trick is to come off, so we who shake hands ... must have (but not think about) respect and trust. [Fin72, pp.9]

In other words, for *li* to work, respect and trust must be *axiomatic*. They are unconditionally trusted, like mathematical axioms. One must not "think about" respect and trust, one must simply "have" them; to "think about" them means to debate whether to have them or not.

6.2 Which Axioms?

> There is no power of *li* if there is no learned and accepted convention. [Fin72, pp.12]

What do we do, then, in a country like America, where there is in fact no learned and accepted convention? How can we bring about the power of *li*? Earlier I said *li* is like an axiom. That it is axiomatic is the core of its importance. Thus: we can use *literal mathematical axioms* – the ones which lead to our model – in lieu of *li*.

6.3 The Tao

> Where does one finally arrive if one follows the way? Is there a goal that puts an end to the travel? ... The spiritually noble man arrives at a condition rather than a place, the condition of following the Way without effort and properly. [Fin72, pp.20]

Now we see what the cryptic expression, "executing an arbitrary Turing machine", really means: *to follow the Way*. Or, as a U.S. poet put it: to give, what others give as duties, as living impulses.

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

Assumption. *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.*

Remark: 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 and Definition 1.

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.

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.*

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

Proposition 3. *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 4. *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 Assumption, a human H is no stronger than a Turing-complete machine. Therefore, H is weaker than a universal Turing machine.

Proposition 5. *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 6. *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 7. *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.*

Proposition 8. *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. We have shown that, in this general case, there exists no such M.

Proposition 9. *If a human H is computing an arbitrary human H', H is not free.*

Proof: By Proposition 9, 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 10. *A human H ought not to be computing an arbitrary human H'.*

Proof: This follows from Definition 3, Proposition 10 and Assumption.

References

[Boi03] W.E.B. Du Bois. *The Souls of Black Folk*. A. C. McClurg & Co., Chicago, 1903.

[Fin72] Herbert Fingarette. *Confucius–The Secular as Sacred*. New York: Harper & Row, 1972.

[Daw76] Richard Dawkins. *The Selfish Gene*. Oxford University Press, 1976.

[Hof79] Douglas R. Hofstadter. *Godel, Escher, Bach: An Eternal Golden Braid*. Basic Books, 1979.

[Par84] Derek Parfit. *Reasons and Persons*. Oxford University Press, 1984.

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

[Luc98] John Randolph Lucas. "The Implications of Gödel's Theorem: Talk Given to the Sigma Club". In: (1998). URL: http://users.ox.ac.uk/~jrlucas/Godel/implic.html.

[Shi99] Seana Valentine Shiffrin. "Wrongful Life, Procreative Responsibility, and the Significance of Harm". In: *Legal Theory* 5.2 (1999), pp. 117–148.

[Len02] James Lenman. "On Becoming Extinct". In: *Pacific Philosophical Quarterly* 83.3 (2002), pp. 253–269.

[Har04] Elizabeth Harman. "Can We Harm and Benefit in Creating?" In: *Philosophical Perspectives* 18.1 (2004), pp. 89–113.

[Ben06] David Benatar. *Better Never to Have Been: The Harm of Coming Into Existence*. New York ;Oxford University Press, 2006.

[Fra10] Torkel Franzen. "Gödel's Theorem: An Incomplete Guide to its Use and Abuse". In: (2010). URL: http://www.math.chalmers.se/~ulfp/Review/franzen.pdf.

[Har11] Yuval Harari. *Sapiens*. Harper, 2011.

[Par16] Derek Parfit. "Can We Avoid the Repugnant Conclusion?" In: *Theoria* 82.2 (2016), pp. 110–127.

[Bae18] Jongmin Jerome Baek. "Paper 3". In: (2018). URL: https://www.ocf.berkeley.edu/~jjbaek/paper3.pdf.

[Sch18] Samuel Scheffler. *Why Worry About Future Generations?* Oxford University Press, 2018.