Moral Experiments (draft)

Rio Popper* Kaarel Hänni†

December 2023

Abstract

I want to argue that we should 'experiment' morally. In particular, we should make moral decisions that are more likely to yield new information about the moral domain, even if those decisions are more likely to be morally wrong. We should therefore encourage moral risk-taking and punish moral wrongs less harshly in contexts where they yield some information (e.g. in nascent fields such as gene editing or artificial intelligence, or where the experimenter knows little about the broader moral domain because she is young or especially unworldly).

My argument has several parts. I begin by arguing that taking actions can in many cases give us moral information that we couldn't get without taking those actions. I then argue that, under several compelling assumptions, this information is morally significant and can eventually lead to making better subsequent moral choices. This is sufficient to establish the conclusion I sketched in the previous paragraph. In the next section of the paper, I show that (under some assumptions) moral experimentation is a public good—so we should societally subsidize it. I then discuss objections. All that done, I finally close by sketching what I think this should look like: by setting out, or at least starting to set out, what I want to change about our moral practice.

1 Introduction

Consider your friend, Alice, who asks you how she looks in an unfortunate new dress. In general, in these situations, let's say you lie: that seems morally better to you. This time, instead, you tell the truth. As (let's say) you predicted, she is a little hurt, but thanks you and changes into a different outfit. Crucially, you predicted exactly what Alice's reaction would be. But this doesn't rule out that you might update your moral beliefs — that is, your correct empirical prediction doesn't necessarily imply that you could have predicted your post facto moral beliefs. Maybe this experience shifts your intuitions that you should more often tell the truth in these cases.[1]

In a similar case, maybe you told Alice the truth specifically to test the claim that 'if I tell the truth, I'll feel that I've committed a (small) moral wrong'. If, despite having the right empirical predictions (that is, you correctly predict how Alice will respond), you don't feel you've committed such a moral wrong, then you might lessen your credence in your initial claim. This case is an example of a moral experiment. You formed a hypothesis, tested it, and revised your beliefs accordingly. And the hypothesis is a moral one — as I define it — because the beliefs in question are moral ones.

*rio.popper@philosophy.ox.ac.uk , Global Priorities Institute, University of Oxford

†kaarelh@gmail.com, Caltech and Cadenza Labs

Thanks to Paddy Allan, Catherine Brewer, Owen Cotton-Barratt, Andreas Mogensen, Sami Petersen, Helen Popper, and Tim Williamson for comments and suggestions.

1. Of course, one could say that intuitions are themselves empirical observations, but I omit a full discussion of that here.

The above is a more clear-cut 'moral' experiment than are most practical experiments. Many experiments mix the empirical and the moral. I think, for example, of something closer to an edge-case moral experiment: giving blunt critical feedback.[2] Sometimes my empirical predictions are wrong (the recipient is more upset than I expected) but still my moral intuitions shift: I do not require my moral experiments to cleanly isolate moral uncertainty — at least not in practice. Mixed together in this case are two experiments, moral and empirical. But the complicated and mixed nature of our moral beliefs does not undercut the distilled theoretical notion of a moral experiment. Often, moral uncertainty is simply mixed with other forms of uncertainty. While the remainder of this paper is concerned with strictly moral experiments, and central cases of those experiments, the work applies also to these kinds of edge cases where the empirical and the moral mix together.

Some understanding of moral experimentation — and the price we ought to pay to enable it — is already in the proverbial philosophical water supply. Most famously, J.S. Mill advocated for experiments in living, despite the 'inconvenience' to society that those experiments might engender (Mill 1859).[3] More recently, while discussing Mill's work, Anderson (1991) sets out an epistemic justification of such experiments in living.

But, just as moral experiments are not just empirical experiments, so too are they not just experiments in living. Experiments in living — while partly moral — are largely empirical. They resolve questions that might relate, for example, to how fit one is for some particular professional enterprise or how much one enjoys some type of relationship. And they usually have a wider scale than do moral experiments — although I do think that the line between the two is not entirely clear. A moral experiment might test a proposition that it is wrong to lie in such and such a kind of situation; an experiment in living might test a claim that people of my sort are happier in such and such a profession. And the feedback loops are likewise different. Experiments in living and in morality share one central kind of feedback: they both help us become aware of the relevance of new dimensions of the domain in which they operate. For example, a moral experiment might make me aware of the importance of — say — mercy, where before I only attended to justice; and in this way an experiment in living might function analogously. But in moral experiments, the feedback most of all is (as I discuss later) our moral intuitions, which I take as — admittedly messy and potentially worryingly biased — data about the moral domain.[4]

Another central difference between experiments in living and moral experiments is the cost potentially incurred by allowing them. Mill, for example, is quick to say that what 'inconveniences' society faces from 'self-regarding' experiments in living are easy for society to 'bear' (Mill 1859). When an experiment in living involves the interests of others and is worthy of 'moral disapprobation', Mill allows it to be prohibited on the basis of the harm principle.

Experiments in morality are different. Almost always, one kind of potential result of a moral experiment is a moral intuition of wrongness — that a particular action was *wrong*. Sometimes, one can conduct experiments where none of the potential results would imply that one committed a wrong, but I doubt these are the central cases of moral experimentation. A central case, as I see it, involves performing some action to learn whether it is wrong or not.[5] If there's a significant amount to be learned (about the wrongness of the action) from such an experiment, one must think there's a reasonable chance one will learn that the action is wrong (otherwise, one should already be essentially sure that the action is not wrong before one conducts the experiment). Since learning

2. Thanks to Owen Cotton-Barratt for crisply setting out this example to me.

3. Mill does caveat this with the harm principle: "When, by conduct of this sort, a person is led to violate a distinct and assignable obligation to any other person or persons, the case is taken out of the self-regarding class, and becomes amenable to moral disapprobation" (Mill (1859) at 153)

4. An experiment in living might also affect our moral intuitions. But if that is the case, then that specific experiment in living would have been in part comprised by an experiment in morality.

5. As I discuss later, one can set oneself up to later passively 'morally experiment' — that is, set oneself up to be in situations where one tends to make decisions that give information about the moral domain without at that later time intending to. This setup has various attractive properties, especially to the deontic objector, and I discuss it more fully in Section 5.1 below.

that the action is wrong is coupled to the action in fact being wrong, there must thus be a significant chance that, running the experiment, one commits a wrong. For example, after you conduct your experiment involving Alice's dress, it is plausible you conclude that your initial hypothesis was correct — it was *wrong* to tell the truth.

Of course, if the argument in this paper goes through, then running an experiment that results in an intuition of wrongness is often not an all things considered moral wrong: the possible expectation of committing a local wrong is often outweighed by the expectation of gathering morally valuable information. It might be a harm, but it often isn't a wrong. Nevertheless, there is an important sense in which this is a local wrong. To see this, consider the case where this is the last time you'll ever face a similar case to Alice and her dress (you will drop dead as soon as you finish your dress-advising). In that case, you should clearly not experiment, since the information you gather wouldn't be morally valuable (stipulate also that no one else will observe and update their moral beliefs based on what you do). If barring informational value an action is a moral wrong, then there is an important sense that — even though should you agree with this paper the action is not a global wrong — it is a 'local' wrong, and maybe a wronging, depending on the action. To morally experiment, we pay a price in local immorality.

In what follows, I argue that it is often a reasonable price to pay. That is, I want to argue that we should include significant moral experimentation in our moral practice. The argument has several parts. In Section 2, I discuss the epistemic value of empirical moral experimentation (compared to, say, thought experiments). In Section 3, taking the epistemic value of such experiments as given, I argue that such epistemic information is morally valuable. This is sufficient to give a pro tanto reason to engage in moral experimentation. This pro tanto reason, of course, doesn't get us very far, since the cost of such experimentation might be reasonably large and the pro tanto reason might be reasonably weak. So, in the next two sections, I first argue that moral experimentation is a public good (Section 4); and second I take up various objections (Section 5). Finally (in Section 6), I discuss what we should take away from all this — that is, how it should change our moral practice.

Some eventual practical implications include (1) that we should incentivize moral experimentation and risk-taking in new fields, such as gene editing and artificial intelligence; (2) that we should punish moral pioneers relatively lightly, even when they make large mistakes;[6] and (3) we should set up young people, especially those young people who are likely to become future leaders, to have a wide range of moral experience and reward them for moral experimentation.[7] I discuss other practical implications as well, particularly those applying not only to large issues or the legal systems, but also those that mostly should change how we engage in day-to-day morality with ourselves and those close to us.

2 Empirical experiments and moral epistemics

Do actions in the external world really resolve moral — as opposed to simply empirical — uncertainty (assuming the two are distinct)?[8] And could we not get this information in some other way, e.g., through thought experiments? These are the two questions this section attempts to answer.

Answers are plentiful in the literature. Aristotle, for example, argues that a young person is unfit to study ethics because "he lacks experience of the actions in life, which are the subject and premises of our arguments" (Irwin (2000) I.3). Other more recent work (see in particular Railton (2017)) takes a reinforcement learning approach to moral learning — an approach that takes sensory moral experience as one of the central components of moral learning.

6. This might jar with intuitions that the arrogance of these people should itself be disincentivized. As I discuss below, I largely disagree with punishing their arrogance, at least given their moral pioneering.

7. We already do this to some extent, e.g. by punishing young offenders less harshly.

8. Note that it is also a sensible, and separate, pro tanto argument for many of the practical actions I argue for (Section 6) that they provide straightforwardly empirical information that nevertheless helps one evaluate the morality of actions in the future — primarily, information about the consequences of an action.

But very much of moral philosophy depends not on empirical experiments but instead on thought experiments or deductive reasoning.[9] So in this section I argue, in several ways, for the claim that empirical moral experimentation is an epistemic tool that can take us significantly beyond where we'd get without it.[10] These arguments are independent — you need not accept all of them to accept my argument. Some of these arguments work particularly for moral 'experiments' — that is, cases where an agent updates her credence in a moral claim based on the result of a test. And some arguments apply also to moral observations that do not come out of deliberate experiments.

The first of these arguments is, fittingly, an empirical one — taken from the literature on experimental philosophy. Specifically, I discuss some experiments about trolley problems and suggest that these show people behave differently in real cases than in imaginary ones (Section 2.1). The second argument comes from the necessary structure of imaginary cases; I discuss this structure and suggest that necessary features of thought experiments make them depart from empirical experiments in ways that influence how we think about them (Section 2.2). And finally, I briefly argue that gaining significant moral information from experience makes sense under a few metaethical views — including nonrealist ones (Section 2.3). I finish this section by addressing three objections to the claim that empirical moral experiments yield information about the moral realm (Section 2.4). Throughout, I do not argue anywhere for the moral significance of the information we acquire from such experiments — I leave that for Section 3 — but simply that we do get information from moral experiments.

Before I begin, I want to set out some basic starting points. First, I take it as a premise that our intuitions can tell us something about the moral domain. I don't take a position as to where those intuitions come from (experience or inborn reasoning), and I don't make many further claims about these intuitions. I do not, for example, think intuitions are always correct. But I do think of them as a kind of information we have about the moral domain.[11] And I also take it as given that thinking about cases, in some form, imaginary or real, is useful for developing those intuitions. The previous sentence might seem like a more substantive premise, and it is. But arguing for the case method from scratch is outside the scope of this paper, and it has already been done in various ways throughout the literature.[12] My goal in this section is to build on the vast philosophical literature on the importance of thought experiments to make the distinct point that — given the role of thought experiments — empirical experiments are also valuable in their own right.

2.1 Differences between judgements on theoretical and empirical trolley cases

In a series of experiments, experimental psychologists and philosophers have tested peoples' behavior in theoretical, virtual reality (VR), and ostensibly real (conducted with mice) trolley problems. Patil et al. (2014) find that people are significantly more consequentialist in VR setups than in theoretical ones—although the sample-size is quite small. And, in a similar study, Bostyn, Sevenhant, and Roets (2018) find that in a case where subjects believed the trolley setup was real (and subjects could choose whether to shock one mouse or allow five to be shocked), subjects were similarly more likely to turn the proverbial trolley in the 'real' setup than in a theoretical one.

These studies are small; and it is unclear how far on the spectrum from thought experiment to empirical experiment a VR setup would fall. Nevertheless, they suggest that people do not behave

9. For discussions of thought experiments and their centrality, see, e.g., Kagan 2001 and Dancy 2021. For examples of thought experiments at work in moral philosophy, see, among many others, Foot (1967), Thomson (2014), Hare (2016), Harris (1975), Scanlon (1998), Hare (2016), Rawls (1958), Nozick (1974), and Raz (1975).

10. To be clear, by empirical moral experiments I don't mean the experiments often used in 'experimental philosophy' — I don't mean surveying people on their judgments of cases or the like. I rather mean using actual moral issues that come up in life as experiments against which to test moral claims.

11. There is a great deal of literature on the importance, role, and justification of moral intuitions. For example, see McMahan (2000) and Beckstead (2013) at chapter 2.

12. See, e.g., Kagan (2001), Dancy (2021), and Ross (2006).

in thought experiments the same way they would if faced with a similar choice in a non-thought world. No experiments have yet been done, as far as I am aware, of moral learning: do people behave differently in cases where they've faced a similar moral dilemma before? If so, is that behavior on the second moral dilemma different than if they'd faced a similar dilemma but only in a thought experiment? If the answer to that last question is 'yes', then I think that would be sufficient to empirically demonstrate a good bit of my conclusion in this section. But we need not rely solely on empirics: the theoretical structure of thought experiments can also shed light on their limitations.

2.2 The structure of thought experiments

When we engage in a thought experiment, there are some necessary departures from reality. There is always an implicit third option: I can let the trolley take its course, turn the trolley, or stop thinking about the whole thing. If I stop thinking, the trolley does not take its default path: the five are not killed. Instead, I'm let off the hook. This has a few effects.

It means, most obviously, that I am under much less pressure to put in the effort to come to a correct conclusion. If I fail to properly think through what I should do in a moral thought experiment, this might be its own kind of moral failing — a failing of my moral-epistemic responsibility. But it is surely (barring some edge cases) a less serious moral failing than if I — out of laziness or aversion — failed to put in the requisite effort to think through what I should do in an analogous empirical case. So, in short, the pressures for me to have good epistemics are much weaker in thought experiments.

This can also lead to a relatively stronger consideration of the aversive consequences of thinking a certain thing. It might be that I want to reject, say, consequentialism, so I wish I were the sort of person who would not pull the trolley's lever (to switch it to kill the one rather than the five). In a thought experiment, the price I pay for being this kind of person is rather low. In the real case, the price of being this kind of person is actual lives. In short, answering a certain way in a thought experiment does have certain consequences: I show what kind of person I am. Sometimes, for whatever reason, that's costly. Empirical moral experiments make us 'put our money where our mouths are'.

It's certainly true, as an objection to the above, that one could just commit to doing the job right — to thinking fully through a thought experiment and not flinching away from the conclusion. On the margin, one should probably do this; and the point of this paper is not to discount the important role that thought experiments should play in our moral judgements. But it seems unlikely that we can always do this fully, so — given our human limitations — it seems experimental cases may sometimes lead to better epistemics, simply by virtue of the stronger pressures to get them right.

2.3 The moral information one gets from experience under various metaethical views

In the above two subsections, I explicitly compared thought experiments to empirical ones. In this one, I want to provide a somewhat broader argument for there being moral information to be gained from experience — namely, by briefly considering how such information can make sense given a few metaethical views. Note that I do not claim to give anything approaching a full metaethical account of moral experimentation. I would like the claims of this paper to be as independent of one's metaethics as possible — I would like moral experimentation to have a place under many metaethical umbrellas. To provide a full metaethical account of moral experimentation conditional on many metaethical views would be a massive undertaking,[13] and it would require me to make some sense of many metaethical views I do not consider sensible. To provide a full metaethical account of moral experimentation conditional on just my own metaethical view would perhaps not be as Herculean, but, roughly in proportion to its smaller Herculeanity, it would also do less to further

13. This isn't to say that it isn't a worthwhile one. But my goals in this paper are more modest.

the general argument for moral experimentation. So I'll only provide a brief sketch of how moral experimentation makes sense given some metaethical views.

2.3.1 Perception

Most simply, if one believes in moral perception — that is, that one observes moral facts through perception, just as one observes empirical facts — then putting oneself in the position to perceive various new morally-relevant empirical phenomena would in itself give perceptual access to moral information.[14]

For example, if I am fighting in a war and see others die around me, I might perceive something about — say — the value of life. This might be because life is made more salient to me by virtue of it seeming more precarious. But here I can reasonably be seen as learning something about the moral domain by perceiving these wartime deaths. Note here that my learning about the moral domain in this way does not commit me to the position that moral truths are not a priori. In particular, I can learn about a priori truths through empirical means. (To see this, consider how we often learn basic mathematics: with our fingers, with simple examples of shapes around us, etc.)[15]

2.3.2 Intuitionism

I now provide an even briefer (and slightly overlapping) treatment of how moral experimentation might be seen by several variants of ethical intuitionism. The role I ascribe to it in this subsection is very similar to the one in the subsection above.

Broadly, ethical intuitionists take intuitions to be a major source of information about ethical matters. A common intuitionist claim is that it is sound to take the ethical proposition an intuition points to as a (perhaps fallible) proposition in one's ethical reasoning system — or even that intuitions provide the unique kind of starting points for such reasoning. Moral experiments can plausibly be accommodated by a few variants of such views. If one takes there to be a way in which our emotions track the moral truth (Prinz 2006) or if one considers us to have something akin to a special sense that perceives morality (Dorsey 2021), it seems likely that moral experiments would give us moral information about their contents. Or, even if one takes the intuitions that can ground ethical thought to be initial seemings of a more intellectual kind (Huemer 2005), it still seems a priori possible that one would get access to certain new such seemings by having things happen that such new seemings might be about. In fact, as we discussed in the previous subsection, even if one considers ethical reasoning to be purely a priori and one thinks that all the relevant starting points are accessible independently of one's experiences, one might see moral experiments as playing the prosaic role of getting us to reason through certain ethical matters.

2.3.3 Thick concepts

Moral experiments can be useful for considering thin moral facts in several ways, as discussed here. But they might be particularly helpful when considering so-called 'thick' concepts, such as courage or cruelty. Trivially, empirical experiments can help dispel empirical uncertainty about such concepts. But I think experiments can also help dispel moral uncertainty. If, as Putnam (2002), Williams (1985), Kirchin (2010), and others argue, the evaluative and the descriptive in a thick concept are not simply added but are entangled in inextricable ways, then to empirically clarify is likewise to morally clarify (and not just to morally clarify by empirically clarifying). For instance, if one

14. For a nuanced discussion of moral perception and its chief rivals, see McGrath (2018), McGrath (2019), and Audi (2013).

15. Most of the ideas in this paragraph are taken from McGrath (2019). McGrath also gives several other careful and nuanced ways in which we can morally learn from our perceptions, and she takes care to clarify the ways in which this does *not* commit us to the position that moral knowledge itself is not a priori — though she also leaves open the possibility that it might not be.

thinks one's boss is operating immorally and one interrupts them during a meeting to criticize them, one might later be significantly better able to tell whether in doing so one acted courageously or arrogantly (or, of course, more precisely, the degree to which one did each). This holds even if one knows the empirical consequences exactly.[16]

2.3.4 Antirealist versions of the above

In much of the above, I present the arguments as if they refer to some exterior moral domain. So — as explicitly stated — these arguments might only appeal to (certain kinds of) moral realists. But I think parallel arguments can be given to the antirealists: one can reframe each of the above arguments as if it dealt with making better sense of a constructed or subjective moral domain, or as describing desirable properties of one's ethical reasoning.

For example, the first brushes of an anti-realist version of Section 2.3.1 might look as follows. Instead of taking there to be moral properties in the world that one can perceive about acts or outcomes by coming into contact with them, an anti-realist might make the same ethical update in the same situation, but see himself instead as taking a step in some process of figuring out what he cares about by seeing what he cares about in particular situations (Carlsmith 2023), or perhaps (overlappingly) see himself in the process of fitting an ethical curve through one's moral intuitions (Beckstead 2013). But again, providing a full account of moral experimentation within any such position is beyond what I wish to do in this paper.

2.4 Objections to moral experimentation as a source of information

There are a range of objections to the thesis of this paper — that we should perform and reward moral experiments — most of which I address in Section 5 below. In this subsection, I address two objections — objections only to the claim that empirical moral experiments are a useful epistemic tool. I do not address other objections, e.g., of the form that 'while they do provide information, they aren't worth the tradeoffs'.

2.4.1 Correlated errors in moral experiments

One of the main objections runs as follows. Take our moral intuitions as some kind of 'data' about the moral realm.[17] In any experimental setting, if one performs ten experiments — but the experiments are all flawed in precisely the same way — then one should not update as far as if the ten experiments had independent errors and still returned roughly the same results. In short, moral experiments — like scientific experiments — can give us systematically biased results. And this implies we get less information from experimental results than it might at first appear.

This is a good objection. And it does mediate the extent to which we should rely on moral experiments and the intuitions that flow from them. It also suggests we should consider particular ways in which our 'moral data' might be systematically biased and correct for those biases.[18] Certainly my proposal is not to replace deductive moral reasoning with moral experiments up the wazoo. And there is undoubtedly an important point to be made that only some experiments are workable. It might be difficult, for example, to get good intuitions about actions that affect the far future, since good 'data' might not be available until that far future. So there might be areas of the moral domain where experiments are less fruitful, and in these domains we should incentivize moral experimentation less in these settings than in other ones. But these limitations do not undercut the usefulness of experiments in a wide swath of morality.

16. This argument is a distilled version of some of what I take to be between the lines of Murdoch (1970).

17. For a fuller articulation of this setup and objection, see Beckstead (2013), chapter 2. For a broader objection that points at a similar thing, see Singer (2005).

18. Some examples of such biases include, e.g., the observable victim effect.

2.4.2 One could — theoretically — perfectly simulate the experiment

Another reasonable objection asks whether any moral experiment could just be replaced by thinking about the same case as if it were really happening. Perhaps if I were a kind of moral-simulating computer, with near-unbounded compute resources, or if I had access to such a simulation machine, then non-simulated experiments would offer little beyond what would be offered by simulated ones. It's plausible a perfect computer simulation could replace a real empirical experiment.[19] But — setting aside the obvious point that we're still very far from such computing machines — if a simulation were advanced enough to give salient moral information, it's unclear in what morally salient ways experimenting on a simulation would differ from experimenting in the non-simulation world. If a computer, say, simulated a conscious agent, the best simulation would potentially be a computer-originated conscious agent — in which case I (or a version of myself that I am simulating) might reasonably have the same relationship to that agent, in most morally relevant ways, as I'd have towards any other non-computer-originated conscious agent. So the simulation-reality distinction might tend to fade away, at least in the limit.

Relatedly, the limits of human cognition give us an overlapping reason to empirically experiment. If I imagine the consequences — moral and empirical — of some action, I'm reasonably likely to get something wrong.[20] For this reason alone, often empirical experiments can be epistemically useful, especially if we morally care about empirical aspects of our choices. What's more, in thought, I might miss certain morally salient features of a decision situation that reality would remind me of in a real experiment, and I might misapprehend the relative moral salience of different features. For example, I might round off considerations to zero in a thought experiment that I would not so round in reality.

2.4.3 Thought experiments allow us to distill cases to their morally relevant features

The final objection I discuss in this section points out that thought experiments (or deductive moral reasoning) allow us to distill cases to their most relevant moral features. Empirical cases, in contrast, can be convoluted.

Let me begin my reply by clarifying that I don't propose we replace other kinds of moral reasoning with empirical moral experiments — and distilling reality to thought experiments can indeed (I think) improve our moral judgements.[21] Instead, I propose that moral experiments should form an important part of our moral epistemics — but a part working in conjunction with thought experiments and deductive reasoning. While thought experiments seem reasonable devices for, e.g., weighing up the tradeoffs we'd make between different values, empirical moral experiments seem unusually good tools for other parts of the moral-epistemic process. For example, empirical moral experiments seem particularly able to point out what features of a situation are most morally salient. Empirical experiments might call my attention to some previously-unnoticed dimension of the situation — a dimension I might not have included in a thought-experiment distillation. In this way, empirical and thought experiments might complement each other.

Other devices — apart from thought experiments or empirical experiments or even computer simulations — can also be useful. I think, for example, of the role of literature to our moral epistemics.[22] All these tools give us different access points to information about the moral domain. In this paper, I argue for more moral experimentation — but I don't mean to imply that these other tools are not also useful. My only claim is that we get information out of empirical moral experimentation even given these other tools: because it gets us in situations where pressures to make good moral judgements are strong enough, because it provides quick feedback on moral claims,

19. Some have discussed computer simulations in the context of thought experiments or arguments. See, e.g., Sanjay Chandrasekharan (2013), Skaf and Imbert (2013), Saam (2017), and Schulzke (2014).

20. For a more general discussion of the importance of experience to moral learning, see Railton (2017).

21. For a plausible discussion of how this works, see Dancy 2021.

22. For work on the moral-epistemic role of literature, see, e.g., Bal and Veltkamp (2013), Swirski (2007).

because it exposes us to a variety of moral observations packaged in epistemically useful ways (especially given our human cognitive limitations), because it lets us acquire moral representations.

Moreover, I think it might be the case that empirical experiments don't just separately add to other forms of moral learning and knowledge, but also enhance different forms — such that deductive moral reasoning, say, becomes a better tool than it was before if we pair it with empirical experimentation. In short, empirical moral experimentation seems like a valuable part of a reflective moral equilibrium.[23]

3 Morally valuable information

In the previous section, I argue that one can gain moral information. In this section, I argue that moral information is morally valuable and seek to more concretely understand its value. The very basic idea is that moral information guides our future actions — it lets us bring these actions into greater coherence with the correct moral theory.

3.1 Intuitions on the usefulness of moral information

One should, all else equal, try to make morally good choices. If one could, say, pay $100 to make a life of significantly better moral choices, this is probably — I think — a morally good buy. And it seems like some kind of access to the moral domain (as discussed above) is precisely the kind of thing that would make one better at making moral choices.

We make these kinds of trades (giving up something to get additional access to the moral domain) all the time, both on an individual and on a societal level. I might — say — spend some time reading about different moral causes, so I can better prioritize among them. I pay, with my time, for this better prioritization. A different example might come from criminal law. A judge might sentence an offender to community service to rehabilitate that offender by, say, giving her an increased sense of belonging to a community or an increased feeling of responsibility to that community. Here, the judge is forcing an offender to give up something for this access to the moral domain. On a societal level, we encourage sometimes-disruptive protests that might lead to greater societal awareness of various issues, such as racism or climate change. Society is giving something up (smoothly running roads, tranquility, etc.) for this potential moral progress. Examples are plentiful.

So, by and large, we are willing to give up some things to get better moral epistemics. What, formally, is the value of that information? And how can we update on information gained from moral experimentation? These are the questions taken up in the remainder of this section.

3.2 A specific example, and how not to make sense of the value of moral information

Before I describe a formalism that appropriately captures the value of moral information, I provide a specific example — the Butchery Case — to have in mind, and a formalism that I think handles this example poorly. I hope this gives some sense of the space of available formalisms.[24]

3.2.1 The Butchery Case

Suppose I am uncertain between the following two moral theories.

Theory 1: a version of utilitarianism that cares somewhat about non-human animals

Theory 2: a version of utilitarianism that only cares about humans

23. For more on the aspects of reflective equilibrium, and on how moral perception fits into this picture, see McGrath 2019 at chapter 2.

24. The mathematical shape of this example is taken from Russell (forthcoming) and Podgorski (2020).

More precisely, suppose I assign 50% probability to each being correct. And suppose I face a choice between the following two options:

Random: A fair coin is flipped. If it lands on heads, I will be vegan for the rest of my life. If it lands on tails, I will be an omnivore for the rest of my life.

Class: I pay $100 to attend a class that involves butchering and cooking a chicken. Suppose I know ahead of time that I will learn which of Theory 1 and Theory 2 is correct as a result of this experience, each with probability 50% (since I think each has 50% probability of being correct, and I think I will learn which one is correct). I accordingly proceed to either become vegan or not.

Suppose that Theory 1 says that in each option, a life as a vegan has utility 1 000 000 and a life as an omnivore has utility 2 000 000, and that Theory 2 says that in each option, a life as a vegan has utility 1 000 000 and a life as an omnivore has utility 0. Suppose both theories think that beyond that, outcomes where one pays $100 have their utility decreased by 100.

3.2.2 A poor argument for picking Random

If I take the two theories to simply be saying that each kind of life has a certain utility and I treat the uncertainty in each case as simply a usual 50% chance of each kind of life, then the resulting expected utilities the options lead to according to each of the two theories are as given in Table 1.

	Theory 1	Theory 2
Random	1 500 000	500 000
Class	1 499 900	499 900

Table 1: Expected utilities of the two options according to the two theories

In particular, in this sense, both theories say that Random is better than Class. One might then think that any reasonable aggregation rule would then say that I should pick Random. For instance, the expected choiceworthiness (see Lockhart (2000) and MacAskill, Bykvist, and Ord (2020)) of Random is 750 000, whereas that of Class is 749 900.

The issue is that, intuitively, Class is much better than Random: for a price of 100 according to each moral theory — minuscule compared to the stakes of 1 000 000 at play, again, according to each moral theory — I could find out the moral truth on the matter and lead a life that accords with it, instead of choosing at random. To make this more precise: conditional on Theory 1 being correct, Random has expected utility 1 500 000 and Class has expected utility 1 999 900; also, conditional on Theory 2 being correct, Random has expected utility 500 000 and Class has expected utility 999 900. Either way, the expected utility of Class is higher.[25] I should choose Class.

3.3 Decision-making frameworks that prefer morally informative events

In this subsection, I first describe a formalism from Podgorski (2020) and Russell (forthcoming) and also essentially from Sepielli (2009) that makes sense of the value of moral information, which includes a decision rule I call Maximizing Expected Objective Value — MEOV (Section 3.3.1). I then describe a small extension of this formalism providing a consequentialist grounding of MEOV, mentioning an issue one runs into when attempting to ground it in preferences alone (Section 3.3.2), as well as an alternative framing from Hänni and Popper (2023) that takes the perspective of each moral theory on what outcomes actions will lead to as an alternative grounding (Section 3.3.3). I

25. The theories even happen to 'agree' that the value of Class is higher than that of Random by 499 900. The better formalism below will say that this is exactly the relevant number to calculate to compare the two options — it is the difference between their expected true utilities.

finish by, following Russell (forthcoming), describing how this formalism sees the Butchery Case and stating a Pareto condition that lends some support to MEOV (Section 3.3.4). So, the point of this section is to formally describe the value of moral information. If this seems intuitive to you and you don't want to see it formally described, then you may wish to skip to 3.4 below.

3.3.1 Maximizing expected objective value

For the first formalism, following Russell (forthcoming) (and Savage (1954)), let $\mathcal{O}$ be a set of outcomes (think of an outcome as specifying everything that comes to matter), let $\mathcal{S}$ be a set of states (think of each state as specifying the values of all background variables relevant to determining the outcome of a decision conditional on each choice), and let $\mathcal{A}$ be the set of actions — each $a \in \mathcal{A}$ is a map of states to outcomes, $a\colon \mathcal{S} \to \mathcal{O}$, or possibly a map of states to probability distributions over outcomes, $a\colon \mathcal{S} \to \Delta(\mathcal{O})$.

Let's say that I'm uncertain between n competing moral theories, $t_1, \ldots, t_n$. Think of $\mathcal{S}$, $\mathcal{O}$, $\mathcal{A}$ as capturing the decision problem that I see myself as facing. In fact, assume that each state $s \in \mathcal{S}$ determines not just all physical data relevant to determining the consequences of a decision, but also which of $t_1, t_2 \ldots, t_n$ is the correct moral theory, so we can accordingly partition $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2 \cup \cdots \cup \mathcal{S}_n$. Even if all worlds that are (in some sense that I do not attempt to make precise) objectively possible have the same correct moral theory, as we think of this setup as capturing my perspective of the decision problem I face and given that I think each of $t_1, \ldots, t_n$ might be right, each S_i is non-empty.

We think of each t_i as (minimally) specifying a value function v_i that, given any background probability distribution μ on $\mathcal{S}$, tells us what the value of each act $a \in \mathcal{A}$ is — we denote this $v_i(a|\mu) \in \mathbb{R}$. If you are concerned about this requiring theories to have opinions about the value of actions when they are wrong, don't be — for the story below, we in fact only need each theory t_i to specify the value of any act conditional on any background probability distribution that is 0 outside of S_i. Russell (forthcoming), Podgorski (2020) and more or less also Sepielli (2009) then propose the following rule for evaluating the choiceworthiness of acts:

$$v(a|\mu) = \sum_{i=1}^{n} \mu(S_i) v_i(a|\mu_{S_i}),$$

where μ is the probability distribution on $\mathcal{S}$ that I have, we assume $\mu(S_i) > 0$ for all i (otherwise t_i would not be a live moral theory for me), and we use μ_{S_i} to denote the probability distribution on S_i that results from 'conditioning μ on S_i', i.e., for each $s \in S_i$, defining $\mu_{S_i}(s) = \frac{\mu(s)}{\mu(S_i)}$. And of course, the corresponding decision rule is then to pick the act which has the highest choiceworthiness (or, more precisely, one such act — there might be ties). Following Sepielli (2009), let's call $v(a|\mu)$ the Expected Objective Value (EOV), and let's call this decision rule Maximizing Expected Objective Value (MEOV). The reason for these names is that, since we take each $s \in S$ to specify (among other things) the correct moral theory to be some t_i, the number $v(a|\mu)$ defined above is the expectation when the state $s \in \mathcal{S}$ is taken at random from μ of the value of the action a according to the moral theory that is correct.

3.3.2 Maximizing the expected objective utility of the outcome endorses MEOV

One (indeed, perhaps a central) way these functions $v_i(\,\cdot\,|\mu)\colon \mathcal{A} \to \mathbb{R}$ could come about is if each theory t_i assigns a utility to each outcome in $\mathcal{O}$ — i.e., there's an underlying function $v_i\colon \mathcal{O} \to \mathbb{R}$.[26][27] That is, we'd let $v_i(a|\mu)$ be the expectation of $v_i(o)$ where o is obtained by drawing a state s from μ and then drawing an outcome o from the distribution induced by the action a in the state s. And

26. Or really, one only needs to include those outcomes in the domain of v_i that 'are coherent with t_i being correct', i.e., which have nonzero probability given some choice of a state $s \in S_i$ together with some action $a \in \mathcal{A}$.

27. I've abused notation here by letting v_i denote both the value function on actions and also the underlying utility function that cares about outcomes.

one (again, perhaps central) way the functions $v_i\colon \mathcal{O} \to \mathbb{R}$ could in turn come about is if each t_i has preferences over distributions on $\mathcal{O}$ which satisfy certain canonical rationality conditions (namely, von Neumann and Morgenstern's), so its preferences between distributions must be as if it were maximizing the expectation of some function $v_i\colon \mathcal{O} \to \mathbb{R}$ (von Neumann and Morgenstern 1947).[28] Though unfortunately, this only pins down v_i up to rescaling by any positive constant (and shifting by another constant, but the shift is unproblematic as, fixing μ, whether a decision is endorsed by the decision rule above does not depend on this shift). This is a standard issue — see, e.g. Sepielli (2009) and Demski (2020).

Given this set of assumptions, $v(a|\mu)$ is the expectation of the utility of the outcome according to the true moral theory. That is, $v(a|\mu)$ is the expectation of the utility of the sampled outcome according to the theory that is correct in the sampled state. In this sense (and given our further assumptions), MEOV is equivalent to optimizing the expected true utility of the outcome.[29]

3.3.3 MEOV from the point of view of the moral theories

Instead of thinking of the decision-maker as directly having a distribution μ over states that specify all physical and moral background information relevant to a decision, Hänni and Popper (2023), following the mathematical formalism of Desai, Critch, and Russell (2018), take the moral theories themselves to have differing views about how things will play out (in the simplest case, what the distribution over outcomes is) if I take a certain action. In particular, if I take an action from which I can learn morally, each theory t_i expects me to update toward that theory t_i. In this setup, the natural version of maximizing expected choiceworthiness (Cotton-Barratt and Greaves 2023) is to pick the action which maximizes the sum over theories of the credence one has in each theory times the expected value that theory assigns to the action (with the expected value calculated according to its own credences about morally informative events).

We can transition between this picture and the one presented previously, where a decision-maker had a joint probability distribution on moral theories and physical states, by taking μ_{S_i} to capture what theory i takes the mapping from actions to outcomes to be. The decision rule specified in the previous paragraph then precisely turns into MEOV.[30]

28. And again, for theory t_i, one really only needs to include those outcomes which have nonzero probability given some choice of a state $s \in S_i$ together with some action $a \in \mathcal{A}$.

29. In this version, the random variable whose expectation $v(a|\mu)$ is is defined on $S \times \mathcal{O}$ (or, if you like, just the subset of those state-outcome pairs that have a nonzero probability). One can change the formalism slightly to make it depend on the outcome alone. To do so, further assume (perhaps by modifying what kind of data one takes an outcome to specify) that each outcome $o \in \mathcal{O}$ has a unique moral theory compatible with it (i.e., for each $o \in \mathcal{O}$, there is a unique i such that there is some $s \in S_i$ and $a \in \mathcal{A}$ such that the distribution on outcomes given by being in state s and taking action a assigns nonzero probability to o). This lets us sensibly talk about the moral theory which is true in a particular outcome, and it makes the EOV $v(a|\mu)$ just the expected true utility of the outcome if we sample a state from μ and then an outcome from the distribution given by taking the action a in that state. I.e., $v(a|\mu)$ is the expectation of the utility of the outcome according to the theory that is correct in that outcome. In this sense (and given our further assumptions), MEOV is equivalent to optimizing the expected true utility of the outcome 'on the outcome's own terms'.

30. In fact, given a sequential decision-making setup, Hänni and Popper (2023) argue for a decision-making process of a specific form — in particular, with a specific way of updating upon morally informative events. Namely, they argue that any decision-making process which is Pareto optimal with respect to the theories has the following form: start with a particular set of schmedences in the theories; treat the theories as hypotheses that sometimes (in morally informative cases) make predictions about the world, updating your schmedences in these hypotheses according to Bayes' rule; and on each step, decide according to MEOV with the schmedences you have at that step. However, they provide no argument that these schmedences have to match one's credences. (Indeed, if there is nothing that fixes the scaling factors on the theories' utility functions, then if one's overall decisions ought to be anything in particular, there could not be anything which would generally (i.e., for any scaling factors on utilities) cause the schmedences to match one's credences, because this would lead to different decisions for different scaling factors. See Fallenstein and Stiennon (2014) for more on the entanglement between the credences and utility scaling factors assigned to theories in such an aggregation rule.)

3.3.4 Returning to the Butchery Case: MEOV and Pareto optimality

MEOV endorses Class in the Butchery Case. The numbers $v_i(a|\mu_{S_i})$ for the two theories i and the two actions are as given in Table 2.

	Theory 1	Theory 2
Random	1 500 000	500 000
Class	1 999 900	999 900

Table 2: The expected objective values of the two options conditional on each of the two theories being correct

The EOV of Random is thus $0.5 \cdot 1\,500\,000 + 0.5 \cdot 500\,000 = 1\,000\,000$ and the EOV of Class is $0.5 \cdot 1\,999\,900 + 0.5 \cdot 999\,900 = 149900$. So MEOV says I should pick Class — the intuitively right result.

In fact, as (Russell, forthcoming) notes, any decision rule that satisfies a Pareto condition — namely, that if conditioning on any theory being right, an option would be better than another option, then the option is all-things-considered preferred over another option — would choose Class. In fact, such a Pareto condition can be translated into a constraint on one's decision rule: that it is from a particular family of decision rules that includes MEOV. See Russell (forthcoming) or Hänni and Popper (2023) for details of arguments for two claims of this kind.

3.4 Exploration vs. exploitation

To show this formalism in some more action, as well as to present an important consideration for assessing the value of a given moral experiment, we now use it to make sense of the tradeoff between exploration and exploitation. The idea is simple: as the information one gains from a moral experiment is often distributed across the future at some particular density, performing a moral experiment early on is often better than performing it later. In other words, one should often be willing to pay more (in local morality) for moral experiments performed earlier — one should be less willing to sacrifice one's local morality for the same amount of moral information when one has fewer relevant future decisions left to make. The balance between exploration and exploitation should slide toward exploitation as time goes on. To illustrate this principle, we consider the T-Delayed Butchery Case.

3.4.1 The T-Delayed Butchery Case

As in the Butchery Case, I assign 50% credence to each of the following moral theories.

Theory 1: a version of utilitarianism that cares somewhat about non-human animals

Theory 2: a version of utilitarianism that only cares about humans

The options I now face a choice between are just variations of the options in our original Butchery Case where my actions start to correlate with the moral truth at time T. Suppose that there are 10000 days in my life. On each day, a fair coin is flipped. By default, on each day that the coin lands on heads, I am an omnivore; on each day that it lands on tails, I am a vegan. On day T, I am going to be offered a chance to pay x to attend a class that involves butchering and cooking a chicken, and, as before, I know ahead of time that I will learn which of Theory 1 and Theory 2 is correct as a result, each with probability 50% (since I think each has 50% probability of being correct, and I think I will learn which one is correct). If I attend the class, I accordingly proceed to be vegan or to be an omnivore on days $T+1, T+2, \ldots, 10000$ (otherwise, I proceed to be guided by the coin flips).

Suppose for simplicity that both theories agree that utility is additive over days and that paying x decreases utility by x. Theory 1 says that a day as a vegan has utility 100, and a day as an omnivore

has utility 0. Theory 2 says that a day as a vegan has utility 100, and a day as an omnivore has utility 200. Given these assumptions, what should my willingness to pay for the class be?

Well, according to Theory 1, ignoring the cost of the class, the difference between the utility of a life in which I attend the class and a life in which I do not is $100(10000 - T)$. Theory 2 agrees. So, still ignoring the price of the class, the expected objective value of a life wherein I attend the class is larger by $100(10000 - T)$. So MEOV says I should pay for the class (and attend it) if and only if $x < 100(10000 - T)$. I should be willing to pay more if T is smaller — getting moral information earlier is more valuable. And, to look at the same idea in a different way, note that if we take the cost of the class, x, as fixed, there is a cutoff in the time T — namely when it becomes larger than $10000 - \frac{x}{100}$ — after which I should no longer be willing to pay x to attend.

3.5 Moral information the formalism fails to capture (and the value of that information)

An ideal framework that makes sense of the value of the moral information gathered from moral experimentation (above that gathered from thought experiments or other moral-epistemic tools) would fully capture the value of such information. But, assuming the framework above captures all the value there is to be gained from moral experiments, I think there is a reasonable objection (which I present below) to the claim that they get us farther than thought experiments. As I think there is in fact more moral information to be gained from moral experiments than from thought alone, I see this as a reasonable objection to the above framework capturing all the moral information one gets from moral experiments.

The objection runs as follows. Often, if I think about a moral experiment before carrying it out sufficiently clearly to have a clear picture of the set of possible updates I would make and of more or less precisely when I would make each of these updates (which seems necessary for the moral experiment I am about to run to be well-described by the framework above), then I can go ahead and make the updates already without having to carry out the experiment. In other words, often, I can make the same updates from thought alone.

One response to this begins by noting that this worry is similar to a worry one might have about updating on matters of physical fact — it is often also the case that if one could just understand the experimental setup and all the possible conclusions clearly enough, then one could immediately see the correct answer. It seems likely that one could have convinced oneself that special relativity must be correct from the totality of empirical observations prior to the Michelson–Morley experiment — indeed, plausibly, for someone that could clearly compare the alternatives, the case for special relativity over classical mechanics would be overwhelming. And yet, it is significantly easier to get a sense for things after running the Michelson-Morley experiment. Plausibly, something similar is true in the ethical case.

A way to make sense of this while maintaining that the formal framework captures something important about the value of moral information that one gets from moral experiments is to take the formal framework as, instead of capturing the ex ante position as presented above, capturing something more like an ex post view on what was a correct way to see the moral problem. One often sees the options that ought to have been live, and which uncertainties ought to have been resolved, more clearly after performing an experiment. For instance, if one is prompted by the initially perplexing result of the Michelson-Morley experiment to figure out special relativity, one can come to later see the result of the experiment as implying one should update toward special relativity compared to some reasonable hypothetical ex ante position. Crucially, I claim that it is not necessary to ever have in fact occupied this experimental position.

That said, I think there are kinds of moral information one acquires from moral experiments that the formal framework above plausibly fails to capture, setting aside the above objection. In particular, I think the framework fails to capture how moral experiments let one do moral 'representation learning': learning how to appropriately represent a situation in terms of specific concepts within

the moral domain. Before I learn how to trade off between moral considerations, I must learn how to see moral problems in terms of reasonable considerations in the first place. And even if I come to see a past moral experiment in terms of reasonable moral considerations and can then come to consider it informative about some moral dilemma (as described in the previous paragraphs), the information it provides about this moral dilemma is not all of the information it provides — it also teaches me to see the moral dilemma in terms of various moral considerations in the first place. In fact, it seems plausible that in thought experiments, one is likely to set up situations that appear neat in terms of one's pre-existing system of moral concepts. This is an obstacle to easily facing thought experiments that make one see moral problems in different terms. So the formal framework above fails to capture an important kind of moral information — and, in fact, a kind of information that might be particularly hard to learn from thought experiments alone: information about the moral terms in which to perceive a situation.

4 Moral experimentation as a public good

Much like individuals can learn from a moral experiment, so can groups of humans: families, friend groups, broader moral-epistemic communities, humanity as a whole.[31] In the above sections, I argue that moral experimentation is a good for the experimenter. Here, I argue moral experimentation is also a public good. If moral experimentation is a public good, this has important implications — namely, that it will be naturally undersupplied, so we (qua societies or qua individuals acting for the common good) should take extra actions to encourage it.

4.1 Public goods: a quick summary

A good is standardly considered a public good if and only if it meets two conditions: that it is 'non-rivalrous' and that it is 'non-excludable'. I discuss each in turn.

A good is non-rivalrous if and only if one person's consumption of that good does not deplete it (Samuelson 1954). That is, if I use it, my usage doesn't make you any less able to use it: the marginal cost of additional usage is 0. For example, the contents of a book is a public good: my reading it doesn't make you any less able to read it. In contrast, a physical book—that is, the paper object that contains the contents of the book—is rivalrous. Either I can use it, or you can use it, but we can't typically both use it together (at least not at one time).

A good is non-excludable if and only if it is impossible to prevent some individual from having the good (Musgrave 1959). For example, let's say a certain community has no rules about which community members can or cannot fish in a local lake. Here, the fish supply is a non-excludable good. If instead community members were required to buy fishing licenses, the fish-stock would be excludable.

Certain goods are rivalrous but non-excludable, and certain goods are excludable but non-rivalrous; but we are concerned with the set of goods that are both non-rivalrous and non-excludable. This set of goods are public goods, and they have the interesting property that they are under incentivized by normal market pressures.[32]

31. Learning from others' experiments might at first seem worryingly like moral deference. I might defer to what you say your new intuitions are, and this might be epistemologically troubling. Against this, there are two natural rejoinders. First, what deference there is here seems to be 'impure' moral deference — much as I might defer to an animal psychologist on the moral rightness of leaving a dog alone in the house. (For more on impure moral deference, see McGrath (2009).) Second, it might not be that I morally defer at all. It could instead be that I simply have — because of your experiment — a wider set of moral experience from which to build intuitions.

32. This is a well-known result explored in many works of economics, as early as Smith (1976 (1776)) book 5, chapter 1 and Mill (1963) at 968. Essentially, because of the positive externalities at play, the provision of public goods ends up looking like a k-player version of the prisoners' dilemma. For a full explanation, see Varian (1992) at 416. For the purposes of this paper, it suffices to understand that public goods will systematically be underincentivized and therefore undersupplied by normal market forces.

4.2 Moral experiments: non-rivalrous and non-excludable

In this subsection, I discuss how moral experiments fit into the public-goods framework set out above.

Moral experimentation, prima facie, can be considered a part of research more broadly. And research is generally considered to be a public good (Stiglitz (1999) and Romer (1986)). It is non-rivalrous once produced, in that my noting a theory you derived and inventing some product relying on that theory does not at all make the theory less available to others. And it is — at least in some important cases — non-excludable. In some cases, intellectual property rights mean that the benefits of some kinds of research are excludable (if I patent my chemical discovery, you can't independently profit from it for some years). But other forms of knowledge are not so excludable. It is not usually excludable that a patent was made, for example, and that alone can be useful information (Stiglitz (1999)). So, research is generally considered *impurely* non-excludable.

Moral experimentation seems to fit this mold well. It is non-rivalrous in the same way that any method of experimentation is generally non-rivalrous. If I engage in a moral experiment, you — a passive observer — may not learn as much as I do (there seems to be something subjective about designing and learning from one's own experimentation). But it seems that you would still benefit, from observing more portions of the moral domain by virtue of my experiment, by hearing of my intuitions from me, or by constructing your own less costly experiments derivative off mine (e.g. if she does X, my intuitions will be Y; tested if I do X), etc. Note that — while Mill (1859) advocates for experiments in living on grounds of liberty, he also justifies them by recourse to the public's ability to learn from observing such experiments. Experiments in morality, on this axis, are similar. And while I can imagine cases of excludable moral experimentation (e.g., if I kept my experiments and learning secret), these do not seem to me to be central cases.

So moral experimentation is non-rivalrous and non-excludable. When I engage in a moral experiment, my action has positive externalities: those around me benefit in ways the incentive landscape doesn't account for. You can morally learn from my experiments. So, as with other public goods, we should consider possible interventions to align the incentive landscape with social welfare.

4.3 Implication: society should subsidize

In the case of other public goods, such as public parks or national defense, governments often explicitly step in to fund the goods. Some research is similar. But other ways governments encourage the provision of public goods are less obvious. Patents are one common way. But neither direct funding to moral experimenters (what would that look like?) nor 'patents' on moral intuitions seem like the right approach here.

But just as — in Section 3 above I discuss how we pay a price (in money, time, or something else) for moral knowledge — here we can reassign some of that price from the experimenter to society at large. Most obviously, this means lightening punishments for moral transgressions when they were made in service of a moral experiment.[33] It also means providing the venues for non-costly experimentation — say, sponsoring clubs where people can run experiments in the morality of leading communities, or sponsoring ideological movements that might run more dramatic moral experiments.

Finally, we as individuals can play a role here. It might be that each of us should locally incentivize moral experimentation in those around us, by being forgiving to our friends, by encouraging them to try things out, and by helping them find new venues to morally explore.

33. There are plenty of issues here. For one, it seems like a somewhat unintuitive defense — at least unintuitive to enshrine. For another, the value of moral experimentation varies dramatically depending on various other factors (e.g. how much information we're likely to get, how valuable that information is, whether there are other negative externalities at play — such as the effects of normalizing a kind of moral wrong by experimenting with it, particularly if the experimenter is in a position of authority or admiration), so experimentation should be more or less incentivized in different areas. And, these various other factors might be hard for an incentivizing body to assess.

5 Objections

In this section, I discuss what I see as the most important objections to incorporating a significant amount of moral experimentation into our moral practice.

5.1 The deontic

Most central cases of moral experimentation mean that we increase the likelihood of acting *locally* immorally during that experiment for the sake of better future moral conduct. The natural objection to this runs that — no matter how the tradeoffs shake out — it's wrong to maximize my expected long-term morality in this way.

In many cases, doing so means we victimize someone locally for the sake of others in the future. My own morality is the proverbial trolley going down the track towards my many victims in the future, and my moral experiment switches the trolley onto a track with someone else — more proximate, at least in time, to me — on it.[34] It might be wrong to do this.

One objection to switching the trolley — that doing harm, as I do when I switch it onto a different track, is different from allowing harm, as I do when I simply let it speed on ahead — doesn't prima facie apply here. Either I *do* harm now (at least in expectation), or I allow myself to *do* harm in the future (at least in expectation) by not doing harm in the present. Either way, I do harm — it is just the time (now or later) and the expected magnitude (less now or more later) in question.[35] If one thinks that — in some relevant set of cases, even if not across the board — it is wrong to do harm at time t instead of allowing at time t oneself to do harm at time $t+1, t+2, \ldots t+n$, then one might reasonably prima facie reject all of moral experimentation (although I below propose one more rejoinder to this position). But it seems to me that allowing oneself to do harm is often worse than simply doing harm.

For seminal discussions of 'allowing oneself to do harm', see Hanna (2015a), Hanna (2015b), and Persson (2013). These authors discuss the distinction only when the harm that one allows oneself to do was originally committed in the past, e.g., a poisoner drops poison in a teacup and doesn't intervene before a victim drinks, thus currently allowing (by not intervening) her past behaviour (putting poison in the cup) to result in harm. The cases I discuss are relevantly different because the harm that one now allows oneself to do would also happen in the future. It is as though the poisoner could have, before putting poison in the cup, warned the victim not to drink from any cup offered to her. Hanna (2015a) and others argue that these cases show that deontology has a kind of present-tense perspective. That, while harm is always bad, one might take a present perspective on the whole thing — caring most about what one does now. While this seems plausible in the case when harm was committed in the past and is allowed now, the temporal considerations are less clear-cut when the harm that is allowed would be done in the future.[36]

Another rejoinder to the above deontic objection is that one could — instead of actively morally experimenting — choose to put oneself in a position where one is likely to morally experiment passively: to be in morally novel situations (where plausibly one is more likely to make wrong choices than otherwise) where experiments are constructed naturally without one's locally constructive influence. I might, for example, take on a leadership role in a local club with the explicit goal of putting myself in morally new situations (maybe to prepare for some potentially more important

34. Given the direct causal power I have here, other variants of the trolley problem might be more analogous, e.g., the version where I must push a fat man onto the track to stop the trolley (Thomson 2014).

35. Of course, one could morally experiment in a high-stakes case for the sake of some future low-stakes cases, but this is a case, I think, of moral experimentation that we shouldn't do: that is, it isn't a counterexample to this point because it is just an inappropriate application of the idea.

36. My use of doing harm in the 'future' in this paragraph might be confusing. Trivially, in the first poisoner case, the would-be victim wouldn't be poisoned until the future, so in some sense the harm isn't actualized until the future in either case. The relevant thing here is that the active *doing* was in the past, pre-allowing, in one case, and the active *doing* was in the future, post-allowing, in the other.

leadership role in future, where I'd need to make good moral calls). Here, I do not ever actively choose to make (in expectation) any single worse moral decision. But I do actively put myself in a position where I'm later more likely to make worse moral calls.

The deontic response might then be that the wrong I've done is simply on the next level (one level broader than local): I ought not to put myself in positions where I'm more likely to do these things unintentionally. But one cannot opt out of allowing oneself to do harm in the future — that is the whole motivation for moral experiments. Let's say the deontic objection is that I cannot, at T, allow myself to do harm at $T+1$ just to stop myself from doing harm at $T+2$ (harm that would have been prevented by moral experimentation at $T+1$). But a moral experimentalist might reply that the deontic objection just advocates at T allowing oneself to do harm at $T+2$ to stop oneself from doing harm at $T+1$. That is, since all one does in this version of moral experimentation actively is put oneself in positions where later one will passively experiment, then the deontic objection just prioritizes $T+1$ over $T+2$, and since the present-tense perspective only justifies the prioritization of T, this is uncompelling.[37]

A final rejoinder to the above deontic arguments could run that we can get some of the benefits of moral experimentation without any active efforts to experiment at all. We can, first, get some moral data from regular moral experience, and that experience — even if not designed to test claims — can nevertheless also serve as (admittedly often inferior) moral experiments. Similarly, we can learn from the moral acts of others. Moral intuitions gained from others' choices can allow us to get a wider range of data about the moral domain than we could get if we were constrained only to act at every local point in the most moral way we could. I do think there is some value in this kind of learning. But I also think we give up a lot of the value of experiments if we only do this. Part of the unique epistemic value, I think, is given up if we give up moral exploration.

A separate deontic thought might run that if I do harm now to prevent myself from potentially doing harm in the future, I've in some sense destroyed the potential worlds where I never do harm. And it is these worlds that I should most act to instantiate. There is something aesthetic about this picture. It would be nice, I think, not to have to embroil oneself in immorality just to later maximize one's morality. I think of puritanical young people I knew who deliberately kept themselves naïve; and I think of their more worldly contemporaries who acted (sometimes immorally) to shelter them and protect that pure aesthetic. There is a kind of unblemished morality that moral experimentation almost fundamentally rejects. But I think that rejection is often right. Optimizing for never doing harm, would — if not, under most assumptions, leading to total paralysis, as Mogensen and MacAskill (2021) argue — certainly lead to a minimization of kinds of actions that we usually consider moral or at least part of the 'good life'.

5.2 We already do this enough: this is all well and good, but we have just about enough of this without more on the margins

This objection mirrors one of those given at the end of Section 2 above. This one is the slightly broader version of the one made there. Above, I set out the objection that moral experimentation (or moral experience more broadly) might lead one to have systematically wrong moral judgements. Recall that we can think of moral intuitions as data about the moral domain — data we can acquire more of through moral experimentation, but data that might be systematically biased. Crucially, our principles-based deductive moral reasoning (another kind of access to the moral domain) might be less biased. So, if I experiment more, and base my moral judgements more on experience and less on non-experiential reasoning, my moral beliefs might get worse. So goes the objection discussed above.[38]

37. Thanks to Catherine Brewer for pushing these objections.

38. This objection was most famously made by Singer 2005. A version made more in the framework of this paper is in chapter 2 of Beckstead 2013.

The crucial difference here is that — if the above objection goes through, even in part — then it lessens the societal obligation we have to enable and encourage moral experimentation.

My only additional rejoinder to this is that we ought to carefully consider justifications for why our moral judgements would be systematically in error. It seems that certain evolutionary arguments would convince many of us to trust experiments that test claims related to genetic continuation less. And this is a large part of the moral domain. We might not want to trust (or trust to the same degree) intuitions related to the importance of family, or those close to us in space or time. But it seems that these limitations of experiments do not undercut many of the most important applications of moral experimentation, including in new scientific fields and within small groups of those close to us (e.g. families or groups of close friends). In these settings, such biases we might have reflectively come to view as suspect seem less strong; and where they do come up, we can hopefully address them sufficiently using other moral-epistemic tools.

So, society (and us as individuals incentivizing other individuals) should carefully consider what the underlying epistemic landscape of the particular part of the moral domain looks like. We should — and this seems a somewhat thin answer — incentivize experimentation where it is likely to be a good epistemic tool.

5.3 The marginal cost outweighs the marginal benefit

Let's say that moral experimentation can get us moral information we couldn't easily otherwise get, and that such information is morally valuable. It does not follow from this that most or every experiment should be done. It might be the case — as this objection argues — that the marginal benefits of each individual experiment are simply outweighed by the costs.

The information gathered from each experiment might be quite small. It's not obvious, say, that lying to your friend Alice about her dress (as I discussed at the very beginning) gives you much generalizable information. It might give you some information about how morally to act with Alice's fashion dilemmas, but maybe not even with Bridget's — let alone in other cases of lying and truth-telling unrelated to fashion or friends. And, at the same time, each experiment has an associated cost. Maybe what information one can get is not worth that cost. It might be that to get the real benefits of moral experimentation, one needs to do very many experiments, each with a cost. Maybe it is only after sufficiently many experiments that the information begins to generalize properly and be significant enough to update one away from one's moral priors.[39] Against this objection, I have two points.

First, one could potentially construct an experiment (or series of experiments) to test or refine specific claims. If addressing specific claims, the chance of useless information — information one might get from an experiment but never again use — is lower. This raises the expected marginal benefit of an experiment, since it raises the chance that information obtained is later useful. Think of moral experiments less as going around committing random acts of immorality, and more as deliberate practice involving surgical precision in one's choice of which ways to risk being locally immoral.

Second, it may not be the case that only a vast number of experiments can help us improve our moral epistemics. In particular, we already encounter a vast array of moral data in our day-to-day lives — so the moral experimentation we consider engaging in would start from an existing model of the moral world. For this response, let me consider two ways in which moral experimentation plausibly contributes to my moral understanding: (1) it lets me better understand how to trade off between various values or reasons that I already grasp, especially in somewhat novel situations, and (2) it teaches me to notice new moral concepts and considerations.

In their first capacity, moral experiments 'fine-tune' a relatively small number of parameters (namely, those that specify how to trade off between values), and we might expect that one does

39. Consider the case of training an advanced artificial intelligence. Such training takes enormous quantities of data before the AI begins to generalize appropriately.

not need a huge amount of data to determine the values of a small number of parameters — in particular, much less than to perform moral learning from scratch. To make an analogy to artificial intelligence, it takes very much information to train a base AI model to do well on a range of tasks. But it takes comparatively little additional information to drastically improve a well-trained model's performance on a somewhat novel task — to fine-tune the model (Devlin et al. (2018) and Radford et al. (2018)) — perhaps because one needs to just select appropriate behaviors already 'contained in the model', or to combine existing high-level concepts in the right way. It is of course unclear how much AI analogies apply to humans. But it seems plausible that we too are trained naturally on a wide array of data (from natural observations, thoughts, and even genetics) and the main role of moral experimentation would be a fine-tuning one. Here, the marginal benefit of each additional moral experiment is in fact quite high.

In their second capacity, moral experiments make new aspects of the moral domain salient to us. Even if a single experiment is not sufficient to significantly update my belief about some concrete moral proposition, that single experiment might be plenty to make me see some new part of the moral domain as salient — to grasp a new moral concept or consideration, or to think of a particular situation in a novel way. If we were to continue with the artificial intelligence analogy, we might think that learning representations would in fact be worryingly ineffective — unsupervised representation learning is in fact the data-intensive step of many modern (as of 2023) machine learning pipelines. However, humans learn novel concepts from very few examples (Lake, Salakhutdinov, and Tenenbaum 2015). It seems plausible that the human ability to learn concepts quickly extends to the moral case.

So, although some experiments certainly should not be done because the local moral costs outweigh the future moral benefits from moral information gained, I do not think that a meaningful amount of moral information is so scarcely found in even well-selected and well-designed moral experiments as to imply that one should only very rarely engage in moral experimentation in general.

6 Moral experimentation in practice

Roughly, the thrust of this paper is that we should morally experiment in some situations. This section is a small sketch of where in our moral practice these points seem most obviously applicable.

I begin with something we — at least to an extent — already do: punish young offenders (who, on average, should be morally experimenting more) less harshly than older ones. In most penal systems, young people who break the law are already dealt with more leniently than their older equivalents. There have been several philosophical justifications given for this difference in treatment (e.g., Brink (2020), Yaffe (2018)), none of which are explicitly about moral experimentation. But, though this treatment isn't explicitly about decreasing the cost of moral experimentation for young people, it has that effect. I think this should continue and increase. In particular, usually young-offender protections quickly fall away when the offender reaches the age of majority; but it is doubtful whether this accurately tracks the stages of moral experimentation. So moral experimentation might be another reason to agree with Brink (2020) and others that the ages at which young offenders are treated the same as older populations should increase; or more precisely that the severity of treatment should increase slower than at present.[40]

The same, I think, should hold in new fields, such as genetic engineering or artificial intelligence. In each of these cases, the stakes for 'getting it right' will only rise across time as the technology develops, opportunity for moral action or inaction is a repeated one, and (often) experiments can be conducted in public, so the information can be beneficial not just to those who conducted the experiments. It can, for example, be written about in the press. So we should postpone passing legislation on new fields longer than we currently do and take a prima facie more experimental

40. The same might be true of the age at which people are first held criminally responsible at all.

attitude.[41] This should, I think, also become a principle of legal interpretation. Now, it often seems as though we want to punish the moral arrogance of a moral pioneer; but even when they get it wrong we should be lightening sentencing or opting not to consider such conduct blameworthy.[42] We might also consider other policies to encourage moral risk taking in these new domains, such as disposing of some ethics review procedures, or — perhaps better — adding into the ethics review processes consideration of the moral information that might be exposed through the experiment.

The above is mostly about policy. But I think this work has implication for how we — as people far away from gene-editing policy and young offender sentencing — should live our lives. Where possible, I think we should indeed try to morally experiment and be open to updating on the results of those experiments — partially because of experimentation's intrinsic value (getting us moral information) and partially to help societally normalize such experimentation. But this is hard and I think the deontic objection gets at an important thing here: it is a tricky thing to make locally worse moral choices in the service of ultimate moral knowledge. So, probably most of all, we should set ourselves up, ex ante, to be put in positions where we naturally fall into such situations. This might happen by meeting new kinds of people, taking on positions of responsibility, or similar.[43]

And we should set up our social and familial lives as to encourage moral experimentation. We should be forgiving of moral mistakes, interested and pleased when those close to us change their moral views based on new information about the moral domain, and happy to share the results of experiments with those around us.[44] I take experimentation — both in our personal lives and in new fields that will influence humanity's future trajectory — to be a moral prerogative.

41. Some, e.g. Amodei et al. (2016), Dung (2023), Yampolskiy (2020), and Vold and Harris (forthcoming), have argued that certain new technologies — in particular artificial intelligence — might pose an existential threat to humanity in the near future. The point about moral experimentation is not particularly compelling in the case of existential risks: if we were to increase the risk of an existential catastrophe to get some moral information, this seems an unreasonably high price to pay for that information. Nevertheless, I think we ought still morally experiment — for now — in the case of AI. Even more, we should experiment especially intensely now, before stakes get quite as high as the existential stakes we may face in a few years.

42. I think, for example, of the sense of vindication at the trial of certain tech founders or financial fraudsters. It often seems that their arrogance itself is considered — quite apart from their crimes.

43. A friend reminds me that this is the point of one's time as an undergraduate.

44. Before we close, let me return briefly to saliency. Part of the value of moral experimentation done empirically, instead of only in one's mind, is that it helps one notice concepts with both a descriptive and an evaluative piece. It tunes us in to these parts of the interwoven normative and empirical domains. So, as moral experimentation is done around one (by one or by others) I think one should attend in particular to what aspects of the world this makes salient.

References

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul F. Christiano, John Schulman, and Dan Mané. 2016. "Concrete Problems in AI Safety." *CoRR* abs/1606.06565. arXiv: 1606.06565. http://arxiv.org/abs/1606.06565.

Anderson, Elizabeth S. 1991. "John Stuart Mill and Experiments in Living." *Ethics* 102 (1): 4–26. ISSN: 00141704, 1539297X, accessed November 29, 2023. http://www.jstor.org/stable/2381719.

Audi, Robert. 2013. *Moral Perception.* Princeton University Press.

Bal, P. Matthijs, and Martijn Veltkamp. 2013. "How Does Fiction Reading Influence Empathy? An Experimental Investigation on the Role of Emotional Transportation." *PLoS ONE* 8 (1): e55341. https://doi.org/10.1371/journal.pone.0055341. https://doi.org/10.1371/journal.pone.0055341.

Beckstead, Nicholas. 2013. "On the overwhelming importance of shaping the far future," https://doi.org/doi:10.7282/T35M649T.

Bostyn, Dries H., Sybren Sevenhant, and Arne Roets. 2018. "Of Mice, Men, and Trolleys: Hypothetical Judgment Versus Real-Life Behavior in Trolley-Style Moral Dilemmas." PMID: 29741993, *Psychological Science* 29 (7): 1084–1093. https://doi.org/10.1177/0956797617752640. eprint: https://doi.org/10.1177/0956797617752640. https://doi.org/10.1177/0956797617752640.

Brink, David O. 2020. "The Moral Asymmetry of Juvenile and Adult Offenders." *Criminal Law and Philosophy* 14 (2): 223–239. https://doi.org/10.1007/s11572-019-09518-4.

Carlsmith, Joe. 2023. "Seeing More Whole," February 17, 2023. Accessed December 19, 2023. https://joecarlsmith.com/2023/02/17/seeing-more-whole.

Cotton-Barratt, Owen, and Hilary Greaves. 2023. "A Bargaining-Theoretic Approach to Moral Uncertainty." *Journal of Moral Philosophy.*

Dancy, Jonathan. 2021. "The Role of Imaginary Cases in Ethics." In *Practical Thought: Essays on Reason, Intuition, and Action.* Oxford University Press, July. ISBN: 9780198865605. https://doi.org/10.1093/oso/9780198865605.003.0006. eprint: https://academic.oup.com/book/0/chapter/340190015/chapter-pdf/42882966/oso-9780198865605-chapter-6.pdf. https://doi.org/10.1093/oso/9780198865605.003.0006.

Demski, Abram. 2020. "Comparing Utilities." Accessed December 3, 2023. https://www.alignmentforum.org/posts/cYsGrWEzjb324Zpjx/comparing-utilities.

Desai, Nishant, Andrew Critch, and Stuart J Russell. 2018. "Negotiable Reinforcement Learning for Pareto Optimal Sequential Decision-Making." In *Advances in Neural Information Processing Systems,* edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, vol. 31. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2018/file/5b8e4fd39d9786228649a8a8bec4e008-Paper.pdf.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *arXiv preprint arXiv:1810.04805.*

Dorsey, Dale. 2021. "Francis Hutcheson." In *The Stanford Encyclopedia of Philosophy,* Summer 2021, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Dung, Leonard. 2023. "Current Cases of AI Misalignment and Their Implications for Future Risks." *Synthese* 202 (5): 1–23.

Fallenstein, Benja, and Nisan Stiennon. 2014. "Loudness Priors." Writeup from the May 2014 MIRI Workshop, May. https://intelligence.org/files/LoudnessPriors.pdf.

Foot, Philippa. 1967. "The Problem of Abortion and the Doctrine of the Double Effect." *Oxford Review* 5:5–15.

Hanna, Jason. 2015a. "Doing, Allowing, and the Moral Relevance of the Past." *Journal of Moral Philosophy* (Leiden, The Netherlands) 12 (6): 677–698. https://doi.org/https://doi.org/10.1163/17455243-4681049. https://brill.com/view/journals/jmp/12/6/article-p677_1.xml.

———. 2015b. "Enabling Harm, Doing Harm, and Undoing One's Own Behavior." *Ethics* 126 (1): 68–90. ISSN: 00141704, 1539297X, accessed November 29, 2023. http://www.jstor.org/stable/10.1086/682190.

Hänni, Kaarel, and Rio Popper. 2023. "Constraints on rational decision-making under moral uncertainty."

Hare, Caspar. 2016. "Should We Wish Well to All?" *Philosophical Review* 125 (4): 451–472. https://doi.org/10.1215/00318108-3624764.

Harris, John. 1975. "The Survival Lottery." *Philosophy* 50 (191): 81–87. https://doi.org/10.1017/s0031819100059118.

Huemer, Michael. 2005. *Ethical Intuitionism.* New York: Palgrave Macmillan.

Irwin, T. 2000. *Nicomachean Ethics (Second Edition).* Hackett Publishing Company. ISBN: 9781603845687. https://books.google.com/books?id=-Mf5XV8q6CgC.

Kagan, Shelly. 2001. "Thinking about Cases." *Social Philosophy and Policy* 18 (2): 44–63. https://doi.org/10.1017/S0265052500002892.

Kirchin, Simon T. 2010. "The Shapelessness Hypothesis." *Philosophers' Imprint* 10.

Lake, Brenden M., Ruslan Salakhutdinov, and Joshua B. Tenenbaum. 2015. "Human-level concept learning through probabilistic program induction." *Science* 350 (6266): 1332–1338. https://doi.org/10.1126/science.aab3050. eprint: https://www.science.org/doi/pdf/10.1126/science.aab3050. https://www.science.org/doi/abs/10.1126/science.aab3050.

Lockhart, Ted. 2000. *Moral Uncertainty and its Consequences.* New York: Oxford University Press.

MacAskill, William, Krister Bykvist, and Toby Ord. 2020. *Moral Uncertainty.* Oxford University Press.

McGrath, Sarah. 2009. "The Puzzle of Pure Moral Deference." *Philosophical Perspectives* 23 (1): 321–344. https://doi.org/10.1111/j.1520-8583.2009.00174.x.

———. 2018. "161Moral Perception and Its Rivals." In *Evaluative Perception.* Oxford University Press, June. ISBN: 9780198786054. https://doi.org/10.1093/oso/9780198786054.003.0009. eprint: https://academic.oup.com/book/0/chapter/152817924/chapter-ag-pdf/44969019/book_7713_section_152817924.ag.pdf. https://doi.org/10.1093/oso/9780198786054.003.0009.

———. 2019. "106C4Observation and Experience." In *Moral Knowledge.* Oxford University Press, December. ISBN: 9780198805410. https://doi.org/10.1093/oso/9780198805410.003.0004. eprint: https://academic.oup.com/book/0/chapter/348817409/chapter-pdf/43311501/oso-9780198805410-chapter-4.pdf. https://doi.org/10.1093/oso/9780198805410.003.0004.

McMahan, Jeff. 2000. "Moral Intuition." In *The Blackwell Guide to Ethical Theory,* edited by Hugh LaFollette -, 92–110. Blackwell.

Mill, John Stuart. 1859. *On Liberty.* Broadview Press.

———. 1963. *Principles of Political Economy.* Edited by J. M. Robson. Vol. 2-3. Collected Works of John Stuart Mill. Toronto: University of Toronto Press.

Mogensen, Andreas, and William MacAskill. 2021. "The Paralysis Argument." *Philosophers' Imprint* 21 (15).

Murdoch, Iris. 1970. *The Sovereignty of Good.* New York, Routledge.

Musgrave, R.A. 1959. *The Theory of Public Finance: A Study in Public Economy.* International student edition. McGraw-Hill. ISBN: 9780070855311. https://books.google.co.uk/books?id=GQMdAAAAIAAJ.

Nozick, Robert. 1974. *Anarchy, State, and Utopia.* New York: Basic Books.

Patil, Indrajeet, Carlotta Cogoni, Nicola Zangrando, and Luca Chittaro. 2014. "Affective basis of judgment-behavior discrepancy in virtual experiences of moral dilemmas." PMID: 24359489, *Social Neuroscience* 9 (1): 94–107. https://doi.org/10.1080/17470919.2013.870091. eprint: https://doi.org/10.1080/17470919.2013.870091. https://doi.org/10.1080/17470919.2013.870091.

Persson, Ingmar. 2013. *From Morality to the End of Reason: An Essay on Rights, Reasons, and Responsibility.* New York, NY: Oxford University Press.

Podgorski, Abelard. 2020. "Normative Uncertainty and the Dependence Problem." *Mind* 129 (513): 43–70. https://doi.org/10.1093/mind/fzz048.

Prinz, Jesse. 2006. "The Emotional Basis of Moral Judgments." *Philosophical Explorations* 9 (1): 29–43. https://doi.org/10.1080/13869790500492466.

Putnam, Hilary. 2002. "The Collapse of the Fact/Value Dichotomy." In *The Collapse of the Fact/Value Dichotomy and Other Essays.* Cambridge, MA: Harvard University Press.

Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. "Improving Language Understanding by Generative Pre-Training."

Railton, Peter. 2017. "Moral Learning: Conceptual foundations and normative relevance." Moral Learning, *Cognition* 167:172–190. ISSN: 0010-0277. https://doi.org/https://doi.org/10.1016/j.cognition.2016.08.015. https://www.sciencedirect.com/science/article/pii/S0010027716302050.

Rawls, John. 1958. "Justice as Fairness." *Philosophical Review* 67 (2): 164–194. https://doi.org/10.2307/2182612.

Raz, Joseph. 1975. *Practical Reason and Norms.* London: Hutchinson.

Romer, Paul M. 1986. "Increasing Returns and Long-Run Growth." *Journal of Political Economy* 94 (5): 1002–1037. ISSN: 00223808, 1537534X, accessed November 30, 2023. http://www.jstor.org/stable/1833190.

Ross, Jacob. 2006. "Rejecting Ethical Deflationism." *Ethics* 116 (4): 742–768. ISSN: 00141704, 1539297X, accessed May 14, 2023. http://www.jstor.org/stable/10.1086/505234.

Russell, Jeffrey Sanford. Forthcoming. "The Value of Normative Information." *Australasian Journal of Philosophy.*

Saam, Nicole J. 2017. "What is a Computer Simulation? A Review of a Passionate Debate." *Journal for General Philosophy of Science / Zeitschrift für Allgemeine Wissenschaftstheorie* 48 (2): 293–309. https://doi.org/10.1007/s10838-016-9354-8.

Samuelson, Paul A. 1954. "The Pure Theory of Public Expenditure." *The Review of Economics and Statistics* 36 (4): 387–389. ISSN: 00346535, 15309142, accessed November 30, 2023. http://www.jstor.org/stable/1925895.

Sanjay Chandrasekharan, Vrishali Subramanian, Nancy J. Nersessian. 2013. "Computational Modeling: Is This the End of Thought Experiments in Science?" In *Thought Experiments in Philosophy, Science, and the Arts,* edited by James Robert Brown Mélanie Frappier Letitia Meynell, 239–260. London: Routledge.

Savage, Leonard J. 1954. *The Foundations of Statistics.* Wiley Publications in Statistics.

Scanlon, Thomas. 1998. *What We Owe to Each Other.* Cambridge, Mass.: Belknap Press of Harvard University Press.

Schulzke, Marcus. 2014. "Simulating Philosophy: Interpreting Video Games as Executable Thought Experiments." *Philosophy and Technology* 27 (2): 251–265. https://doi.org/10.1007/s13347-013-0102-2.

Sepielli, Andrew. 2009. "What to Do When You Don?T Know What to Do." *Oxford Studies in Metaethics* 4:5–28.

Singer, Peter. 2005. "Ethics and Intuitions." *The Journal of Ethics* 9 (3-4): 331–352. https://doi.org/10.1007/s10892-005-3508-y.

Skaf, Rawad El, and Cyrille Imbert. 2013. "Unfolding in the empirical sciences: experiments, thought experiments and computer simulations." *Synthese* 190 (16): 3451–3474. ISSN: 00397857, 15730964, accessed December 9, 2023. http://www.jstor.org/stable/24019869.

Smith, Adam. 1976 (1776). *An Inquiry Into the Nature and Causes of the Wealth of Nations (Ed. R.H. Campbell, A.S. Skinner, and W. B. Todd).* Oxford University Press.

Stiglitz, Joseph E. 1999. "308Knowledge as a Global Public Good." In *Global Public Goods: International Cooperation in the 21st Century.* Oxford University Press, July. ISBN: 9780195130522. https://doi.org/10.1093/0195130529.003.0015. eprint: https://academic.oup.com/book/0/chapter/192852283/chapter-ag-pdf/44628084/book_25545_section_192852283.ag.pdf. https://doi.org/10.1093/0195130529.003.0015.

Swirski, Peter. 2007. *Of Literature and Knowledge: Explorations in Narrative Thought Experiments, Evolution, and Game Theory.* New York: Routledge.

Thomson, Judith Jarvis. 2014. "Killing, Letting Die, and The Trolley Problem." *The Monist* 59, no. 2 (December): 204–217. ISSN: 0026-9662. https://doi.org/10.5840/monist197659224. eprint: https://academic.oup.com/monist/article-pdf/59/2/204/4172191/monist59-0204.pdf. https://doi.org/10.5840/monist197659224.

Varian, Hal R. 1992. *Microeconomic Analysis.* Third. New York: Norton. ISBN: 0393957357 9780393957358.

Vold, Karina, and Daniel R. Harris. Forthcoming. "How Does Artificial Intelligence Pose an Existential Risk?" In *Oxford Handbook of Digital Ethics,* edited by Carissa Véliz.

von Neumann, John, and Oskar Morgenstern. 1947. *Theory of games and economic behavior.* Second edition. Princeton University Press.

Williams, Bernard Arthur Owen. 1985. *Ethics and the Limits of Philosophy.* London: Fontana.

Yaffe, G. 2018. *The Age of Culpability: Children and the Nature of Criminal Responsibility.* Oxford University Press. ISBN: 9780198803324. https://books.google.ee/books?id=XWpNDwAAQBAJ.

Yampolskiy, Roman V. 2020. "On Controllability of AI." *CoRR* abs/2008.04071. arXiv: 2008.04071. https://arxiv.org/abs/2008.04071.