The Meta-Ethics of Artificial Intelligence: Are We Beholden to Normative Joints?

Shamik Dasgupta
Draft of November 2020

How should we treat artificially intelligent beings (AIs)? What obligations do they have to us, and to each other? How does their presence affect our responsibilities to fellow humans? These questions were once the province of speculative fiction, but with the stunning technological progress of recent years they now loom very much on the horizon.

I don't have answers to these questions. Instead, my thesis is that there are no right or wrong answers to be had. There are no facts to discover about the ethics of AI, it is more a matter of inventing a new way of life.

The idea that ethics is more invention than discovery is nothing new. It is associated with certain anti-realist views on which there are no ethical facts in the first place. But my thesis does not rest on those views—I'm happy to grant a realist picture on which there are ethical facts out there to discover. Still, I'll argue that those facts don't settle AI ethics, for there are too many of them out there and it is up to us which ones to follow. The argument rests partly on considerations specific to AI and partly on general meta-ethical considerations. Regarding the first, I'll argue that AI is so unlike anything we've encountered before that we must ask which ethical concepts to use when dealing with them, our received concepts or new ones better tailored to our brave new world. And regarding the second, I'll argue that there is no right or wrong answer to *that*. All ethical concepts are on a par; we must simply choose which to live by.

1. Artificial intelligences

I should start by saying what I mean by 'AI'. Commentators typically distinguish domain-specific AIs, which excel at specific tasks but not much else (think of self-driving cars); general AIs, which would posses more general-purpose human capacities such as the ability to invent, plan, and reason about a variety of domains; and superintelligent AIs, whose capacities would vastly exceed ours in ways we cannot imagine. I'll use 'AI' broadly to include all these, even the rudimentary domain-specific systems found in contemporary sexbots. Indeed, sexbots are likely to be the locus of our first serious confrontation with questions about the ethical status of AIs, given the consumer demand for sex and the complex ethical terrain surrounding it. Already there are discussions about whether using a sexbot could count as cheating (Cheok 2017, Ziaja 2011) and whether issues of consent and age apply to them (Levy 2007, Sullins 2012).

Current sexbots are nothing like real humans, of course. But with technological developments we could in principle produce artificial entities that are almost indistinguishable from us in both cognitive and physical respects—just think of the "replicants" in the *Blade Runner* film franchise. These raise a host of ethical questions. Is it permissible to produce replicants for labor, or would that be a form of slavery? Is destroying a replicant like destroying a laptop or more like murder? Does the fact that we created them give us leeway to treat them as we please or additional responsibility for their welfare?

This is an easy case—the franchise invites us to think of replicants as our equals and it's hard to disagree. Less clear are AIs that differ vastly from us in cognitive architecture and physical implementation. Consider Hal from *2001: A Space Odyssey*, a supercomputer that

controls most aspects of a spacecraft singlehandedly and communicates with humans through a microphone and speaker. Physically speaking, Hal is a room-sized piece of silicon circuitry. And given Hal's super-human capacities, we can only imagine that his cognitive architecture bears little resemblance to ours. Now the ethical questions are less clear. Do these differences leave Hal outside the moral circle, or does his superintelligence render him *more* ethically significant than us? If Hal "intentionally" kills a human, is that like a mechanical accident or more like murder? If the latter, could Hal absolve himself of responsibility by "wiping his hard-drive" after the event (perhaps keeping a backup of his former self for future use)?[1]

Hal doesn't have a human-esque body, but he might find one useful to perceive and manipulate the environment as we do. We might therefore equip systems like him with robotic "bodies" that they control by radio-transmitter. Or they might prefer to rent one by the hour when needed—a "zip body", if you like. Now the ethics gets murkier still. If Hal rents a body which then destroys a human, could we hold the body itself responsible? What if the body is then rented by another system—is it ethically speaking a new individual or do its past responsibilities remain?

We can even imagine that Hal is capable of renting thousands of bodies at once. The bodies would live out separate, human-like lives around the globe, each one with a unique "personality" and distinctive quirks. But unbeknownst to them their cognitive processing is done not in their heads but by Hal in a warehouse in Houston. Rather than process each "stream of thought" separately, Hal might collate the information from each body into a central store, process it together, and then deliver personalized instructions back to each body. Now the units of ethical evaluation start to break down. Given their separate lives and personalities, is each body a distinct ethical unit with something like the status of a human being? Or given the computational connectedness, is this ethically speaking a single individual scattered across the globe?

These fanciful scenarios illustrate that questions of AI ethics may concern entities that live a very different "form of life" from us. Indeed, it's not obvious whether the bodies just described fall under our ordinary psychological categories. Do they count as having beliefs, desires, and emotions? Do they act with intentions or reasons in anything like the way we do? Given their radically different cognitive architecture, it's not clear that these descriptions are appropriate in anything but the most attenuated sense (perhaps these artificially intelligent systems don't count as "intelligent" in the ordinary sense of the term!). Yet for all that there is clearly room to ask ethical questions about our responsibilities to them and theirs to us.

And this is just the beginning. AI systems could in principle be implemented in patterns of radio waves propagating through the universe. They might have the ability to clone themselves at will, rendering them (in effect) immortal. They might experience time and causation very differently from us—ordinary causal interactions we attend to might be unrecognizable to them, or at least insignificant to their form of life.[2] All this means that AI ethics is quite unlike anything we've done before. Even the ethics of non-human animals concerns entities that resemble us enough to be at least *loosely* describable in familiar terms. The wilder AIs just described are a very different beast altogether.

---

[1] See DiGiovanna (2017) for a discussion of this last question.

[2] See Bostrom and Yudkowsky (2011) for a discussion of some of the ethical issues concerning the experience of time.

2. Cosmopolitan ethics

How then are we to get a handle on AI ethics? For concreteness, focus on questions of how to treat AIs like Hal. Imagine you're working with Hal and become annoyed with his computerized voice. You consider shutting him down like a laptop but wonder whether that would be like murder. What do you do?

Philosophers typically approach this kind of question by asking whether the entity in question has *moral status*. The idea is that things with moral status, such as humans, matter in themselves and for their own sake, whereas laptops don't—you can do what you like with them without concern for their well-being. So, if Hal has moral status it would be wrong to shut him down.

But how are we to determine whether Hal has moral status? Two methods suggest themselves. The first is to settle on a general theory of what makes for moral status, of the form

> x has moral status if and only if x has feature F

and then ask whether Hal has F. There are many such theories—including ones on which F consists in sentience, or rationality, or membership of the species *Homo sapiens*—so it's unsurprising that they've been used to address questions of how to treat AIs like Hal.[3] The trouble is, there's little consensus on which theory is right. Indeed, the abstract question of what makes for moral status would appear no more tractable than our original question of how to treat Hal.

The second method starts with the fact that humans have moral status and then argues that anything resembling us in some respect would have moral status too. This doesn't need a general theory of moral status, just a sufficient condition of the form

> If x resembles humans in respect R then x has moral status.

For example, Schwitzgebel and Garza (2015) argue that anything resembling us in social and psychological respects would have moral status.[4] But the trouble with this method is that it says nothing about the interesting cases. It implies that the *Blade Runner* replicants have moral status, but we suspected as much already. What about Hal controlling 8000 zip-bodies from Houston? It doesn't resemble us in any obvious social or psychological respects, so their sufficient condition doesn't imply whether it has moral status or not.[5]

These points are straightforward. But I want to raise a deeper problem for both methods, namely they are objectionably *parochial*. They take our folk concept of moral status for granted and use it uncritically to evaluate how to treat Hal. This strikes me as backwards. Our concept of moral status has a particular genealogy, a messy history of cultural and evolutionary events that made it what it is today. It was acquired and refined over the years as we interacted with things

---

[3] See for example Basl (2013), Ashrafian (2015), and Talbot *et al* (2017). There are also views on which things can have an 'erzatz' moral status due to the fact that it would be good for humans to *treat* it as having moral status; Darling (2016) applies this view to robot ethics.

[4] This approach is also pursued by Bostrom and Yudkowsky (2011) and LaBossiere (2017).

[5] This is no objection to Schwitzgebel and Garza *per se*, for they were just arguing that AIs with moral status are *possible* and their sufficient condition is enough to establish that.

we encountered along the way—things like humans, dogs, deer, and dung-beetles. Perhaps it's well-suited to guiding our treatment of them, but that doesn't mean it's appropriate when it comes to radically different forms of life like Hal.

There are a number of worries in this vicinity. One is semantic, of whether our concept of moral status has a determinate extension beyond the familiar things we've encountered so far (by an extension, I mean a set of things to which the concept truly applies). Another is epistemic, of whether the candidates for moral status we've encountered so far is too biased a sample to justify a general theory. But the problem I want to press is different. Suppose for the sake of argument that 'moral status' has a determinate extension, the set of sentient beings. And suppose we know this. Still, sets are cheap: there are innumerable sets out there, some largely overlapping the set of sentient beings but differing at the fringes, others very different. For every set, there's a possible concept with that set as its extension. Does *our* concept of moral status pick out an important set? Given all the concepts we *could* use to evaluate our treatment of Hal, is ours the right one? The methods discussed above are objectionably parochial insofar as they unreflectingly assume that it is.

To illustrate, imagine we had introduced our concept of moral status with the stipulative definition that it is true of all and only *carbon-based* sentient beings. And suppose Hal and his silicon pals had introduced a corresponding concept of 'shmoral status' with the stipulative definition that it is true of all and only *silicon-based* sentient beings. When evaluating the action of destroying Hal, our standard is whether Hal has moral status while theirs is whether he has shmoral status. Finally, assume that Hal is sentient. Then we'd say

> "Hal doesn't have moral status."

and they'd say

> "Hal does have shmoral status."

and both statements would be true. They'd therefore evaluate the action of destroying Hal negatively, condemning it fiercely, and that would be correct given *their* standard of shmoral status. By contrast, we'd evaluate the action as being fine, and that would *also* be correct given our standard of moral status. It would then seem objectionably parochial to soldier on and use our standard without reflecting on *whether it's the right standard to use*. It may be appropriate when dealing with familiar things like dogs and deer, since every candidate for moral concern we've encountered so far has been carbon-based. But that doesn't mean it's suitable when it comes to things like Hal—indeed it would seem unsuitable precisely because it rules him out of the "moral circle" by *fiat*. Thus, to unreflectingly use our received concept of moral status would make us guilty of a kind of parochialism—of assuming that concepts tailored to our familiar pastoral life are apt for wider purposes.

But is anyone guilty of this? We presumably use our concept of moral status because we assume that things falling under it *deserve respect*, or that there are *reasons* to treat them well, and so on. If there are reasons to treat Hal well, everyone will agree that a concept of moral status defined in terms of carbon isn't the right one to use. And indeed the literature on moral status has proceeded along more or less these lines, asking which account of moral status connects with other considerations of respect and reasons in the right way.

But seen like this, the question of whether to destroy Hal ultimately hangs on these other considerations of respect and reasons. And then the same issue arises over again. Our concept of something's being a 'reason' to treat Hal well has a certain extension, but Hal and his AI pals may use a concept of 'shmeason' with a different extension as their standard by which to evaluate action. Hal's superintelligence is a shmeason to treat him well, let's suppose, but not a reason. Thus, we'll say

"There is no reason to treat Hal well"

and they'll say

"There is a shmeason to treat Hal well"

and both statements are true. They'll then condemn the action of destroying Hal; we'll evaluate it as fine; and both evaluations are correct given our respective standards. It then seems objectionably parochial to act on the basis of our received standard without pausing to ask whether it's the right one to use.

Ethics involves asking what to do, how to live, and who to be. It therefore requires evaluating actions, projects, and characters. *Parochial* ethics uses our received normative concepts as the evaluative standards. It evaluates the action of destroying Hal according to whether Hal has moral status, whether there are reasons to treat him well, and so on. By contrast, *cosmopolitan* ethics involves asking which concepts to use as our standard in the first place, e.g. reasons or shmeasons. I claim that the ethics of AI must be cosmopolitan. We must recognize that the standards that have suited our limited lives so far might not be suited to the wider world.

This point is familiar in scientific reasoning. Suppose one community uses color concepts like 'green' to categorize gemstones while another uses concepts like 'grue'.[6] In parochial science, each community uses their received concepts. Given two green gemstones only one of which is grue, the first community will say they're the same while the second will say they're different, and both are correct relative to their respective category concepts. In cosmopolitan science, by contrast, each community would ask which concepts to use in the first place. Science is often cosmopolitan in this sense. Our folk concepts of space, time, and matter are well-suited to ordinary life but not for describing quantum states or black holes. Scientific progress often involves fashioning new concepts better suited to unfamiliar domains. Ethical progress sometimes requires this kind of conceptual revision also. Williams (1985) recognized this when he noted that "thick" concepts like chastity are tailored to particular cultural contexts and should be questioned as times change. But cosmopolitan ethics as I conceive it is more general insofar as it applies to *all* normative concepts, thick and thin. No concept is sacrosanct, all are potentially under scrutiny.[7] Eklund (2017) discusses this project of cosmopolitan ethics in the abstract; my claim here is that it is central to first-order ethical questions concerning AI.

I'll soon argue for this, but first it will help to introduce a simple framework for discussing cosmopolitan ethics. Focus just on the evaluation of actions, and consider the vast range of

[6] This example is from Goodman (1955), who defined 'grue' thus (for some future time t): x is grue if and only if x is first observed before t and green, or not first observed before t and blue.
[7] Though not necessarily all at once; one could proceed piecemeal *a la* Neurath's boat.

standards against which they can be evaluated as meeting or not meeting. There's the *utilitarian* standard of maximizing pleasure, the *Kantian* standard of consistency with the categorical imperative, the *Humean* standard of satisfying one's desires, and (if God exists) the *divine* standard of promoting God's will. Suppose we happen to use the utilitarian standard: we praise actions that meet it, we call them the "right" thing to do, and so on. Assume further that because of this practice our term "right" has come to be true of an action if and only if it maximizes pleasure.[8] Nonetheless, actions can also be evaluated against the Kantian standard. This is not to ask whether they're *right*, by assumption. Still, say that an action is *kwight* if and only if it meets the Kantian standard; then the point is that an action can be evaluated for whether it's right *or* whether it's kwight. Likewise, say that an action is *hwight* iff it meets the Humean standard, and *dwight* iff it meets the divine standard. Then an action can evaluated as to whether it's hwight or whether it's dwight. While parochial ethics just asks whether an action is *right*, cosmopolitan ethics first asks which standard to use and then evaluates action accordingly.

If an action is right it's right *simpliciter*; by assumption, it makes no sense to say that it's right according to one standard but not another. But it will be useful to make inter-standard comparisons like these. I'll use "correct" as a term-of-art for this purpose. Thus, an action can be correct by the lights of one standard but not another, meaning just that it meets the first and not the second.

I said that cosmopolitan ethics asks which standard to use, but what exactly is the question? It isn't which standard is *right*, for that's just to ask which standard meets the parochial standard *we* express by 'right' and the whole point was to put our parochial standards under scrutiny. Nor is it which standard we *should* use, or which is *best*, for the same reason. And nor is it which standard is *correct*, for something is correct only relative to a choice of standard and we're asking which standard to use in the first place. Rather, the question must be

> Which standard *shall* I use?

where this is answered not by forming a belief about which I *should* use, or a prediction of which I *will* use, but by making a *decision* to use a certain standard. If this sounds artificial, remember that practical reasoning begins with questions like this. Shall I volunteer at the local food bank? Shall I destroy Hal? These questions are answered by making a decision: one decides to volunteer or not, to destroy Hal or not.[9] Parochial ethics is the business of evaluating these decisions as correct or incorrect according to our received standards—e.g. whether volunteering is *right*. The cosmopolitan question takes the same form; the only difference is that it asks us to decide not on an action but on a standard by which to evaluate them.

3. Normative joints

But how are we to evaluate *that* decision? Are some standards distinguished over the rest as *the ones to use*? If so, we must ask what makes them normatively significant in this regard.

---

[8] Here I don't assume that usage alone fixes truth-conditions, just that it's a contributing factor.

[9] Alternatively, one might say that these questions are answered with an *intention*. But this subtlety doesn't matter: the point is that these questions are answered with a conative state akin to decision. This state is sometimes characterized as judging that the action is "the thing to do" (see e.g. Gibbard (2003)), but that's not quite right. You may just think it's *a* thing to do and you've decided to do it.

Is it something about *us* that distinguishes them, or do they have this status independently of us?

The latter view is what Eklund calls "ardent realism". As he puts it, 'reality itself favors certain ways of valuing and acting' (2017, p. 1); that is, certain standards are distinguished as normatively significant by reality itself, independently of us. I'll call these standards *normative joints*. Suppose an action maximizes pleasure but frustrates God's will. Then it's correct by the lights of the utilitarian standard but not the divine standard—it's right but not dwight (on current assumptions about 'right'). But according to ardent realism there's a human-independent fact about which standard is privileged; about whether 'right' or 'dwight' carves at the normative joints. Thus, in addition to being correct by the lights of some standards and not others, there's a further fact as to whether the action is what I'll call "*really*-correct" or "*really*-right"; i.e., correct by the lights of the jointy standard. If that standard is the divine one, the action is right but not really-right.

This is analogous to the more familiar idea that reality favors certain ways of *describing* it. Suppose one community uses 'green' while another uses 'grue'. Both may form true beliefs with their respective concepts, but according to Sider (2011) there is a further fact as to which concept carves at the natural joints—or "descriptive joints", as I'll call them—and the aim of inquiry is not just to believe truths but to do so in terms of concepts that carve at those joints.[10] Descriptive and normative joints therefore play analogous roles in theoretical and practical reasoning, respectively, as human-independent constraints on which concepts to use. But they should not be conflated. Even if green is a descriptive joint, it certainly isn't a normative joint—reality doesn't favor painting the town green! Likewise, if pleasure is a descriptive joint it doesn't follow that reality favors actions that maximize pleasure. In the other direction, can something be a normative but not a descriptive joint? This is less clear: one might think that, say, pleasure-maximization can be a normative joint only if pleasure (or perhaps sentience) is a descriptive joint.[11] But this would be substantive connection in need of argument; it doesn't follow just because both are called "joints".

If the divine standard of promoting God's will is a normative joint, that's presumably because *God* has a distinguished normative status. I'll therefore use "normative joint" broadly to describe the components of a standard, such as God, as well as the standard itself. I'll also talk at times as if at most one standard is a normative joint. This is a simplification as there could in principle be many jointy standards, each one relevant to a different context. But the simplification is harmless; if you like, just disjoin the standards and treat them as one.

I've focused on 'right', but the same goes other normative concepts like 'good', 'virtue', 'justice', 'reasons', 'obligation', 'permission', and so on. For these concepts intertwine—*virtuous* people tend to do the *right* thing, for example, and typically for good *reasons*—so a distinction between right and really-right will induce a distinction between virtue and real-virtue, reasons and real-reasons, etc. Precisely how these concepts intertwine is controversial but matters little for our purposes. To see this, consider the "reasons-first" view on which the fundamental

[10] I call these "descriptive joints" rather than "natural joints" because the latter can lead to conflating the general thesis that reality itself prefers some ways of describing it with the specific thesis that those ways are determined by which properties are "perfectly natural" in David Lewis's (1984) specific sense. As I use the term, descriptive joints could instead by determined by which properties are *ungrounded*, or which correspond to *universals* or *God's ideas*, and so on.

[11] Lee (forthcoming) argues for something in this vicinity.

normative notion is whether a fact F is a *reason* for someone S to do A, and other normative matters like the good and the right derive from this. To assess F as a reason is to evaluate whether it "supports" or "speaks in favor of" S's doing A, and with evaluation comes standards. There is the utilitarian standard on which F must imply that A maximizes pleasure, the divine standard on which F must imply that A promotes God's will, and so on. Suppose our parochial term 'reason' expresses the utilitarian standard, so that F is a reason for S to do A iff F implies that A maximizes pleasure. Still, let F be a *dweason* for S to do A iff F implies that A promotes God's will. The cosmopolitan question, then, is which standard to use when evaluating whether F supports S's doing A, the reasons-standard or the dweasons-standard. And the thesis of normative joints is that one standard is normatively significant independently of us. Thus, F "*really*-supports" S's doing A—or is a "*real*-reason" for S to do A—if it supports S's doing A by the lights of the privileged standard. This distinction between reasons and real-reasons will then induce an analogous distinction for any normative notion that depends on reasons—which, according to the reasons-first view, includes them all.

Don't confuse the thesis of normative joints with what typically gets called 'realism' in the meta-ethics literature, for the latter is a view about the status of *parochial* ethics; of parochial normative judgments about right, good, reasons, etc. More fully, realism is typically characterized as holding (i) that these normative judgments are beliefs that can be true or false; (ii) that some normative beliefs are true; and (iii) that normative truths are mind-independent, i.e. they hold independently of our beliefs, values, and other attitudes. This leaves open whether these normative truths involve properties that are ultimately natural, i.e. reducible to properties akin to those of the empirical sciences (naturalism), or whether they constitute a *sui generis* domain over and above the natural order (non-naturalism). Either way, realism likens parochial ethics to astronomy insofar as both concern truths that are "out there anyway", independent of us. By contrast, views like subjectivism agree that there are normative truths but claim that they depend in some measure on our values and other attitudes, *contra* (iii). Yet other views deny that there are normative truths in the first place. These include error-theory, which rejects (ii) and claims that normative beliefs are systematically false, and non-cognitivist views that reject (i) and claim that normative judgments aren't truth-evaluable beliefs in the first place.

The point is that these views concern the status of parochial judgments and are, therefore, independent of the question of normative joints. To see this, note that realism does not imply that there are normative joints. All it implies is that parochial concepts like 'good' express mind-independent properties, but it doesn't follow that those properties are normative joints or that there are normative joints in the first place.[12] Compare with the case of descriptive joints: the property of being an electron or a cow is mind-independent, but it's certainly no descriptive joint! This goes even for non-naturalist realism: that view implies that our parochial concepts like 'good' express properties that aren't naturalistically reducible, but it doesn't follow that they are normative joints. For all non-naturalism says there may be gazillions of other non-natural properties out there and nothing significant about the few we express other than that *we* organize our lives around them. Admittedly, non-naturalist philosophers tend to talk as if their non-natural properties as special, if only implicitly. If that's what they mean, fine—that just

[12] Eklund (2017, chapter 1) argues for this more thoroughly than I do here. Admittedly, some discussions of realism assume implicitly that the mind-independent properties would be normative joints—Mackie (1977) and Street (2006) may be two examples. And of course 'realism' is a term of art that different authors may use differently. My point is just that the definition I present in the text—which is hardly idiosyncratic—does not imply that there are normative joints.

means they embrace the thesis of normative joints. My point is just that non-naturalism as typically defined—that 'good' and 'right' express non-natural properties—does not imply that they are normative joints. Nor does the thesis of normative joints imply that the joints involve non-natural properties: the jointy standard could involve wholly natural properties such as pleasure-maximization.

Still, the thesis of normative joints does imply a non-naturalism of sorts, for the property of *being a normative joint* is not naturalistically reducible. This may not be obvious. To say that a standard S_1 is a normative joint is to say that it is normatively significant independently of us. But why couldn't S_1 be significant in virtue of some natural property P_1 it has independently of us? Well, it could. But properties are cheap—there's a property for every set. For any standard S_2 there'll inevitably be some property P_2 that stands to S_2 just as P_1 stands to S_1. So, for S_1 to be normatively significant over S_2, P_1 must have some normative significance that P_2 lacks. If this fact about P_1 is irreducible, then the fact that S_1 is a normative joint bottoms out in this *normative* fact about P_1 and is not naturalistically reducible. If instead some further property P_1* confers this significance on P_1, there'll be another property P_2* that stands to P_2 just as P_1* stands to P_1… and so on. But for S_1 to be significant over S_2, this cannot go on forever. At some point, *something* must be normatively significant all on its own, not in virtue of anything else. Thus, if S_1 is a normative joint this is either a primitive fact about S_1, or else S_1 inherits its significance from something else and it's a primitive fact that *that* is normatively significant. Either way, there's irreducible normativity somewhere. The difference between this and what is typically called 'non-naturalism' is that it's non-naturalism about the property of *being a normative joint*, not the properties expressed by parochial terms like 'good'.

This then is the thesis of normative joints: there are human-independent facts about which standards are normatively privileged and the aim of cosmopolitan ethics is to discover them. On this view, cosmopolitan ethics is akin to astronomy insofar as both aim to discover truths that are "out there anyway". And what if there are no normative joints? Then there are two possibilities: either something about *us* distinguishes a standard as normatively privileged, or *nothing* does and all standards are on a par. I'll call the former view "anthropocentrism" and the latter "egalitarianism".

I'll soon argue for egalitarianism, so let me emphasize the striking effect it has on practical reasoning. Imagine you're wondering whether to destroy Hal. You start by treating the question parochially, asking what the right thing to do is. And assume *per* moral realism that there are mind-independent facts about right and wrong to discover. This is typically seen as the pinnacle of objectivity in ethics. As Korsgaard put it, "moral realism conceives ethics on the model of applied knowledge" (2008, p. 317)—knowledge of facts that are out there anyway.

But when ethics goes cosmopolitan the picture changes dramatically. If there are mind-independent facts about right and wrong, there are also mind-independent facts about dwight and dwong. Destroying Hal may be right but not dwight. There are countless standards out there and any action will inevitably be correct by the lights of some but not others. Cosmopolitan ethics asks which standard to go by, yet according to egalitarianism they're all on a par. But then all *actions* are on a par too, for *whatever* you do is correct by the lights of *some* standard and on the egalitarian view that is all there is to say. In this way the egalitarianism about standards seeps down to the actions themselves: anything goes, there is no possibility of error whatever you do! And now this looks very much *unlike* applied knowledge of mind-independent

facts are more like an unconstrained choice of how to live. Yet according to egalitarianism, this is what cosmopolitan ethics amounts to *even if moral realism is true!*

The thesis of this paper is that AI ethics has precisely this status. My argument rests on three premises:

(1) The ethics of AI should be cosmopolitan.

(2) If there are no normative joints then all standards are on a par (egalitarianism).

(3) There are no normative joints.

It follows that there are no facts to discover about how to treat AIs like Hal, it is a matter of freely inventing a way of life. And this is so even if moral realism is true, as just indicated.

The rest of this paper will support these premises. But first, it's worth emphasizing the sterility of moral realism just exposed. The view is *supposed* to imply that ethics is as "objective" as astronomy in that its questions have mind-independent answers—answers that hold independently of our attitudes and values. But without normative joints it leaves ethics no more an objective enterprise than sport. To illustrate, in soccer a player is offside when closer to the opponent's goal than the second-last opponent. Being offside is therefore a mind-independent property in the sense that it depends just on physical distances and not on anyone's attitudes or values—we should be realists about offsides.[13] Still, the offside rule was obviously invented by us. We can and do change it as we go along. The rules of 1863 stated that a player was offside if closer to the opponent's goal than the *third*-last opponent—call this the offside* rule—but in 1990 the rule was revised to encourage more attacking play. We use the current rule because we prefer it: it makes for a more enjoyable game given human athletic abilities and aesthetic sensibilities. There are obviously no human-independent facts that distinguish it as *the* rule to use—there are no "sporting joints"! Thus, while being offside and being offside* are both mind-independent properties, the rules are nonetheless are up to us to invent. Likewise for moral realism without normative joints: being *right* may be a mind-independent property, but it remains up to us whether to use the standard of rightness or dwightness as a guide to life.[14]

Of course, we might choose a standard because *we like* to live by it, just as we choose the offside rule because we like the resulting game. On the anthropocentric view, our preference for that standard gives it normative significance—it makes it *the one to use*. But for egalitarians, our preference gives it no such significance: to choose another standard—one we *don't* like—would be no mistake, just a different choice. The fact that we tend to choose ones we like is, on this view, simply an empirical fact about *what we do*. To be clear, then, egalitarians needn't deny that in choosing a standard we'll likely think carefully about which one we prefer, which one will serve us best given the particular problem at hand, and so on. In this sense our choice may be

[13] To be clear, I'm talking about actually *being offside*, not the referee's call. While the referee determines how the play is recorded, they can make mistakes and hence don't determine whether the player was in fact offside.
[14] Enoch (2011) agrees that moral realism doesn't suffice for real objectivity. But we disagree on what would suffice. Enoch claims that non-naturalist realism suffices, but we've seen why that's not enough: even if "good" and "right" expresses non-natural properties, they need not be normative joints in which case the discussion in the text goes through just the same. Relatedly, Dunaway and McPherson (2016) argue that realism properly construed requires that moral properties are descriptive joints. But this doesn't suffice for real objectivity either: the property of being an electron is arguably a descriptive joint, but that doesn't make it an objective constraint on practical reasoning.

said to involve “discovery”. But the same is true of other creative acts like inventing a sport or a new style of music: they too involve careful experimentation which could be described as “discovering” the one we want. The point is that for egalitarians, this is just an empirical fact about our habits of choice. To choose differently would be no mistake, just a different form of life. Premises (1)-(3) imply that AI ethics is like *that*. We must simply do what we do and there is no possibility of error.

Let me now turn to defending these premises in turn.

4. Alternative standards

Start with (1), the claim that AI ethics should be cosmopolitan. Here I mean ‘should’ in the parochial sense. Thus, (1) states that when it comes to AIs *our parochial standards themselves* require that we ask which standards to use when evaluating how to treat them. Since this is a claim of parochial ethics my argument will sometimes be parochial too, resting on commonsense moral judgement and the like. Hardly water-tight, but such is the nature of parochial ethics.

The argument for (1) is straightforward if there are normative joints. For in that case the sole aim of morality is to conform to those joints: to do what’s *really*-right, to act on the basis of *real*-reasons, and so on. But there’s no guarantee that our parochial concepts carve at those joints—it’s always possible that what’s right is not what’s really-right. Hence we must always be awake to the cosmopolitan question of whether our parochial standards are the ones to use, regardless of whether we’re dealing with AIs. If there are normative joints, *all* ethics should be cosmopolitan.

What if there are no normative joints? Then (1) can be motivated by noting that human moral cognition has a genealogy that renders it suitable for some environments but not others. I mentioned this briefly in section 2 but let me expand. We don’t yet know what the exact genealogy is, of course. Evolutionary psychologists tend to promote the idea that our ancestors during the Pleistocene era faced environmental challenges that led to new kinds of social behavior. Survival and reproduction required living in bands of around 150, engaging in coordinated activities like group hunting, sharing food and other resources, and competing with other groups. On this view, moral cognition associated with *fair* distribution, *punishment* of free-riders, and *loyalty* to one’s in-group, are adaptations that conferred a fitness advantage by producing the social behavior required for survival. By contrast, some anthropologists emphasize the role of cultural evolution, a process whereby individuals do better by copying behaviors and traditions that were refined over generations than re-inventing the wheel themselves. Moral cognition associated with *tradition* and *conformity* would facilitate this mimicry, while concepts of *prestige* and *social status* would help determine who to mimic. Yet others propose that moral concepts are a more recent invention that post-dated the agricultural revolution.[15]

[15] The literature on evolutionary psychology is nicely summarized in Cosmides, Guzman, and Tooby (2019); see also Green (2007) and Tomasello (2016) for distinctive takes. For an introduction cultural evolution, see Richerson and Boyd (2005) and Henrich (2015). For an example of the idea that morality post-dated the agricultural revolution, see McCullough (2020). All these approaches rest on game-theoretic ideas summarized in Skyrms (2014).

But the case for (1) doesn't hang on the details. The point is that our moral concepts didn't come from nowhere. They have a causal history and are what they are in large part because of the vagaries of that history. They were shaped by challenges faced in ecological and social environments that were particular to a time and place.[16] Therefore, we cannot assume that they are well-suited to new environments containing AIs. Just as our craving for high-calorie foods was beneficial when food was scarce but not in a world of cheap fast-food, concepts like in-group loyalty that were useful in the Pleistocene could lead to ruin in the modern era.

The analogous point about *physical* concepts is uncontroversial. Our parochial concepts of space, time, and matter helped a bipedal ape gather mushrooms and avoid predation, but they are notoriously *un*helpful for understanding quantum tunneling and black holes. Our physical concepts are therefore not sacrosanct: we can and should reflect on whether they're well-suited to new domains. Why should moral concepts be any different? The analogous point about sport is less controversial still. If humans started growing regularly to 15 feet, the National Basketball Association would increase the regulation height of the basket else the game would become silly. The rules of basketball are what they are because they solve a problem: they make for an interesting game given constraints like body sizes, so when the constraints change we must ask whether the rules remain fit-for-purpose. Again, why should moral concepts be different? Without normative joints our moral concepts are no more sacrosanct than the rules of sport, as we saw in the last section.

This suggests that ethics should *sometimes* be cosmopolitan. Why the ethics of AI? Just because the difference between AIs like Hal and the social apes of the Pleistocene strikes me as akin to the difference between black holes and mushrooms, or between 15-foot giants and contemporary humans with respect to basketball. Admittedly, I have no well-defined metric and rest on commonsense moral "intuition" here. But if you agree that we should *sometimes* question our parochial standards, it would be strange if our dealings with AI is not such a time. Remember, the claim is not that new standards are definitely needed, just that the question should be raised. Nor is the claim that the question is *only* pertinent to AI: it may be independently reasonable in our current environment of nation states, legal structures, and financial institutions so different from our Pleistocene past. But a world with wild AIs like Hal strikes me as *so* alien as to render the cosmopolitan question inescapable.

This is an evolutionary debunking argument of sorts, but not the epistemic variety often wielded against moral realism. The latter argues that given the evolutionary origins of moral cognition, there's no reason to expect that beliefs about mind-independent moral properties would be reliably true. But my argument is practical, not epistemic. I emphasize the genealogy of moral cognition not to raise a question about moral knowledge, but a *practical* question of whether to continue living by past standards. Responses to the epistemic debunking argument tend to focus on epistemic matters and to that extent do not speak to my argument here.[17]

---

[16] To be clear, the fact that moral cognition has a causal history does not directly imply that its nature depends on that history, for it could be that all (or most) histories would lead to the same kind of cognition. But this possibility can be ruled out by noting the diversity of social organizations found throughout the animal kingdom, each one adapted to a particular environmental challenge. That empirical fact alone shows that our manner of social organization wasn't inevitable; hence our moral faculties that sustain it weren't inevitable either.

[17] For an epistemic debunking argument, see Street (2006). For responses see Shafer-Landau (2012) and Fitzpatrick (2015). For what it's worth, I think that the epistemic debunking argument fails by the lights of the evolutionary theory it rests on. If we evolved a concept of a mind-independent standard because that aided survival and reproduction,

That completes my argument for (1). The main objection to (1), I believe, is not that the cosmopolitan question is inappropriate but that it is incoherent. It asks which standard to use, yet this presupposes that it's *possible* to use an alternative to our parochial standard in the first place. Some philosophers think it isn't. The worry is that an agent can be said to *use* a standard only if she has a concept of it that plays a certain role in practical reasoning. And according to a thesis Eklund (2017) calls "referential normativity", this practical role fixes the concept's extension. If so, the objection goes, it's impossible to use an alternative standard and the cosmopolitan question doesn't get off the ground.

One might reply that this objection relies on an overly individualist conception of "using a standard". New standards can be implemented not by agents acquiring new concepts but by institutional structures that make agents act in new ways.[18] Still, the objection shouldn't convince even on the individualist conception, for two reasons.

First, referential normativity is hard to believe. To appreciate this, suppose our concept 'right' plays the relevant role in practical reasoning, and suppose the role is that judging an action to be right disposes an agent to perform it. And assume, as before, that the extension of 'right' is the set of actions that maximize pleasure. It follows that we use the utilitarian standard as a guide to action. Still, it seems *clearly* possible for another community to use the divine standard instead: it's just a possibility in which they're disposed to do what they judge to promote God's will instead. Now, fans of referential normativity might reply that the practical role can be played only by a concept that is "thin" in the sense that it has no constitutive links to non-normative contents. It may be true that an action is right iff it maximizes pleasure, but not a conceptual truth. But I didn't suppose otherwise, so their point must be that the possibility of using the divine standard is one in which the practical role is played *not* by the concept 'promoting God's will' *per se* but by some thin, co-extensive concept D that (we can imagine) they express by 'right'. But why is that incoherent? It's clearly possible for them to act *as if* they use the divine standard in the sense that that standard causally regulates their application of D and, consequently, their behavioral patterns. They consider promoting God's will to be the only relevant evidence that an action falls under D, for example, so they tend to do something only when they think it promotes God's will, they encourage others to do the same, and so on. Referential normativity implies that *nonetheless* their action-guiding concept D refers to the set of actions that maximize pleasure! But why on earth think that? Pleasure may play no role in their practical reasoning—they might even be averse to pleasure if they believe God is. If so, pleasure fits their use of D no better than God fits my use of 'right' (I'm an atheist). Why on earth insist that *both* concepts have the same extension of pleasure-maximization nonetheless?

---

one would expect *on evolutionary grounds* that we also evolved a faculty to reliably track the standard (Sterelny and Fraser (2017) develop this point). Indeed, evolutionary psychologists emphasize our evolved capacity to detect cheaters (so as to punish them), high-status individuals (so as to align ourselves with them), and so on. What we probably didn't evolve is a faculty to detect which standards are normative joints, for it's hard to see how *that* would confer a selective advantage. For this reason I believe the real target of the epistemic argument is not moral realism but the thesis of normative joints. Indeed, that may have been Street's (2006) target all along, for she emphasizes in section 7 that her epistemic argument only targets a particular variety of moral realism. While she uses different terminology, I read her as referring to realism conjoined with the thesis of normative joints.

[18] This is, in effect, the point of the Hobbesian social contract. Even if individuals are invariably self-interested, a sovereign can change incentive structures in such a way that self-interested individuals make "altruistic" choices.

Perhaps because of Horgan and Timmons' (1991) "Moral Twin Earth" argument. They note that if we assert "A is right" and the above community denies it, we disagree. And they argue that to disagree, our concepts must have the same extension. But I reject the claim that disagreement requires sameness of extension. True, without sameness of extension there's no single content towards which we take opposing attitudes. Still, there can be disagreement over which standard to use in practical reasoning—indeed, this is what disagreement in cosmopolitan ethics is all about! Compare the parallel issue in science: when gruesome scientists call grue gemstones 'grue', we can object not because they said something *false* but because they use the wrong categories. The fact that we disagree about what 'right' applies to does not, therefore, show that our concepts share the same extension; if anything, it shows the opposite.

This is not meant to refute referential normativity, just to emphasize how strange it is.[19] But my second point is that it doesn't matter: even if it's true the cosmopolitan question still gets off the ground, just in a different guise. To see this, consider again this community that acts *as if* they use the divine standard: they systematically apply 'right' to actions that promote God's will and act on that basis. Referential normativity implies that their concept 'right' has the same extension of ours, which we're assuming is the set of actions that maximize pleasure. It follows that their judgments of what's 'right' are systematically false. Still, there's an interpretation of their concept that's more charitable to their usage. By an interpretation I mean a function from their concept to a set, so the charitable interpretation maps their concept to the set of actions that promote the divine will. Now, if referential normativity is true this set isn't the *extension* of their concept. Fine, call it its *dwextension* instead. Then the situation is this. Here we are, organizing our lives around the *extension* of 'right'; and there they are, using its *dwextension* instead. The cosmopolitan question is then *which way to live*; whether to use the extension or the dwextension of 'right' as a guide to life, the standard by which to evaluate action.[20]

This question might sound odd. We assumed that our judgments involving 'right' are *true* and theirs *false*, so how can there be a further question about whether to live by our judgments or theirs? But the question sounds odd only because of our unfortunate tendency to fetishize truth.[21] Remember, truth and extension go together: 'a is F' is true iff the referent of 'a' is in the extension of 'F'. So, there's a notion of *dwuth* that stands to dwextension just as truth stands to extension: 'a is F' is *dwue* iff the referent of 'a' is in the *dwextension* of 'F'. One and the same judgment has a truth-condition *and* a dwuth-condition at the very same time! Thus, *our* normative judgments are true but not dwue; *their* normative judgments are dwue but not true; and the cosmopolitan question is whether to use truth or dwuth as a guide to life.

At least, that's how things look on an inflationary view of extension and truth-conditions, on which they're fixed by some mixture of use and the world involving causation, naturalness, teleology, or what have you. For such mixtures are cheap: there are *innumerable* facts about the world that fix *innumerable* different extension-like relations and truth-like conditions. The cosmopolitan question is then which mixture to live by.

---

[19] Other arguments for referential normativity can be found in Wedgwood (2007), Dunaway and McPherson (2016), Williams (2018), and Sepielli (forthcoming).

[20] If referential normativity is true there's also an epistemic question of how we could *know* whether our judgments or theirs are true, and perhaps one can raise a cosmopolitan question in that guise instead. Eklund (2017) and Dunaway (manuscript) both explore epistemic issues in this vicinity.

[21] I discuss this fetish in Dasgupta (manuscript).

By contrast, deflationists claim that notions like truth and extension are just convenient logical devices of disquotation. To say that 'S' is true, on this view, is just to say that S; the two statements are "cognitively equivalent" as Field (1994) puts it. Likewise, "the extension of 'F'" is cognitively equivalent to "the set of Fs". What then are we to say about the alternative community above? For deflationists, the sentence

The extension of 'right' = the set of acts that are right

is trivial and analytic in our mouths as well as in theirs. The only further question is whether their term 'right' translates well as ours, and for deflationists this is just a pragmatic question of whether that translation suits the purposes and contexts at hand. The Moral Twin Earth argument tries to elicit a context in which we'd like to translate them as the same thanks to their same normative role. But in other contexts we might translate them differently due to their different pattern of application. Neither translation better reflects some prior fact about *their extension*, for according to deflationism there is no prior fact. All there is, at root, are two communities using different patterns of application to guide their decisions and actions. The cosmopolitan question, then, is which pattern to adopt.

In sum, it doesn't matter whether referential normativity is true. If it isn't, we can state the cosmopolitan question as I originally did, as which evaluative standard to use. If it is, we can state the question in one of the alternative manners just described. Either way, the question is *coherent*; and, I argued, one we must reckon with when it comes to the ethics of AI.

5. No middle ground

Turn now to premise (2), the claim that if there are no normative joints then all standards are on a par (egalitarianism). To unpack this, recall that a normative joint is a standard that has normative significance independently of us. If there are no normative joints then there are two possibilities: either something about *us* distinguishes certain standards as normatively significant (anthropocentrism), or *nothing* does and all standards are on a par (egalitarianism). What premise (2) states, then, is that anthropocentrism is not an option. I'll argue for this by showing not that it is false but that it collapses into egalitarianism.

My argument does not depend on what it is about us that is said to distinguish one standard over the others. It could be something about our values or other attitudes, some feature of our deep-rooted cultural history, or some immutable feature of our psychological or biological make-up. Whatever it is, anthropocentrism holds that *some* property P_1 of ours makes a standard S_1 normatively significant. OK, but since properties are cheap we'll inevitably have some other property P_2 that stands to a different standard S_2 just as P_1 stands to S_1. So, for S_1 to be normatively significant over S_2, P_1 must already have some normative significance over P_2. But P_1 cannot have this significance independently of us, for that's to say that P_1 is a normative joint and anthropocentrism rejects normative joints by definition. So, P_1 must be significant in virtue of some other property of ours, P_1*. But we'll inevitably have another property P_2* that stands to P_2 just as P_1* stands to P_1… and so on. The problem here is not the regress *per se*. The problem is that if nothing in the series

$S_1 \leftarrow P_1 \leftarrow P_1^* \leftarrow P_1^{**}\ldots$

is a normative joint, then nothing gives *that* series normative significance over the series

$S_2 \leftarrow P_2 \leftarrow P_2^* \leftarrow P_2^{**}\ldots$

Hence the collapse. If something about us distinguishes one standard as normatively significant, something *else* about us will distinguish another standard as normatively significant too. Without normative joints, *every* standard is distinguished by *something* about us! And as I tell my children in darker moments, if everyone's special no one's special. All standards are therefore on a par, *per* egalitarianism.

The views are not exactly the same, admittedly, for anthropocentrism has this bewildering array of regresses while egalitarianism has none. But what matters is that both views have the very same upshot for practical reasoning. As we saw in section 3, the egalitarian view that all *standards* are on a par implies that all *actions* are on a par too. Every action is correct by the lights of *some* standard, so if nothing distinguishes one standard as the one by which to evaluate actions, then nothing distinguishes one action as the thing to do either. Since anthropocentrism agrees that all standards are on a par, it implies that all actions are on a par for very the same reason. Thus, my claim is that anthropocentrism collapses into egalitarianism *for all practical purposes*: both views imply that no action is distinguished as the thing to do.

That's the argument in the abstract, but let me illustrate it with specific anthropocentric proposals. Suppose that pleasure-maximization is the normatively significant standard S_1, and consider the anthropocentric view that S_1 is significant because it's what we mean by 'good'. Our property of *meaning S_1 by 'good'* is then the property P_1 that makes S_1 significant. But meaning is just a relation between word and world, and relations are cheap. For any other standard S_2 such as the divine standard, our word 'good' also stands in *some* relation to it; call it the *dweaning* relation.[22] Thus, our word 'good' means S_1 *and* dweans S_2 at the very same time! Our property of *dweaning S_2 by 'good'* is then the property P_2 that distinguishes S_2 as special just as P_1 distinguishes S_1 special. One wants to say that meaning is more significant than dweaning, of course, but that requires saying that meaning is a normative joint. Without normative joints, there's nothing to distinguish S_1 over S_2: both are equally distinguished by *some* property of ours.

To take a second example, consider the view that S_1 is significant because it's the constitutive aim of action. The idea is that for a bodily movement to count as an *action* it must aim at maximizing pleasure; that's why S_1 is the standard by which to evaluate action. To see why this view is anthropocentric, note that it doesn't rule out aiming to meet the standard S_2 of promoting the divine will instead, it just implies that that wouldn't count as *action*. Thus, the view is presumably that S_1 is normatively significant only because *we act*—that's our property P_1 that makes S_1 special. But to say that we act, on this view, is just to say that we *aim at* S_1. And that's just a relation between us and S_1, and relations are everywhere. Thus, we also stand in *some* relation to the divine standard S_2, the "*dwaiming at*" relation—a single bodily movement aims at S_1 *and* dwaims at S_2 at the very same time! Call movements that dwaim at S_2 "dwactions". Then in addition to acting we also *dwact*, and this property P_2 distinguishes S_2 just as P_1 distinguishes S_1. Again, one *wants* to say that acting is more significant than dwacting, perhaps because the aiming-at relation is more significant than the dwaiming-at relation. But that's tantamount to

[22] If you equate meaning with extension, this is just the point in the last section that a word has an extension and an dwextension at the same time.

saying that acting and aiming-at are normative joints. Without normative joints, there's nothing to distinguish S_1 over S_2.

These examples are somewhat artificial, so as a final example consider Street's (2008) constructivist view on which normative facts are fixed by our normative attitudes. The way this works involves the constitutive natures of these attitudes. It's constitutive of the attitude of *valuing* something, says Street, that it aims at means-ends coherence. For example, if you value swimming and you believe that swimming requires a gym membership card, then you also value a membership card. If you really believed that swimming requires a card but you don't value the card, the idea is, it wouldn't be true that you value swimming after all. On Street's view, constitutive principles like these fix normative facts about what's valuable. If you value swimming and swimming *in fact* requires a membership card, then the card is valuable to you even if you don't *believe* it's required.

This is a view about our parochial notion of value. In effect, it says that the standard that determines whether something is valuable is a "Humean" standard involving our valuing attitudes. Our question is why this Humean standard is normatively significant. Street doesn't say, but she does say that her view is "thoroughgoing" in that it accounts for all normativity. Thus, she can't say that the Humean standard is a normative joint, for that entails primitive normativity (see section 3). Instead, she must either be egalitarian and say that all standards are on a par, or be anthropocentric and say that the Humean standard is significant in virtue of the fact that *we* happen to have valuing attitudes.

Once again, though, anthropocentrism collapses into egalitarianism. To see this, suppose I *say* that I value swimming, and suppose this is an honest report insofar as I have inner sensations that typically accompany valuing something such as feeling drawn to it. But suppose I know that swimming requires a card and I don't value the card. According to Street, I don't *really* value swimming because valuing something constitutively aims at means-ends coherence. Fine, but there's a different attitude of *shmaluing* that's just like valuing but lacks this constitutive aim when it comes to matters aquatic. After all, attitudes are cheap: an attitude like "valuing this" or "wanting that" is a property of individuals, and there's a property for every set. Thus, even if I don't value swimming I do *shmalue* it. But if my *valuing* attitudes fix truths about value *per* Street's constructivism, my *shmaluing* attitudes fix truths about *shmalue* in just the same way. My valuing attitudes imply that swimming isn't valuable, sure, but my shmaluing attitudes imply that it is *shmaluable*. In our terms, the Humean standard involving valuing attitudes determines whether its valuable, but a "Shumean" standard involving shmaluing attitudes determines whether it's shmaluable. Thus, if the Humean standard is significant because we have valuing attitudes, then the Shumean standard is equally significant because we also have shmaluing attitudes! Of course, one *wants* to say that the attitude of *valuing* is more significant that that of shmaluing, but that's tantamount to saying that valuing is a normative joint. Without normative joints, both attitudes and the standards they pick out are on a par, just as egalitarians said.[23]

To drive the point home, notice the upshot for practical reasoning. Suppose I'm wondering whether to buy swimming goggles. I discover that swimming isn't valuable, so I hesitate. But I also discover that swimming is shmaluable, so I reach for my wallet. Shall I buy them or not? Heard parochially, the answer is easy: I shouldn't buy them because swimming has no value.

[23] This is, in effect, Enoch's (2006) argument against constructivism.

But once ethics goes cosmopolitan, the question is whether to act on the basis of value or shmalue; that is, whether to use the Humean or the Shumean standard when evaluating the action of buying goggles. Whatever I do is correct by the lights of one of these standards, but since neither standard is distinguished over the other as the one by which to evaluate action, neither action (buying or not buying) is distinguished as the the thing to do either. Both actions are on a par and it is impossible to "go wrong".

The result is that when ethics goes cosmopolitan, Street's "thoroughgoing" constructivism is unstable—at least insofar as it claims to distinguish some actions as the thing to do. Being thoroughgoing, it must reject normative joints. But then the Humean and Shumean standards on a par, in which case all actions are on a par too and none is *the* thing to do.

That completes my argument for premise (2). Either there are normative joints, i.e. standards that are normatively significant independent of us, or all standards are on a par. There is no middle ground on which something about us determines which standards are significant.

6. Against normative joints

It remains to argue for (3), the claim that there are no normative joints. Here I am hindered by the fact I find the idea of normative joints baffling to begin with. If there were normative joints the sole aim of morality would be to reflect those joints, yet as Bernard Williams said 'the point of morality is not to mirror the world, but to change it' (1972, p. 49). Still, let me try to press this feeling into an argument.

Recall that normative joints would determine reality's preferred way of living. They'd determine what's *really*-right, for example, where this may differ from what's right in the parochial sense. They'd determine one's *real*-reasons and *real*-obligations, where again these may differ from one's reasons and obligations. The aim of morality—and practical reasoning more generally—would be to do the really-right thing (not the right thing), to act on the basis of real-reasons (not reasons), and so on.

Recall also that some normative joints can inherit their significance from others. If the divine standard of promoting God's will is a normative joint, that might be because God is a normative joint, which might in turn be due to God's infinite love or what have you. But as we saw in section 3, this must come to an end somewhere: some normative joints must be *primitive* joints, normatively significant in themselves and not in virtue of anything else. Thus, if there are normative joints, at least one of them N is such that (i) it helps determine what's really-right, the real-reasons and real-obligations, etc., yet (ii) there is nothing in virtue of which it plays that role.

Now, just look at how bizarre this is. By (i), N is a *remarkable* bit of reality. It fixes reality's preferred way of living regardless of the various parochial standards we've come to adopt—it's one normative ring to bind us all, as it were. Surely, you'd think, there must be some explanation of why N is so special, something about it that makes it fit to play this extraordinary role. But by (ii), there isn't. Nothing explains why real-reasons and real-obligations spring from *it* rather than anything else, it's simply a brute fact about the universe that they do. To see how bizarre this is, suppose someone said that N is my daughter. Their view is that real-reasons and real-obligations are fixed by what promotes *her* will, not God's. It follows that my daughter is *really* special: it's her will, not yours or mine, that fixes reality's preferred way of living! But if we ask

what makes her so special, there's no answer; there's nothing about her that makes her more fit to play this role than you or me. And that just seems ludicrous. What's ludicrous here is not just the idea that the preferred way of living is fixed by her (though that certainly is ludicrous). It's the idea that the preferred way of living could be so *arbitrary*. It's *reality's preferred way of living*, after all, something far too important for there to be no rhyme or reason why it is what it is.

I claim that the same goes whatever N is: it is no less ludicrous to claim of *anything*—God, pleasure, whatever—that it plays this remarkable role while adding that nothing makes it more fit to do so than anything else. But this is precisely what fans of normative joints must say.

Here I take myself to be recycling the well-known argument from "normative authority" against realism. Nowell-Smith ran it against non-naturalist realism, noting that if right and wrong were *sui generis* properties one could reasonably ask why they'd matter: "Why should I *do* anything about these newly-revealed objects? … [W]hy should I do what is right and eschew what is wrong?" (1954, p. 41). Likewise, Korsgaard said that realism about reasons 'invites the question of why it is rational to conform to those reasons' (1997, p. 240). They argued that realists have no answer to these questions and considered that a serious problem: there must be some explanation why we should do what's right and conform to reasons. My argument against normative joints is analogous. Why is it that I really-should do what's recommended by *N*, not some other aspect of reality? Why does real-rationality require conforming to N and not something else? The thesis of normative joints implies that there is no answer, and I too consider this a serious problem.

Admittedly, their argument is typically construed differently, as asking not why the moral properties themselves would matter but why beliefs about them would motivate action (see Drier 2015 for a discussion of this interpretation). The argument is then that realism has no answer to *that*, thereby violating an "internalist" principle that normative judgments are necessarily motivating. The ensuing debate centered around the internalist principle, with realists arguing that it is false or that they can respect it. I won't speculate which construal Nowell-Smith or Korsgaard had in mind, but one advantage of my construal is that it makes no appeal to internalism and hence sidesteps that debate.[24]

What might fans of normative joints say in response? Not that something explains why N plays its remarkable role after all, for their thesis implies that *some* bit of reality plays that role in virtue of nothing and N is whatever that is. Instead, they must try to show that the lack of explanation isn't ludicrous. And the obvious way to do that is to argue that there's no mystery why N plays this role and so nothing *to* explain in the first place. This idea can be made to appear more promising than it really is, so it's worth taking some time to see why it is, ultimately, an illusion.

To begin, suppose a fan of normative joints says:

> "Look, my view is that N is a *normative joint*. So of course it fixes real-reasons and real-obligations and the like—that's what normative joints do by definition! There's no mystery here, N just does what it says on the tin."

[24] I develop this alternative construal in Dasgupta (2017) as an argument against non-naturalism, but I subsequently realized that the argument is better pitted against the thesis of normative joints.

This is clearly cheating. To call something a "normative joint" *just is* to say that it fixes reality's preferred way of living—the real-obligations and the like. It's *because* N plays that role that it deserves the title. The question then remains why *it* rather than anything else plays that role. Just consider again the view that N is my daughter. It's *clearly* a good question why we really-ought to do what promotes her will rather than yours or mine. Calling her a normative joint doesn't remove the mystery, it just labels it.

Much the same goes for other attempts to remove the mystery. Korsgaard asked realists about reasons why, on their view, we should conform to reasons, and some replied that there's no mystery: it's an *obvious platitude* that we should conform to reasons! Now, whatever the merits of this reply to Korsgaard, it's irrelevant here because we're discussing a theory of real-reasons, not reasons. Real-reasons may differ from reasons, so it's certainly *not* obvious that we should conform to real-reasons. Still, a fan of normative joints might say it's an obvious platitude that we *really*-should conform to real-reasons. Could *this* remove the mystery? To see how it might, suppose their view is that N first determines what counts as a real-reason for a subject S to do action A, which in turn fixes what S really-should do—e.g. S really-should do whatever S has most real-reason to do. My question is then why *N* rather than some other bit of reality fixes this preferred way of acting. But the fan of normative joints might now say "Look, my view is that N is the *real-reasons* standard, the standard in virtue of which something counts as a real-reason for S to A, and it's an *obvious platitude* that we really-should conform to real-reasons. What's the mystery?"

But this is confused thrice over. First, it's not obvious that we really-should conform to real-reasons. Even if it's obvious that we should conform to reasons in the parochial sense of the terms, we cannot assume that their jointy correlates line up in the same way (that would be like assuming that obvious, parochial truths of geometry are preserved in the move to General Relativity). Second, even if it were obvious that we really-should conform to real-reasons, it only follows that we really-should conform to N if we add that N is the real-reasons standard. But that's just the above cheat in a new guise. On the theory of normative joints under discussion, to call something the real-reasons standard *just is* to say that it determines real-reasons and real-shoulds. That is, the statement

For any X, if X fixes real-reasons and real-shoulds, then X = the real-reasons standard.

is an analytic truth that fixes the referent of 'the real-reasons standard'. But our question is why N—my daughter, God, whatever it is—satisfies the antecedent clause in the first place. What makes it fit to play that special role? Calling it the '*real*-reasons standard' doesn't explain *why* it's special, it just *reflects* its specialness. And third, even if it were an obvious platitude that we really-should conform to N, that *still* wouldn't remove the mystery. For my question is not *whether* N fixes what we really-should do, but *what explains why* it does. It's obvious *that* water is a wetting agent, but that does not remove the chemical question of *why* it is. Likewise, even if it's obvious *that* we really-should conform to N, the question *why* N plays that role remains live and unaddressed.

Non-naturalists might dig in their heels here. "If N had a non-normative description involving your daughter or God, we could intelligibly ask why it plays this special role. But on our view N has no such description. We *start* with the concepts 'real-reason' and 'real-should'. We then discover that nothing natural could fix the real-reasons or real-shoulds; hence whatever fixes them, N, must be a *sui generis*, non-natural part of reality that we know only via the description

'that which fixes real-reasons and real-shoulds' *per* the reference-fixing stipulation above. There is then no intelligible question of why N plays this special role of fixing real-reasons and real-shoulds, for it was described as such from the start!"[25] But this is just more smoke and mirrors. For one thing, it ignores the possibility that *nothing* plays this special role—i.e. that there are no normative joints—in which case 'N' has no denotation. But more importantly, the speech at most establishes *that* N fixes real-reasons and real-shoulds, but that wasn't in question. Again, my question is not *whether* N plays this role but *what makes it* fit to do so. Nothing in this speech addresses or dissolves this question.

So far the mystery has not been removed, just buried in language. Perhaps more promising is the idea that there's no mystery why N plays this role because it lies in N's essential nature to do so. For example, suppose that real-reasons and real-shoulds are fixed by the utilitarian standard of pleasure-maximization. Then perhaps this lies in the nature of pleasure itself. On this view pleasure doesn't inherit its normative significance from something else, but nor is it *arbitrary* that it plays this role: the constitutive nature of pleasure leaves this unmysterious and in no need of explanation. This has a plausible ring, for isn't there a natural sense in which it lies in *the very experience of pleasure* that it really-should be maximized? Correspondingly, doesn't it lie in the nature of *pain* that it really-should be minimized? I suspect that non-naturalists would find this kind of explanation particularly appealing, for they often say that their *sui generis* properties are "essentially normative".[26]

As tempting as this idea is, however, it doesn't stand up to scrutiny. Focusing on pleasure, the relevant claim is this:

(*) It is essential to pleasure that one really-should do what maximizes pleasure.

But what does this talk of "essence" mean? There are three salient readings. On one reading, essence is just necessity: (*) means that it's *necessary* that one really-should do what maximizes pleasure. But this can't be the intended reading, for something's being necessarily true doesn't remove the question why it's true. To see this, suppose instead that the divine standard were the normative joint, so that one really-should do that which promotes God's will. If this were true I suspect it would be necessarily true, but that doesn't explain *why* the divine will is special.

A second reading interprets talk of essence in the model of definition. Just as words have nominal definitions, the idea is that worldly entities have real definitions.[27] On this reading, (*) states that pleasure is *by definition* that state such that one really-should do what maximizes it. But this can't be the intended reading either. For on the utilitarian view under discussion, pleasure is supposed to be that which explains real-shoulds: an action A is what one really-should do iff, *and because*, it maximizes pleasure. First come the facts about pleasure, as it were, and then they fix what one really-ought to do. But the current reading of (*) has it exactly the other way around, since it says that pleasure is *defined* in terms of real-oughts. This point is most perspicuous if we consider the non-naturalist view that N is not pleasure but an irreducible non-natural property. To say that N is irreducible is to say that it has no definition, so this non-naturalist obviously cannot say that it's essential to this property that one really-ought do what it

[25] Chappell (2019) offers this kind of response to Korsgaard's normative question.
[26] See for example Chappell (2019).
[27] Fine (1994) champions this reading of essentialist talk.

favors—at least, not if essence is understood in the model of definition. But the point stands whatever N is said to be. For N is, by hypothesis, that which *fixes* real-shoulds; hence N cannot itself be *defined* in terms of real-shoulds else we'd have gone in a circle.

The third reading of essence is the more Scholastic idea that things have an "inner nature" responsible for their attributes. A sleeping pill causes sleep, for example, because its inner nature contains a "dormative virtue". Likewise, we might read (*) as stating that pleasure has an "inner nature" that renders it fit to fix real-shoulds. But I urge us to reject this style of "explanation" as a pernicious remnant of Scholastic metaphysics. No one today explains the effect of a sleeping pill just by positing some "inner nature" that is stipulated to produce the effect. The point is not that there *couldn't* be such a thing, but that it's a black-box; a label for an explanation-to-be-filled-in-later, not an explanation itself. Intellectual progress since the scientific revolution has taught us to reject Scholastic "explanations" such as these and I encourage us to heed this lesson here.

For these reasons, the idea that N is normatively significant because of its essential nature strikes me as unpromising. Still, the discussion has been abstract and one might remain pulled by the idea that there is something intrinsic to the nature of *pain* that makes it a normative joint, something abhorred not just by us but by reality itself.

I suspect this pull is largely due to the fact that pain is typically accompanied by two things: (i) a phenomenal experience, and (ii) various negative attitudes such as fear, intense desire for self-protection, and so on. It's controversial whether 'pain' denotes just the phenomenology or something involving the attitudes too. It's also controversial whether it's possible to have the one without the other—perhaps the phenomenal experience consists partly in the negative attitudes, or vice-versa. But regardless, insofar as we typically experience pain along with these negative attitudes, it's unsurprising that we so find it so natural to think that there's something intrinsically terrible about pain.

But this is, ultimately, a mistake. Consider first the attitudes themselves. If you say that *they* are the normative joint N, what this really amounts to is a Humean view on which real-oughts and the like are determined by the standard of desire-satisfaction. This Humean view can be assessed in its own right, of course, but the idea under discussion—that real-oughts are fixed by the utilitarian standard involving pleasure and pain—has fallen away. And insofar as the phenomenology can be separated from the attitudes, it's hard to see how it could be a normative joint either. Consider an AI system for whom this phenomenology is systematically accompanied by positive attitudes: it finds relief and value in a life full of it, it wants more of it, and so on. For this AI, the phenomenology plays the normative role that pleasure plays in us! It's hard to see why they really-ought minimize this phenomenology nonetheless.[28]

This then is my objection to normative joints. If there were such things, they'd play the remarkably important role of determining reality's own preferred way of living. Yet there'd be no explanation of why they play this role; it would be a brute fact that the preferred way of living

[28] This paragraph is, in effect, a condensed summary of Street's argument that pain is not an objective bad (2006, section 9), transposed into the setting of normative joints. Lee (forthcoming) argues for a similar conclusion from a different angle. On his view, conscious states like pain are not *descriptive* joints and therefore cannot be normative joints either.

springs from them and not anything else. Thus, normative joints render morality so incredibly *important* yet at the same time completely *arbitrary*—two ideas that strike me as incongruent.

7. Conclusion

There we were, wondering what to do with Hal. Shall we destroy him? It's easy to slip immediately into parochial thinking, asking whether Hal has *moral status*, whether there are *reasons* to treat him well, and so on. But these standards that suited us in past epochs might be unsuited to a world of AI. Instead, I argued that we must ask the cosmopolitan question of which standard to use in the first place (premise (1)). I then argued that all standards are on a par: there are no normative joints (premise (3)), and hence nothing to distinguish one standard as normatively privileged over the rest (premise (2)). It follows that all actions are on a par too: each is correct by the lights of some standard, and that is all there is to say. Hence this question of how to treat Hal has no answer out there to discover; it is more a matter of choosing a way of life.

This doesn't mean that we must simply toss a coin. As emphasized in section 3, choosing how to live apparently matters to creatures like us, so—as a matter of empirical fact—we will likely consider the alternatives carefully and choose whatever strikes us as the best solution to the problem at hand. But in this respect, it's no different than other creative acts such as inventing a new sport or a style of music. The point is that the ethics of AI is a creative act akin to these. It's something to be *made*; there are no pre-existing answers to discover.

Let me finish with three general remarks about this line of argument. First, fans of normative joints might object to my egalitarian view on the grounds that it leaves out the importance of morality. This mirrors a common objection to error-theory about parochial ethics: if it's just system of falsehoods, why care about *those* falsehoods over others?[29] Similarly, without normative joints to privilege one standard over the rest, why care about the one we happen to live by? But I think this objection over-states the importance of morality. Regarding parochial ethics, the data is just we *happen* to take it seriously, and error theorists can explain this psychological fact in any number of ways—perhaps it's evolutionarily advantageous for us to do so. It doesn't follow that we *must* take it seriously or that we'll continue to do so after discovering that it's false, but that was never part of the data. Likewise, one doesn't need normative joints to explain the psychological fact that we care about the standard we've come to live by.

Second, my argument against normative joints was "metaphysical" in that it involved considerations about explanation and essence and the like. I therefore used metaphysics to address the ethical question of how to treat Hal, i.e. to show that the answer is up to us. So-called quietists would object, for they think that first-order ethical questions can be addressed only by arguments that are "internal" in the sense of resting on first-order ethical premises. Now, strictly speaking my argument from (1)-(3) *is* internal because (1) is an ethical premise (it states that AI ethics *should* be cosmopolitan).[30] But quietists would reject my metaphysical argument

[29] Boghossian (2006a) develops this line of thought; see also Enoch (2011) chapter 5.

[30] In this regard, Dworkin (1996) emphasizes that quietists can accept the argument that "morality is empty because there is no God" since it "presupposes the substantive view that a supernatural will is a plausible and the only plausible basis for morality" (p. 91). His point, I think, is that this "substantive view" is the *ethical* claim that something is morally right iff God wills it; hence if there's no God then nothing is morally right. But my argument from (1)-(3) has exactly the same structure. To see this, note that premises (1) and (2) imply an analogous "substantive view" that normative joints are the only basis for AI ethics; hence if there are no normative joints then AI ethics is "empty".

against normative joints, for they think that meta-ethical views are really claims of first-order ethics in disguise, in which case arguments for meta-ethical views must also be internal to ethics—arguments from metaphysics don't cut the mustard. My argument against normative joints certainly does not clear this bar!

I cannot fully address this worry here, but let me just observe that it would be surprising if arguments against normative joints must be internal. Consider the analogy with sport. There are no sporting joints—no human-independent facts that privilege the current off-side rule over past iterations. This is obvious, but you couldn't possibly establish it on the basis of the rules of soccer themselves! Arguments against sporting joints would clearly be "external" to sport, appealing to science or metaphysics or what have you. Why would normative joints be different? Perhaps because the thesis of normative joints just is a claim of first-order ethics, quietists might say. But this would be equally surprising. The claim that there are sporting joints is obviously not itself a rule of soccer, so—again—why would the thesis of normative joints be different? Here quietists might say that the thesis of normative joints is unintelligible unless construed as a claim of first-order ethics. Well, they can *say* that, but it hardly advances matters. Claims of unintelligibility have a habit of remaining intractable, especially when skeptics are skilled at misunderstanding their interlocutors. I explained the thesis of normative joints in section 3; if quietists don't understand it the problem may not be mine. More tractable would be a clear statement of where my particular external argument against normative joints runs afoul. In the absence of this, I'm unmoved by their insistence that *no* external argument can succeed.[31]

My third remark concerns the worry that any argument against moral realism is self-defeating, at least if 'morality' is interpreted to encompass normativity writ large. After all, an argument is supposed to be a *reason* to believe the conclusion, yet if realism is false there are no mind-independent facts about what counts as a reason. Why then take the argument against realism seriously? As Dworkin puts it, opponents of realism need

> "some place to stand… They must assume that some of what they think (at an absolute minimum, their beliefs about good reasoning) are not just their own or their culture's invention but are true or valid—indeed, "objectively" so. Otherwise they could only present their views as "subjective" displays in which we need take nothing but a bibliographic interest" (1996, p. 88).

Likewise, if there are no normative joints then there are no human-independent facts that distinguish reasons as normatively significant over dweasons, so one might complain that an argument presenting *reasons* against normative joints undermines its claim to be taken seriously. But this venerable idea rests on a confusion. As Boghossian (2006b) pointed out, it ignores the fact that the argument (against realism or normative joints) might well appeal to inferential rules that all parties to the debate accept. If the argument just used modus ponens, for example, it would be a strange opponent indeed who refuses to take it seriously! So again, I ask: where exactly does my argument against normative joints go awry?[32]

---

[31] In his defense, Dworkin (1996) does criticize a number of particular external arguments against moral realism, including Mackie's (1977) arguments from disagreement and queerness, epistemic arguments that mind-independent moral truths would be unknowable, and psychological arguments that such truths would conflict with known facts about the human psyche. But my argument against normative joints falls into none of these categories, and Dworkin doesn't indicate how his objections might generalize.

[32] Acknowledgements…

References

Ashrafian, Hutan (2015). “Artificial Intelligence and Robot Responsibilities: Innovating Beyond Rights”. *Science and Engineering Ethics* 21: 317-326.

Basl, John (2013). “The Ethics of Creating Artificial Consciousness”. *APA Newsletter* 13(1): 25-30.

Boghossian, Paul (2006a). “What is Relativism?” In *Truth and Realism*, edited by Lynch and Greenough, pp. 13-37. Oxford: OUP.

Boghossian, Paul (2006b). *Fear of Knowledge: Against Relativism and Constructivism*. Oxford: OUP.

Bostrom, Nick and Eliezer Yudkowski (2011). “The Ethics of Artificial Intelligence”. In *Cambridge Handbook of Artificial Intelligence*, edited by W. Ramsey and K. Frankish. Cambridge: CUP.

Chappell, Richard Yetter (2019). “Why Care About Non-Natural Reasons?” *American Philosophical Quarterly* 56(2): 125-134.

Cheok, David (2017). “Lovotics: Human-Robot Love and Sex Relationships”. In *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*, edited by P. Lin, K. Abney, and R. Jenkins, pp. 193-213. Oxford: OUP.

Cosmides, Leda; Ricardo Andres Guzman, and John Tooby (2019). “The Evolution of Moral Cognition”. In *The Routledge Handbook of Moral Epistemology*, edited by Aaron Zimmerman, Karen Jones, and Mark Timmons. New York: Routledge Publishing.

Darling, Kate (2016). “Extending Legal Protections to Social Robots: The Effects of Anthropomorphism, Empathy, and Violent Behavior Towards Robotic Objects”. In *Robot Law*, edited by M. Froomkin, R. Calo, I. Kerr, and E. Elgar.

Dasgupta, Shamik (2017). “Normative Non-Naturalism and the Problem of Authority”. *Proceedings of the Aristotelian Society*, CXVII (3): 297-319.

Dasgupta, Shamik (manuscript). “Undoing the Truth Fetish: The Normative Path to Pragmatism”.

DiGiovanna, James (2017). “Artificial Identity”. In *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*, edited by P. Lin, K. Abney, and R. Jenkins, pp. 307-321.

Dreier, Jamie (2015). “Can Reasons Fundamentalism Answer the Normative Question?” In *Motivational Internalism*, edited by G. Bjornsson, C. Strandberg, R. Francen Olinder, J. Eriksson, and F. Bjorklund, pp. 167-81. Oxford: OUP.

Dunaway, Billy (manuscript). “Realism, Meta-semantics, and Risk”.

Dunaway, Billy, and Tristram McPherson (2016). “Reference Magnetism as a Solution to the Moral Twin Earth Problem”. *Ergo* 3(25): 639-79.

Dworkin, Ronald (1996). “Objectivity and Truth: You’d Better Believe It”. *Philosophy and Public Affairs* 25(2): 87-139.

Eklund, Matti (2017). *Choosing Normative Concepts*. Oxford: OUP.

Enoch, David (2006). “Agency, Shmagency: Why Normativity Won’t Come From What Is Constitutive of Action”. *The Philosophical Review* 115(2): 169-198.

Enoch, David (2011). *Taking Morality Seriously: A Defense of Robust Realism*. Oxford: OUP.

Field, Hartry H. (1994). “Deflationist Views of Meaning and Content”. *Mind* 103: 249-285.

Fine, Kit (1994). “Essence and Modality”. *Philosophical Perspectives* 8: 1-16.

Fitzpatrick, William J. (2015). “Debunking Evolutionary Debunking of Ethical Realism”. *Philosophical Studies* 172: 883-904.

Gibbard, Alan (2003). *Thinking How to Live*.

Greene, Joshua (2007). “The Secret Joke of Kant’s Soul”. In *Moral Psychology: Volume 3*, edited by Walter Sinnnot-Armstrong. MIT Press.

Goodman, Nelson (1955). “The New Riddle of Induction”. In *Fact, Fiction, and Forecast*, pp. 59-83. Cambridge, MA: Harvard University Press.

Henrich, Joseph (2015). *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter*.

Horgan, Terence and Mark Timmons (1991). “New Wave Realism Meets Moral Twin Earth”. *Journal of Philosophical Research* 16: 447-465.

Korsgaard, Christine M. (1997). “The Normativity of Instrumental Reason”. In *Ethics and Practical Reason*, edited by G. Cullity and B. Gaut, pp. 215-54. Oxford: OUP.

Korsgaard, Christine M. (2008). “Realism and Constructivism in Twentieth-Century Moral Philosophy”. In *The Constitution of Agency*, pp. 302-326. Oxford: OUP.

LaBossiere, Michael (2017). “Testing the Moral Status of Artificial Beings; or “I’m Going to Ask You Some Questions…” In *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*, edited by P. Lin, K. Abney, and R. Jenkins, pp. 293-306. Oxford: OUP.

Lee, Geoffrey (forthcoming). “Alien Subjectivity and the Importance of Consciousness”. In *A Festschrift for Ned Block*, edited by A. Pautz and D. Stoljar.

Lewis, David (1984). “New Work for a Theory of Universals”. *Australasian Journal of Philosophy* 61(4): 343-377.

Levy, David (2007). *Love and Sex with Robots: The Evolution of Human-Robot Relationships*. New York: Harper Collins.

Mackie, John L. (1977). *Inventing Right and Wrong*. Penguin.

McCullough, Michael E. (2020). *The Kindness of Strangers: How a Selfish Ape Invented a New Moral Code*. New York: Basic Books.

Nowell-Smith, P.H. (1954). *Ethics*. Harmondsworth: Penguin Books.

Richerson, Peter J. and Robert Boyd (2005). *Not By Genes Alone: How Culture Transformed Human Evolution*. The University of Chicago Press.

Schwitzgebel, Eric and Mara Garza (2015). "A Defense of the Rights of Artificial Intelligences". *Midwest Studies in Philosophy*, XXXIX: 98-119.

Sepielli, Andrew (forthcoming). "Quietism and Counter-Normativity". *Ergo*.

Shafer-Landau, R. (2012). Evolutionary debunking, moral realism and moral knowledge. *Journal of Ethics and Social Philosophy* 7: 1-37.

Sider, Ted (2011). *Writing the Book of the World*. Oxford: OUP.

Skyrms, Brian (2014). *Evolution of the Social Contract*. Cambridge: CUP.

Sterylny, Kim and Ben Fraser (2017). "Evolution and Moral Realism". *The British Journal for the Philosophy of Science*, 68: 981-1006.

Street, Sharon (2006). "A Darwinian Dilemma for Realist Theories of Value". *Philosophical Studies* 127(1): 109-166.

Street, Sharon (2008). "Constructivism About Reasons". In *Oxford Studies in Metaethics: Volume 3*, edited by R. Shafer-Landau, pp. 207-245. Oxford: Clarendon Press.

Sullins, John (2012). "Robots, Love, and Sex: The Ethics of Building a Love Machine". *IEEE Transactions on Affective Computing* 3(4): 398-409.

Talbot, Brian; Ryan Jenkins; and Duncan Purves (2017). "When Robots Should Do the Wrong Thing". In *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*, edited by P. Lin, K. Abney, and R. Jenkins, pp. 258-273. Oxford: OUP.

Tomasello, Michael (2016). *A Natural History of Human Morality*. Harvard University Press.

Wedgwood, Ralph (2007). *The Nature of Normativity*. Oxford: Clarendon Press.

Williams, Bernard (1972). *Morality: An Introduction to Ethics*. Cambridge: CUP.

Williams, Bernard (1985). *Ethics and the Limits of Philosophy*.

Williams, J. Robert G. (2018). “Normative Reference Magnets”. *The Philosophical Review* 127(1): 41-71.

Ziaja, Sonya (2011). “Homewrecker 2.0: An Exploration of Liability for Heart Balm Torts Involving AI Humanoid Consorts”. In *Social Robotics*, edited by S. Sam Ge, O. Khatib, J.J. Cabibihan, R. Simmons, and M.A. Williams, pp. 114-24. Berlin: Springer.