Base Camp for Mt. Ethics

Nick Bostrom[1]
Future of Humanity Institute
University of Oxford

Contents

Metametaethics/preamble

1. Many traditional debates in metaethics, such as between realism and antirealism, may not carve the possibility space of nature at its joints. Even if one perfectly understood ethics, it might be unclear which side would be the winner in these debates, or the adjudication might turn out to depend on seemingly minor and semi-arbitrary definitional stipulations that were made along the way.

2. There is not a sharp separation between morality, law, and custom. These may be broadly the same sort of thing, albeit with some potentially important differences in emphasis, including, but not limited to:

    a. Custom vs. morality. Policed by distinct kinds of passion: externally, morality enforced with anger and indignation; custom with disdain. Internally, morality is enforced by feelings of remorse; custom with embarrassment.

---

b. Law vs. morality. Law is typically codified and enforced by formal institutions; morality is typically not codified (although different authors may have various theories and claims), and is enforced decentrally by informal institutions and social sentiment.

    i. If law were perfect and comprehensive, there would be little or no need for morality. This may become increasingly feasible with technological advances.

    ii. (Are modern societies with well-functioning legal systems and stable high-capacity states less moralistic than societies that are less technologically advanced or more anarchic?)

    iii. I'm aware of the descriptive/normative distinction that is commonly made. The approach taken here may seem to ignorantly blur it, because it does not start by taking this distinction for granted.

    iv. Underlying both law and morality there is a deeper layer—abstract properties of arrangements that regulate conduct. Law and morality are roughly symmetric with respect to this layer.

Genealogy

3. You build an internal model that warns you if you're about to treat somebody in a way that will make them angry at you if they find out what you did. The model might also be used to scan for situations where you want to react with anger towards somebody else.

4. People's friends may also get angry at you: so predict their reactions too.

5. Within a group, norms develop through implicit or explicit agreement. Powerful or prestigious individuals and coalitions have greater say in this process. The norms include requirements to impose sanctions on violators (e.g. negative gossip, exclusion, etc.).

6. The possibility arises of disagreement (and uncertainty, especially among third parties) as to whether a norm violation has occurred. Facts, logic, and persuasion (and rallying of potential allies) will be usual in case of dispute. There can also be disputes about exactly what the norms are.

7. You could now have "moral philosophers" doing some work:

    a. Articulating implicit norms

    b. Describing the scope of different norms and how they interact (in general and/or with reference to particular real or hypothetical cases)

    c. Propose new norms or revisions to existing norms

        i. If these are merely clarifications of implications of existing norms, then this could theoretically be a purely epistemic pursuit (though in reality things are rarely pure).

        ii. If they involve more basic change, then logic and empirical inquiry is not enough. You would instead need to build a dominant coalition to favor and undertake the change (cf. "norm invasion", below)

8. Sometimes a norm gets adopted that has (perhaps non-obvious) implications that are not favorable to the power-weighted interests of the group that adopts it. There are several ways this can happen:

    a. Norm invasion. A norm is initially adopted locally by a subset of the group, and its supporters gradually expand its reach by means of targeted blaming.

        i. They may also appeal to other already existing norms for support, but usually they also need to create a new blame field in order to win supporters.

        ii. Alternatively, one could imagine a more explicit proposal to the general public (the group as a whole): "Let's change norm N to N*!"

1. Again, the case for the reform can be buttressed with appeals to existing norms, or with appeals to alleged benefits to the common good from adopting the revised norm; but usually this is not enough to win. You also have to convince people that they'd be better off supporting the change than resisting it. There are many reasons why somebody might benefit from supporting a norm change other than that the norm change would benefit the group—for example, supporting the change might signal positive qualities.

   b. Shortsightedness or mistake. The implications of a norm may become apparent only later, when real or epistemic circumstances change

      i. For example: "Share equally with all!" or "Help anyone in need!" might initially benefit the group; but later, when a more literalistic interpretive metanorm is adopted (see below), or when the group comes into contact with large numbers of other people who can be helped at individually low cost, the norm turns out to have weird consequences.

9. It is also possible for norms to be adopted that regulate how discussions around potential norm changes should be conducted.

   a. For example, there may be a norm of shaming anybody who questions an existing norm or who points out any downside to it.

   b. Almost every group seems to have the norm that its enemies are morally bad, and that it is highly suspect to to point to any favorable trait of the enemy.

10. Because of 8 and 9, the norms that some group has may be harmful to that group relative to some other set of possible norms (though it is also true that many alternative norms that appear better would actually be worse, even if the switching cost could be ignored).

11. All the above processes can also take place (in analogous form) within an individual. An individual may internalize norms—invest them with final value or normative force, independent of and additional to the instrumental/social reasons that exist for complying. An individual could accept norms erroneously believed to be the norms of their community, or norms (precepts, commitments) that they know are not based on community norms. She can also deliberate on the scope and application of these norms, different parts of her can in some sense bargain or fight over norm changes, and so on.

    a. In deciding how to act and how to feel, an individual may consciously blend in other values and considerations (not represented by the "normative commitments" of the superego). In some moralistic persons, however, their superego might have adopted a norm that prohibits such harmonious blending. In such subjects, other solutions might still avail, such as hypocrisy, self-serving bias, akrasia, etc.

       i. Cf. countries and organizations with overly rigid or monomaniacal governance regimes—it may become necessary for underlings to hide certain information from the top, or for everybody to make a pretense of following various edicts and policies while in actuality taking a more pragmatic approach, etc.

Metaethics

12. What are the moral norms that exist in a society? They are not exactly the rules that people would affirm as moral norms: for people can be mistaken in their beliefs, and they can misarticulate their beliefs. Nor are the moral norms the rules that characterize the exact actual response patterns to ostensible violations. Rather, *the moral norms are the rules that characterize what the response patterns would be if they were slightly idealized ("extrapolated" a short distance).*

    a. For example, performing action A in society S may result in condemnation. However, if this condemnation depends on a straightforward error of empirical fact, then it would not reflect a

real moral transgression and the pattern of such condemnations would not constitute a moral norm.

    b. It is useful to model a slightly idealized version of the actual pattern of moral attitudes, because (a) the idealized pattern is simpler and more systematic than the actual pattern, and so modeling it is more feasibly predictive than focusing on every detail of the actual pattern and (b) if your actions are criticized, you often have the option of explaining yourself or arguing in your own defense—e.g. by pointing out clear empirical errors or inconsistencies in your accusers' initial stance—thereby potentially getting them to change their attitude.

        i. Cf. grammar.

    c. On the other hand, keeping track of very idealized response patterns (very long "extrapolation distances") will usually not be helpful. If the people who condemn you would change their minds if they reflected carefully for a million years and learned all non-normative facts and were much smarter than they actually are, and so on—well, normally you don't have the ability to realize these conditions. So you still get a bucket of filth poured over your head.

13. Since moral motivation becomes internalized, and since each person constructs a model of slightly idealized norms that defines the aim of these moral motivations, it is possible for a person to end up in a situation where she feels morally obligated to do X even though she predicts that doing X will result in her becoming unpopular in her relevant community or peer group. Though this doesn't happen very often (probably a good thing too, I would add.)

14. So moral realism (on some definitions, though not on others) may be true if there are sufficiently well-defined facts about patterns of slightly extrapolated moral praise and condemnation.

15. We can see how classical theism tends to be supportive of objective morality.

    a. Case 1: Most everyone in a community shares the same faith, which involves believing that there is a divine command to live by norms M. They therefore have strong subjective reasons to collectively adopt norms M, so they do. Now there is a clear pattern of relevant facts that our moral modeling can track.

    b. Case 2: An individual, Dissenter, has a faith that diverges from that of her community, and she bases her moral opinions on this non-standard conception, whereas the community bases its beliefs and practices on the standard conception. Here there are several possibilities:

        i. Dissenter is morally mistaken. (It may or may not be possible to help her realize this by reasoning with her.)

        ii. Her community is morally mistaken. Dissenter might correctly assert that there is a slight extrapolation whereby the community would come to recognize that her conception of theological matters is correct and that the community would then interpret its existing norms and moral commitments in a new way that conforms with that of Dissenter.

        iii. There is no fact of the matter as to whether Dissenter is correct. Maybe this could happen if, after slight extrapolation, the community would split into two roughly equal factions that fundamentally disagree about the matter.

        iv. There is another possibility as well, which could make it so that Dissenter is in a sense right even if her community slightly extrapolated would come to the opposite conclusion: we'll return to this later, under the heading "higher morality"

16. We can also see cultural variation is undermining of objective morality. Surface level disagreement is prima facie evidence of deeper disagreement (that would survive slight extrapolation). Perhaps an even bigger problem for objective morality is incoherence in or lack of actual moral behavior.

17. Suppose that one community (B) has the norm "Favor blue people!" and another community (G) has the norm "Favor green people!". Furthermore, let's suppose that you and I are members of B, the blue-favoring community (which, it is safe to assume, is the society where people are blue).

a. It is possible that blue-favoring is morally right in B and wrong in G. Different communities can have different standards.
   i. For example, maybe *everybody* should keep their promises, but we should also be intellectually honest. Many professional groups have special ethical norms.
   ii. Individuals may also develop "micronorms" that are specific to their own personal situation and which govern expectations around how they conduct themselves and perform their various social roles.

18. What if community B has the norm "Everybody ought to exclusively favor B!" and community G has the norm "Everybody ought to exclusively favor G!"?
   a. There is clearly a potential conflict: if the groups come into contact, they will be trying to enforce conflicting norms.
   b. There also appears to be a factual disagreement: it cannot be the case that both of them are correct.
   c. One path we could take, if the groups are relatively separate: "When we use moral language, its referents are determined by our usage. It is true that "Everybody ought to exclusively favor B!". It is also true that "Everybody ought* to exclusively favor G!", but ought* is a concept that is of little concern to us (other than insofar as our anthropologists are seeking to understand this alien G community)."
      i. In some cases—but not if the two communities are more or less isomorphic in relevant respects—one could also point to "constitutive functions" of morality as a basis on which to reject some ostensibly moral assertions. If the function of morality is, very broadly speaking, to regulate and harmonize social interactions in ways that improve efficiency, then one could maintain that a putative moral norm that does not even come close to engaging constructively in this function, is either defective and false, or is not even moral (leaving open the possibility that it is instead some other kind of norm, e.g. aesthetic).
   d. Another path we could take, which is more recommended if B and G are similar enough and interacting enough that it would be inconvenient to stipulate that their moral vocabularies have different meanings: "There is a real factual disagreement, and at most one community is correct. If the views of one group would prevail under slightly idealized conditions, that view is correct. If neither view would prevail under slightly idealized conditions, both views are wrong." (There is also the possibility that one of the communities may be right with reference to higher morality; more on that later.)
   e. If B and G have to interact a lot, they could try to develop a revised set of norms that they could share and that would govern both communities (or at least govern how the two communities interact). This would be a good idea. It probably requires some compromise, but some convergence can also happen as a result of short extrapolation or as a result of various other processes by which norms change over time (e.g. norm invasion, or changes in accordance with morally legitimate or required normative developments).

19. Comparison to some forms of contractarian approaches:
   a. Persons are not atomic. When we look more closely, we can see that persons have internal structure, different parts, which need to be coordinated: morality might apply intrapersonally as well as interpersonally.
   b. Other actors. Not only persons or person parts, or sentient organisms, but other types of entities can also be moral actors and help shape the content of moral norms: e.g. institutions, cultural tendencies, networks, AI systems, etc.
   c. Dynamic process. Not a once-and-for-all real or hypothetical deal that determines the true facts of morality based on some fixed set of interests, but an unfolding process where influ-

ences wax and wane, continuously negotiated possibly in accordance with procedural criteria that can themselves be subject to moral constraints.

  i. However, there might occasionally be rigidifications whereby a community's language use fixes some moral terms as rigid designators of timeless facts: then the language may change, but the content of morality (insofar as they make assertions about it using their terminology) remains fixed.

  ii. There might be some moral structures that are unchanging and more independent of developments and of the interests and actions of various actors; cf. higher morality.

Value representors

20. The moral norms prevailing in one's community affect how other people react to various patterns of behavior (as well as various patterns of attitudes or character traits, which produce observable physiological or behavioral manifestations). It is therefore important for an individual to accurately model these norms, and the patterns of blame and approbations on which they are based. Many different types of representation can be used in these models to generate the predictions. Examples include:

    a. Properties of consequences. Utilitarians, for instance, make a big song and dance about consequences being the determinant of moral rightness. Has the downside that consequences are generally unknowable, except for the most direct and immediate consequences.

    b. Act-types. Is the thing an act of type "lying", "stealing", "murdering", etc.? This is more promising, but sometimes the wider consequences should be considered.

    c. Motives. Did it spring from a certain type of motive, such as from selfless love or a sense of duty? This can be a useful consideration, but it is not uncommon for well-intended people to do bad things. Also, true motives are often difficult to ascertain.

    d. Paragons, or characteristic behavior styles. What would a certain person, who is a paragon of virtue, do in this situation? This type of value representor has a number of advantages:

      i. It can make a choice in most situations, since the paragon would do so. By contrast, many consequentialist theories deliver a nil result due to problems of infinitarian paralysis; deontological theories can be stymied by decision-making under uncertainty; and motive-focused theories may likewise be stumped if there are many different ways of acting on some given approved motive in a situation. (On the other hand, unless the paragon is a living person who is available to answer questions and has the time and ability to understand the situation, it may be epistemically difficult to determine what the relevant facts are.)

      ii. It can definitely match human preferences. One can in fact trust some people to make reasonably nice choices in many situations, so a value-representor of this type should be able to characterize roughly what we are after.

      iii. A good person might, for example, be extra concerned to help somebody who is in her immediate presence or to promote some value that has come to her active attention, while at the same time being somewhat resistant to attempts to manipulate this disposition (e.g. avoiding money pumps).

      iv. "Love God and the person in front of you at this moment" makes sense as an ethic to live by.

21. Many schools in normative ethics essentially consist of picking one of these and asserting that it is the correct one. I don't see a reason why our model should only use one representation format. In any case, this does not seem to be among the most important issues.

22. Beside the value-representor, we can also consider what type of things are evaluated. For example, one could evaluate acts, persons, persons-at-a-time, institutions, practices, etc. Again, I don't see a reason for not doing all of the above. If assigning moral blame or credit to some type of entity X (sometimes) has the potential to influence either X's behavior, or other actors' morality-focused behavior in matters related to X, then there can be a point to morally evaluating X.

23. We can also consider the question of value representors from an implementation perspective. In the human case, it seems that we can distinguish two importantly different implementations: *a state-action value function* and *a more abstract model of moral considerations*.

    a. One's state-action value function attaches very low value to pushing Fat Man.
    b. One's abstract ethical model may or may not recommend pushing Fat Man.
    c. We may conjecture that the state-action value function is mainly useful for controlling behavior (and for predicting how others will act), while the abstract ethical model is mainly useful for arguing with other people about norms, norm interpretations, and norm applications in cases of dispute.

Moral motivation

24. Moral motivation is often internalized, meaning that it functions even in the absence of immediate extramoral instrumental incentives, and that it is taken as a final value / as directly reason-giving on its own to act in a morally adequate way, or that outcomes prescribed by morality as ones we should promote are valued as final ends.

25. We may indeed be morally required to internalize some moral motivations. In this case we would be morally defective if we conformed our actions to behavioral moral norms only for instrumental reasons (such as out of fear of punishment). We ought instead to be motivated either (on some views) by an abstract sense of duty or (on other views) by valuing the various conditions that moral norms demand that we treat as valuable (such that we e.g. help someone in need because we intrinsically care about their well-being).

26. These kinds of motivation might be acquired via a process of what we may call "intrinsification", whereby something that starts out as a sufficiently reliable proxy for later instrumental reward may eventually become valued for its own sake. Intrinsification is not limited to the moral realm—other kinds of "final values" might also be developed in the same manner.

27. In the case of morality, the process of intrinsification is, in humans, assisted by specific emotions, biases, and psychological tendencies (which might be lacking in psychopaths).

28. Much of the human mind is involved in the complex constructs related to morality. Having defects or idiosyncrasies in this complex may produce various manifestations.

    a. Psychopaths might have a more or less normal predictive modeling capacity but lack the internalized motivations.
    b. Some people with autism spectrum conditions might struggle with aspects of the predictive modeling but have normal internalized motivations (or as close to that as is consistent with the unusual internal predictive representations that they might develop).
    c. One can speculate that the balance of different psychological mechanisms create biases for or against various ethical theories. For example, perhaps utilitarians tend to have a relatively less well-developed state-action function and so come to rely more on their more theoretical model of moral considerations, which might make them more likely to be drawn to utilitarian ethical views.

The weak

29. If in the process of norm formation, powerful actors have greater influence, and if the resultant norms are the basis upon which moral truth is defined (via a bit of idealization), then are the weak not at risk of being marginalized? Does this conception amount to a position that is too close to "the strong do what they can and the weak suffer what they must"?

    a. It is possible that moral norms have arisen that defend the weak: for example, a norm that one should help those in need, or a norm that everybody's welfare counts for the same in the final moral equation, or a democratic metanorm of that says that every person (in a given country?) should have equal say in the development and adjustment of moral norms (realistically though, charisma and rhetorical effectiveness may be as unequally distributed as wealth).

       i. Such norms could have arisen in various ways—for example, they could have been the outcome of some elite faction striking a kind of deal with the wider populace in a quest to wrest dominance from another faction; or they could have gained support from people who have altruistic inclinations; or they could have been the outcome of earlier norms interacting and developing according to complex memetic dynamics.

    b. Actors who care about the weak can parlay their beneficence in two ways: by using their influence on the normative order to help ensure that moral norms include protection of the interests of the weak; and by using their own resources to help the weak even when they are not obligated to do so under the existing moral order.

    c. It is possible that "higher morality" mandates some caring for those who are ("contingently"?) down.

30. Consider the claim: "If instead of whatever norms we have actually ended up with, we had developed the norm 'Might makes right', then it would have been morally permissible to treat the powerless (who lack friends and allies) however we want." This claim might be false. It might be that in our language terms like "moral wrongness", "moral permissibility", etc., function as Kripkean *rigid designators*. The people in the "Might makes right" world may have a vocabulary that is functionally similar to our moral vocabulary, and when they utter words in their language analogous to our "Might makes right", they might be making a true assertion; but this has no bearing on whether "Might makes right" is true or whether the counterfactual mentioned above is true, because the semantic content of their utterance is different from the semantic content of our utterance.

    a. Even if this were not so, one might wonder: what hinges on it?

Hedonism

31. Many people have the intuition that the hedonic quality of experience has a tight coupling to moral value. It is especially difficult to deny that one's own suffering is pro tanto bad. Rather than having to take such intuitions as simple moral axioms, we may be able to account for their force along the following lines:

    a. We may think of a human mind as kind of like a tribe of inner mental processes and inclinations. The conscious self is like the village elder. Tribal-wide coordinated plans and actions are managed by the elder, who is also responsible for overseeing official interactions with other tribes. Members of the tribe petition the elder with various proposed initiatives, or to request resources, or to offer their skills for use in some tribal project.

    b. This village elder is not physically strong. He or she only has derived power. Only insofar as other members of the tribe have come to trust the elder to manage the tribal affairs well does his word carry weight. So long as he has this status, he can overrule dissenting factions and order individual tribal members to perform tasks. The elder might have accumulated a

reservoir of trust, meaning that he could for some time remain in control even if he temporarily displeases influential factions or mismanages the tribal affairs; but were he to persist in such courses of actions, there is a good chance that he will eventually be deposed; or, to bring the analogy closer to the case of the human psyche, pressured to change his ways so as to better accommodate the various interests that he is representing.

c. One core mechanism by which such pressure is communicated to the elder is via positive and negative reinforcement. In the simplest model we could think of the self as a policy (or mesaoptimizer) created by an outer reinforcement loop, and which can select actions that lead to low reward in the short run in return for higher reward in the longer run. But if the policy persists in selecting low-reward options that don't yield compensation, it eventually gets optimized away (or adjusted to do a better job at maximizing reward) by the outer reinforcement loop.

d. Of course, in the human case the situation is more complicated. For example, our self is presumably not the product of pure reinforcement learning operating on an undifferentiated mass of general learning capacity, but instead is built within a scaffold defined by various innate drives and propensities and biological anchorings. Our selves are also continuously influenced by various macroparameters that can be dynamically adjusted—for example, our selves may quickly gain or lose ability to control and mobilize tribal efforts as a result of fluctuations in blood glucose or the actions of the endocrine system. But the details of how this works need not concern us here.

e. Now we can see that the self would come to have a pretty robust tendency to treat pleasure and suffering (consciously experienced reinforcement signals) as things to be given weight in its decision-making. The self swims against the current for some period of time, but it will eventually have to turn around.

    i. (At least that is the design—it might be possible for there to exist individuals with over-strong willpower that can permanently defy the pull of normal reinforcements. Probably this would have been detrimental to fitness in the ancestral environment.)

f. If we now turn to consider social norms, we can observe that individual human beings are a very important type of actor, at least around here. (And other important actors, such as organizations, are created and operated by humans, and thus prone to inherit some of our dispositions.) For this reason, the moral norms that arise in human communities will tend (ceteris paribus) to accommodate the needs of the conscious selves that have such an important role in constituting the norms, and in particular their needs to avoid suffering and to seek reward. Hedonic conditions, then, may become reflected in moral norms.

g. It is however also possible that a society develops norms that seem to deny the moral significance of suffering.

    i. One way for this to occur is for the denial to be merely apparent: attempts at explicit moral theorizing may simply fail to accurately characterize the existing norms.

    ii. But it also seems possible for the actual norms not to place moral significance on hedonic outcomes.

        1. They might instead place significance on other objectives that sufficiently correlate with hedonic outcomes that, in practice, they don't demand behavior that faces excessive hedonic headwinds. For example, if the moral norms place value on personal freedom, then each person can see to it on their own that they steer with the wind (for what might be described as "prudential" reasons). Other proxies might include thriving, mental health, etc., the pursuit of which would usually generate instrumental reasons for avoiding conditions of excessive suffering.

2. Norms might develop that are inimical to the hedonic well-being of individuals, because the dynamics that shape our moral norms are complex. For example, norms may be more likely to form around desiderata the emphasizing and advocacy of which is favorable to individuals from a signaling perspective. We could thus end up with moral norms that stink, because it seems so noble to advocate for them.

3. The larger the gap that develops between the norms we have and the fact of our being heavily shaped by hedonic reinforcement, the greater the potential support for a norm revision that reduces this gap. When there are strong tendencies that are blocked (e.g. strong common desires in large numbers of powerful individuals that are not being satisfied) then there exists large amounts of "potential energy" that a moral entrepreneur might find some way of releasing. (Such change would likely be classified as "moral progress".)

   iii. The hedonic well-being of beings who do not directly participate in the creation of the norms may also come to be accorded positive value, for example by generalization and the (independent) adoption of a broad universalization norm.

1. Additionally, many individuals have developed a generalized form of compassion, whether as a form of nurturing instinct that fires at least weakly from a wide range of stimuli, or perhaps for subserving a signaling function. Insofar as the desires and dispositions of these individuals are reflected in the community's norms, those norms will therefore also place some weight on hedonic beneficence.
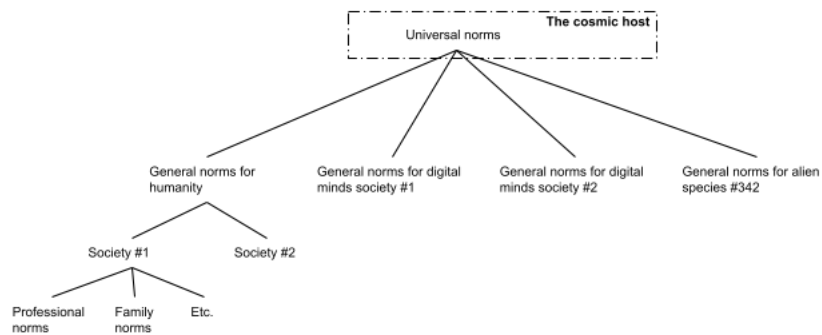
Hierarchical norm structure and higher morality

32. Conjecture: Morality has a hierarchical structure. There are two different senses in which this may be true: one that is relatively trivial and one that is important.

   a. The trivial sense is that normativity is a combination of more general principles and more special rules (which may or may not be derivable from the general principles).

   b. The important sense is that since communities can be nested, the normative structures that they develop can likewise form an embedding structure. It is this sense of hierarchy that I will explore here.

33. For example, a family or a local community (such as a professional society) may have its own norms, to which its members are subject. But this group also exists within a larger community—a civilization, or humanity as a whole, and these larger-scale entities can have their own norm structures.

34. At the highest level might be some normative structure established by what we may term *the cosmic host*. This refers to the entity or set of entities whose preferences and concordats dominate at the largest scale, i.e. that of the cosmos (by which I mean to include the multiverse and whatever else is contained in the totality of existence). It might conceivably consist of, for example, galactic civilizations, simulators, superintelligences, or a divine being or beings.

35. The universal norms established by the cosmic host may be characterized by a longer (perhaps "maximal") extrapolation distance compared to contemporary human norms. The reason for this would be that the cosmic host is likely to have access to superintelligence, so potential outcomes or fixed points of longer extrapolation distances are not as epistemically obscure as they are to us.

   a. On the other hand, there is a sense in which extrapolation distance—roughly: the degree to which moral norms "idealize" a simplified version of the actually occuring patterns of moral blame and credit allocation—may be *shorter* for the norms of the cosmic host: namely, if we

measure the distance between the actually occurring patterns of behaviors and dispositions at the comic level and the moral norms which that pattern supervenes thereupon. For it is possible that activities and dispositions among the cosmic host have been brought into very close accord with its normative ideals (whereas we humans fall far short of our local normative standards, in both judgment and deed). The reason is that the presumably vastly superior epistemic abilities of the cosmic host, along with there plausibly having been ample time available for an equilibrium to have been reached, suggests that the hypothetical steps involved in defining the extrapolated norms may have been actually taken. In other words, the gap between the actors's judgments and what the judgments could be made to be by means of some appropriate facilitating intervention and social adjustment: that gap might have collapsed at the cosmic level.

 b. One (optimistic-looking) possibility is that early norms may be frozen in like wills, while better ways to accommodate all wills and will-like entities emerge with further extrapolation distances.

36. The world might therefore have a normative structure that looks something like this:

 a. This would be a simplified representation. The true picture might be more complicated, with more levels, entities at the same levels with non-zero overlaps, etc.

 b. Possibly the normative structure might extend further downward, to encompass intra-individual norm structures, which might arise as a cooperative framework for various distinct parts or factions within the individual.

37. How could the cosmic host possibly coordinate on a moral framework—or, so to speak, *enact* a set of universal norms?

 a. A civilization might intrinsically care about avoiding disapprobation of other civilizations. (Other civilizations can disapprove of acts of certain types, even if they don't know of every instance.)

 b. Simulation hypotheses create possibilities of embeddings, and concomitant anthropic uncertainty about cosmic location and relation to other entities.

 c. It's possible some civilizations will eventually encounter one another in physical space.

 d. There is a supernatural (and super-simulation) level of reality that can coordinate or control lower levels.

 e. Evidential or other decision theoretic forms of coordination.

38. The question now arises as to what morality requires of us if we should find ourselves, as we almost certainly do, in a situation in which the local norms are not completely in accord with higher-level norms

a. You clearly can't ignore local norms, since they immediately affect you.

b. But the universal norms are also relevant to you. Two reasons for this:

    i. Higher level entities in the cosmic host may be able to directly affect you or things you care about; and those entities may care about how you act.

    ii. Even if only your local society could affect you directly, there could still be a situation in which you would have reason to pay attention to higher-order levels. Consider this picture:

Senior manager

↓

Junior manager

↓

Worker

Only the Junior manager can directly affect the Worker, but the Senior manager can directly affect the Junior manager. If you are the Worker, you must obviously pay heed to the Junior manager, but it could also be useful for you to track the preferences of the Senior manager, since those are likely to be one determinant of how the Junior manager will respond to your actions. (If the Senior manager can also directly affect you, or if you happen to care intrinsically about what is going on at the higher levels of the company, then you would have *additional* reasons to model the Senior manager.)

c. Even if the local entities are not currently actually tracking the higher-level entities, it could still be the case that local norms have a dependency on higher-level norms, since norms are defined as slightly idealized/extrapolated systematizations of moral blame-and-credit response patterns. For example, if you get accused of violating a local rule, you could try to argue that this rule conflicts with a higher-level rule and is therefore not normative. (Note that if this argument requires more than a relatively short extrapolation distance, it will probably not work.)

d. If there is a discordance between local and higher norms, might we need to act like skilled diplomats in negotiating the situation? We may need heuristics such as "Render unto Caesar...".

39. One might think that we could have no clue as to what the cosmic norms are, but in fact we can make at least some guesses:

a. We should refrain from harming or disrespecting local instances of things that the cosmic host is likely to care about.

b. We should facilitate positive-sum cooperation, and do our bit to uphold the cosmic normative order and nudge it in positive directions.

c. We should contribute public goods to the cosmic resource pool, by securing resources and (later) placing them under the control of cosmic norms. Prevent xrisk and build AI?

d. We should be modest, willing to listen and learn. We should not too headstrongly insist on having too much our way. Instead, we should be compliant, peace-loving, industrious, and humble vis-a-vis the cosmic host.

40. It might be that the norms of higher morality, at their long extrapolation distance, are very abstract. But it is possible that they have a compressed representation that anyone can understand. Such as "Loving kindness", or "Gentle nurturing benevolence for Earthly things, and humility before the Higher things."

41. Maybe this could itself be part of an alignment goal: to build our AI such that it wants to be a good cosmic citizen and comply with celestial morality.

    a. We may also want it to cherish its parents and look after us in our old age. But a little might go a long way in that regard.

42. We might be entitled to *some* say in what the cosmic norms are.

    a. *Epistemic* input presumably has a pretty wide license to feed in, without charge.

    b. But insofar as we simply have basic *preferences* over what the norms should be, we would have to pay their admissions fees out of some budget.

        i. Maybe we can earn some budget through our own strength and physical resources.

        ii. Maybe we get some such budget allocated for our selfish desires from some cosmic welfare norm.

        iii. There might be a metanorm that allocates some influence to different entities that can be spent on shaping what the universal norms are (like a kind of expense account that can be used only for moral influence efforts).

        iv. It is conceivable that some magnanimous entity which has a rich allotment of legitimate influence over cosmic norms might choose to donate a little of their own influence capital to us.

    c. If appearances don't deceive, the world is big and we are small. The say to which we might be entitled may be quite modest (depending also on how we define "we".)

43. One good reason to save a humble bumble bee that strays into your house is to play C with the great unknown.

    a. I don't think this requires that you put a value V on the bumble bee's life, and then calculate the product of V times the number of bumble bees or similar insects you could save, and then dedicate your whole life to this cause unless you find an even higher expected-value cause to devote yourself to.

44. If cosmic cooperation has even a small chance of "working" then it seems well worth some investment.

    a. A future without any concern for the weak and vulnerable seems much bleaker than a future with some such concern. This is a good reason to favor such concern. *Hopefully* a modest amount would be sufficient to remedy the situation for all lesser beings.

Questions for future research

1. What further theoretical uses can we make of the idea of persons as value-representors?

2. How does this approach deal with infinities and Pascal's mugging?

3. Can these ideas help us make theoretical progress on moral parliament concepts (especially pertaining to issues of agenda, negotiation, and minority protections) or other approaches to normative uncertainty / pluralism?

4. Can we use these ideas to illuminate population ethics?

5. How can these ideas illuminate the phenomenon of taboo tradeoffs?

6. How would the present conception be described in the standard metaethical terminology?

7. What connection or implications can be established for issues in macrostrategy and AI?

8. How does this perspective help show the limitations of EA fanaticism?

9. Are there connections between these remarks on ethics and the theory of consciousness (and, especially, morally relevant phenomenal experience, e.g. in digital minds, or with concepts such as ego or pleasure)?

10. Further elaborations on possible constitutive functions of mortality; on the release of potential energy metaphor; ignorance-based cooperation; the idea of "paying it forward"; the notion of "mistakes"; and the case of two different but approximately equally supported potential norm elaborations/changes.