A formalization of indirect normativity

Paul Christiano Follow
Apr 20, 2012 · 31 min read

This post outlines a formalization of what Nick Bostrom calls "indirect normativity." I don't think it's an adequate solution to the AI control problem; but to my knowledge it was the first precise specification of a goal that meets the "not terrible" bar, i.e. which does not lead to terrible consequences if pursued without any caveats or restrictions. The proposal outlined here was sketched in early 2012 while I was visiting FHI, and was my first serious foray into AI control.

Introduction

When faced with the challenge of writing down precise moral principles, adhering to the standards demanded in mathematics, moral philosophers encounter two serious difficulties:

- Basic notions, like "agent," "act," and "motive" are themselves almost inherently imprecise. More complex concepts like "well-being," "possible world," and "honesty" are even less approachable from a formal perspective.
- Even expressed in terms of these notions, satisfactory moral principles have proven incredibly elusive. Committing ourselves to any simple, exceptionless theory has proved incompatible with basic ethical intuitions.

In light of these difficulties, a moral philosopher might simply declare: "It is not my place to aspire to mathematical standards of precision. Ethics as a project inherently requires shared language, understanding, and experience; it becomes impossible or meaningless without them."

This may be a defensible philosophical position, but unfortunately the issue is not entirely philosophical. In the interest of building institutions or machines which reliably pursue what we value, we may one day be forced to describe precisely "what we value" in a way that does not depend on charitable or "common sense" interpretation

(in the same way that we today must describe "what we want done" precisely to computers, often with considerable effort). If some aspects of our values cannot be described formally, then it may be more difficult to use institutions or machines to reliably satisfy them. This is not to say that describing our values formally is necessary to satisfying them, merely that it might make it easier.

Since we are focusing on finding any precise and satisfactory moral theory, rather than resolving disputes in moral philosophy, we will adopt a consequentialist approach without justification and focus on axiology. Moreover, we will begin from the standpoint of expected utility maximization, and leave aside questions about how or over what space the maximization is performed.

We aim to mathematically define a utility function U such that we would be willing to build a hypothetical machine which exceptionlessly maximized U, possibly at the catastrophic expense of any other values. We will assume that the machine has an ability to reason which at least rivals that of humans, and is willing to tolerate arbitrarily complex definitions of U (within its ability to reason about them).

We adopt an indirect approach. Rather than specifying what exactly we want, we specify a process for determining what we want. This process is extremely complex, so that any computationally limited agent will always be uncertain about the process' output. However, by reasoning about the process it is possible to make judgments about which action has the highest expected utility in light of this uncertainty.

For example, I might adopt the principle: "a state of affairs is valuable to the extent that I would judge it valuable after a century of reflection." In general I will be uncertain about what I would say after a century, but I can act on the basis of my best guesses: after a century I will probably prefer worlds with more happiness, and so today I should prefer worlds with more happiness. After a century I have only a small probability of valuing trees' feelings, and so today I should go out of my way to avoid hurting them if it is either instrumentally useful or extremely easy. As I spend more time thinking, my beliefs about what I would say after a century may change, and I will start to pursue different states of affairs even though the formal definition of my values is static. Similarly, I might desire to think about the value of trees' feelings, if I expect that my opinions are unstable: if I spend a month thinking about trees, my

current views will then be a much better predictor of my views after a hundred years, and if I know better whether or not trees' feelings are valuable, I can make better decisions.

This example is quite informal, but it communicates the main idea of the approach. We stress that the value of our contribution, if any, is in the possibility of a precise formulation. (Our proposal itself will be relatively informal; instead it is a description of how you would arrive at a precise formulation.) The use of indirection seems to be necessary to achieve the desired level of precision.

I. The Proposal

Our proposal contains only two explicit steps:

1. Obtain a precise mathematical characterization of a particular human's brain.
2. Precisely define a completely abstract environment containing an idealized and unbounded computer. Adopt whatever utility function the human would decide on, if allowed free reign in this abstract environment. (With the technical modification that the utility is restricted to take on values between 0 and 1.)

Each of these steps requires substantial elaboration, but we must also specify what we expect the human to do with these tools.

1. Build a community of copies of herself and an idealized environment for interaction. Deliberate on critical scientific and philosophical questions, and decide how to proceed (possibly in the spirit of what follows).
2. Search through the space of all possible worlds to find a simulation of our own. By interacting with this simulation, incorporate simulations of many living humans into her environment.
3. Continue to make scientific and philosophical progress, as safely as possible. Possibly engage in principled self-modification or understand and build machine intelligences which reflect relevant ethical principles.
4. Output a utility function. This function may be a compact set of ethical principles, or it may require many simulated humans or machine intelligences to apply their own (radically modified)

common sense to evaluate states of affairs on a case-by-case basis.

This proposal is best understood in the context of other fantastic-seeming proposals, such as "my utility is whatever I would write down if I reflected for a thousand years without interruption or biological decay." The counterfactual events which take place within the definition are far beyond the realm our intuition recognizes as "realistic," and have no place except in thought experiments. But to the extent that we can reason about these counterfactuals and change our behavior on the basis of that reasoning (if so motivated), we can already see how such fantastic situations could affect our more prosaic reality.

The remainder of this document consists of brief elaboration of some of these steps, and a few arguments about why this is a desirable process.

Brain Emulation

The first step of our proposal is a high-fidelity mathematical model of human cognition. We will set aside philosophical troubles, and assume that the human brain is a purely physical system which may be characterized mathematically. Even granting this, it is not clear how we can realistically obtain such a characterization.

The most obvious approach to characterizing a brain is to combine measurements of its behavior or architecture with an understanding of biology, chemistry, and physics. This project represents a massive engineering effort which is currently just beginning. Most pessimistically, our proposal could be postponed until this project's completion. This could still be long before the mathematical characterization of the brain becomes useful for running experiments or automating human activities: because we are interested only in a definition, we do not care about having the computational resources necessary to simulate the brain.

An impractical mathematical definition, however, may be much easier to obtain. We can define a model of a brain in terms of exhaustive searches which could never be practically carried out. For example, given some observations of a neuron, we can formally define a brute force search for a model of that neuron. Similarly, given models of individual neurons we may be able to specify a brute force search over all ways of connecting those neurons which account

for our observations of the brain (say, some data acquired through functional neuroimaging).

It may be possible to carry out this definition without exploiting any structural knowledge about the brain, beyond what is necessary to measure it effectively. By collecting imaging data for a human exposed to a wide variety of stimuli, we can recover a large corpus of data which must be explained by any model of a human brain. Moreover, by using our explicit knowledge of human cognition we can algorithmically generate an extensive range of tests which identify a successful simulation, by probing responses to questions or performance on games or puzzles.

In fact, this project may be possible using existing resources. The complexity of the human brain is not as unapproachable as it may at first appear: though it may contain 1014synapses, each described by many parameters, it can be specified much more compactly. A newborn's brain can be specified by about 109bits of genetic information, together with a recipe for a physical simulation of development. The human brain appears to form new long-term memories at a rate of 1–2 bits per second, suggesting that it may be possible to specify an adult brain using 109additional bits of experiential information. This suggests that it may require only about 1010bits of information to specify a human brain, which is at the limits of what can be reasonably collected by existing technology for functional neuroimaging.

This discussion has glossed over at least one question: what do we mean by 'brain emulation'? Human cognition does not reside in a physical system with sharp boundaries, and it is not clear how you would define or use a simulation of the "input-output" behavior of such an object.

We will focus on some system which does have precisely defined input-output behavior, and which captures the important aspects of human cognition. Consider a system containing a human, a keyboard, a monitor, and some auxiliary instruments, well-insulated from the environment except for some wires carrying inputs to the monitor and outputs from the keyboard and auxiliary instruments (and wires carrying power). The inputs to this system are simply screens to be displayed on the monitor (say delivered as a sequence to be displayed one after another at 30 frames per second), while the outputs are the information conveyed from the keyboard and the

other measuring apparatuses (also delivered as a sequence of data dumps, each recording activity from the last 30th of a second).

This "human in a box" system can be easily formally defined if a precise description of a human brain and coarse descriptions of the human body and the environment are available. Alternatively, the input-output behavior of the human in a box can be directly observed, and a computational model constructed for the entire system. Let H be a mathematical definition of the resulting (randomized) function from input sequences (In(1), In(2), ..., In(K)) to the next output Out(K). H is, by design, a good approximation to what the human "would output" if presented with any particular input sequence.

Using H, we can mathematically define what "would happen" if the human interacted with a wide variety of systems. For example, if we deliver Out(K) as the input to an abstract computer running some arbitrary software, and then define In(K+1) as what the screen would next display, we can mathematically define the distribution over transcripts which would have arisen if the human had interacted with the abstract computer. This computer could be running an interactive shell, a video game, or a messaging client.

Note that H reflects the behavior of a particular human, in a particular mental state. This state is determined by the process used to design H, or the data used to learn it. In general, we can control H by choosing an appropriate human and providing appropriate instructions / training. More emulations could be produced by similar measures if necessary. Using only a single human may seem problematic, but we will not rely on this lone individual to make all relevant ethical judgments. Instead, we will try to select a human with the motivational stability to carry out the subsequent steps faithfully, which will define U using the judgment of a community consisting of many humans.

This discussion has been brief and has necessarily glossed over several important difficulties. One difficulty is the danger of using computationally unbounded brute force search, given the possibility of short programs which exhibit goal-oriented behavior. Another difficulty is that, unless the emulation project is extremely conservative, the models it produces are not likely to be fully-functional humans. Their thoughts may be blurred in various ways, they may be missing many memories or skills, and they may lack important functionalities such as long-term memory formation or

emotional expression. The scope of these issues depends on the availability of data from which to learn the relevant aspects of human cognition. Realistic proposals along these lines will need to accommodate these shortcomings, relying on distorted emulations as a tool to construct increasingly accurate models.

The Virtual Environment

For any idealized "software", with a distinguished instruction return, we can use H to mathematically define the distribution over return values which would result, if the human were to interact with that software. We will informally define a particular program T which provides a rich environment, in which the remainder of our proposal can be implemented. From a technical perspective this will be the last step of our proposal. The remaining steps will be reflected only in the intentions and behavior of the human being simulated in H.

Fix a convenient and adequately expressive language (say a dialect of Python designed to run on an abstract machine). T implements a standard interface for an interactive shell in this language: the user can look through all of the past instructions that have been executed and their return values (rendered as strings) or execute a new instruction. We also provide symbols representing H and T themselves (as functions from sequences of K inputs to a value for the Kth output). We also provide some useful information (such as a snapshot of the Internet, and some information about the process used to create H and T), which we encode as a bit string and store in a single environment variable data. We assume that our language of choice has a return instruction, and we have T return whenever the user executes this instruction. Some care needs to be taken to define the behavior if T enters an infinite loop—we want to minimize the probability that the human accidentally hangs the terminal, with catastrophic consequences, but we cannot provide a complete safety-net without running into unresolvable issues with self-reference.

We define U to be the value returned by H interacting with T. If H represented an unfortunate mental state, then this interaction could be short and unproductive: the simulated human could just decide to type 'return 0' and be done with it. However, by choosing an appropriate human to simulate and inculcating an appropriate mental state, we can direct the process further.

We intend for H to use the resources in T to initiate a larger deliberative process. For example, the first step of this process may be to instantiate many copies of H, interacting with variants of

messaging clients which are in contact with each other. The return value from the original process could then be defined as the value returned by a designated 'leader' from this community, or as a majority vote amongst the copies of H, or so on. Another step might be to create appropriate realistic virtual environments for simulated brains, rather than confining them to boxes. For motivational stability, it may be helpful to design various coordination mechanisms, involving frameworks for interaction, "cached" mental states which are frequently re-instantiated, or sanity checks whereby one copy of H monitors the behavior of another.

The resulting communities of simulated brains then engage in a protracted planning process, ensuring that subsequent steps can be carried out safely or developing alternative approaches. The main priority of this community is to reduce the probability of errors as far as possible (exactly what constitutes an 'error' will be discussed at more length later). At the end of this process, we obtain a formal definition of a new protocol H+, which submits its inputs for consideration to a large community and then produces its outputs using some deliberation mechanism (democratic vote, one leader using the rest of the community as advisors, etc.)

Finding Earth

The next step requires our community of simulated brains to construct a detailed simulation of Earth which they can observe and manipulate. Once they have such a simulation, they have access to all of the data which would have been available on Earth. In particular, they can now explore many possible futures and construct simulations for each living human.

In order to locate Earth, we will again leverage an exhaustive search. First, H+ decides on informal desiderata for an "Earth simulation." These are likely to be as follows:

- Comprehensible: the code for the simulation should relate to the best understanding of physics which H+ can develop. This may require operating at a coarser level than the "real" physics underlying our world, and it probably requires structuring the code for the simulation in a way that is not maximally efficient. This requirement is designed to ensure that the 'views' and 'interventions' produced below are acting the expected way when they are being tested for accuracy, so that generalization to new situations will work as expected.

- Supports viewing: H+ should decide on an event to use as a landmark (such as a particular data entry process involved in the definition of H), and search for simulations which support viewing scenes defined relative to that landmark. For example, they might require the simulation to provide a function View(x, y, z, t), which provides a "human's view" from a point specified by the coordinates (x, y, z, t), defined relative to the landmark.

- Accurate: the simulation should consistently conform to expectations about Earth. The history of the individual simulated in H should conform to H's simulated memories; the simulation should follow the same broad historical course as described in data and remembered by H; all of the other data (for example, the complete snapshot of the Internet) should also be in agreement. Perhaps most importantly, the details of whatever system's behavior depends on U should be recorded as part of the data.

- Supports intervention: H+ should decide on a particular class of interventions (such as non-physically setting the output of a particular invocation of one program to be equal to S), and search for simulations which support intervening in that way. For example, they might require the simulation to provide a function Modify(S) which causes all future View() instructions to show what would have happened if the output of the intervention corresponding to S had been performed.

Once H+ has decided on the desiderata, it uses a brute force search to find a simulation satisfying them: for each possible program it instantiates a new copy of H+ tasked with evaluating whether that program is an acceptable simulation. We then define E to be a uniform distribution over programs which pass this evaluation.

We might have doubts about whether this process produces the "real" Earth–perhaps even once we have verified that it is identical according to a laundry list of measures, it may still be different in other important ways. There are two reasons why we might care about such differences. First, if the simulated Earth has a substantially different set of people than the real Earth, then a different set of people will be involved in the subsequent decision making. If we care particularly about the opinions of the people who actually exist (which the reader might well, being amongst such people!) then this may be unsatisfactory. Second, if events transpire significantly differently on the simulated Earth than the real Earth, value judgments designed to guide behavior appropriately in the

simulated Earth may lead to less appropriate behaviors in the real Earth. (This will not be a problem if our ultimate definition of U consists of universalizable ethical principles, but we will see that U might take other forms.)

These concerns are addressed by a few broad arguments. First, checking a detailed but arbitrary 'laundry list' actually provides a very strong guarantee. For example, if this laundry list includes verifying a snapshot of the Internet, then every event or person documented on the Internet must exist unchanged, and every keystroke of every person composing a document on the Internet must not be disturbed. If the world is well interconnected, then it may be very difficult to modify parts of the world without having substantial effects elsewhere, and so if a long enough arbitrary list of properties is fixed, we expect nearly all of the world to be the same as well. Second, if the essential character of the world is fixed but detailed are varied, we should expect the sort of moral judgments reached by consensus to be relatively constant. Finally, if the system whose behavior depends on these moral judgments is identical between the real and simulated worlds, then outputting a U which causes that system to behave a certain way in the simulated world will also cause that system to behave that way in the real world.

Once H+ has defined a simulation of the world which permits inspection and intervention, by careful trial and error H+ can inspect a variety of possible futures. In particular, they can find interventions which cause the simulated human society to conduct a real brain emulation project and produce high-fidelity brain scans for all living humans.

Once these scans have been obtained, H+ can use them to define U as the output of a new community, H++, which draws on the expertise of all living humans operating under ideal conditions. There are two important degrees of flexibility: how to arrange the community for efficient communication and deliberation, and how to delegate the authority to define U. In terms of organization, the distinction between different approaches is probably not very important. For example, it would probably be perfectly satisfactory to start from a community of humans interacting with each other over something like the existing Internet (but on abstract, secure infrastructure). More important are the safety measures which would be in place, and the mechanism for resolving differences of value between different simulated humans.

The basic approach to resolving disputes is to allow each human to independently create a utility function U, each bounded in the interval [0, 1], and then to return their average. This average can either be unweighted, or can be weighted by a measure of each individual's influence in the real world, in accordance with a game-theoretic notion like the Shapley value applied to abstract games or simulations of the original world. More sophisticated mechanisms are also possible, and may be desirable. Of course these questions can and should be addressed in part by H+ during its deliberation in the previous step. After all, H+ has access to an unlimited length of time to deliberate and has infinitely powerful computational aids. The role of our reasoning at this stage is simply to suggest that we can reasonably expect H+ to discover effective solutions.

As when discussing discovering a brain simulation by brute force, we have skipped over some critical issues in this section. In general, brute force searches (particularly over programs which we would like to run) are quite dangerous, because such searches will discover many programs with destructive goal-oriented behaviors. To deal with these issues, in both cases, we must rely on patience and powerful safety measures.

Extrapolation

Once we have a formal description of a community of interacting humans, given as much time as necessary to deliberate and equipped with infinitely powerful computational aids, it becomes increasingly difficult to make coherent predictions about their behavior. Critically, though, we can also become increasingly confident that the outcome of their behavior will reflect their intentions. We sketch some possibilities, to illustrate the degree of flexibility available.

Perhaps the most natural possibility is for this community to solve some outstanding philosophical problems and to produce a utility function which directly captures their preferences. However, even if they quickly discovered a formulation which appeared to be attractive, they would still be wise to spend a great length of time and to leverage some of these other techniques to ensure that their proposed solution was really satisfactory.

Another natural possibility is to eschew a comprehensive theory of ethics, and define value in terms of the community's judgment. We can define a utility function in terms of the hypothetical judgments of astronomical numbers of simulated humans, collaboratively evaluating the goodness of a state of affairs by examining its history

at the atomic level, understanding the relevant higher-order structure, and applying human intuitions.

It seems quite likely that the community will gradually engage in self-modifications, enlarging their cognitive capacity along various dimensions as they come to understand the relevant aspects of cognition and judge such modifications to preserve their essential character. Either independently or as an outgrowth of this process, they may (gradually or abruptly) pass control to machine intelligences which they are suitably confident expresses their values. This process could be used to acquire the power necessary to define a utility function in one of the above frameworks, or understanding value-preserving self-modification or machine intelligence may itself prove an important ingredient in formalizing what it is we value. Any of these operations would be performed only after considerable analysis, when the original simulated humans were extremely confident in the desirability of the results.

Whatever path they take and whatever coordination mechanisms they use, eventually they will output a utility function U'. We then define $U = 0$ if $U' < 0$, $U = 1$ if $U' > 1$, and $U = U'$ otherwise.

II. Desirability

At this point we have offered a proposal for formally defining a function U. We have made some general observations about what this definition entails. But now we may wonder to what extent U reflects our values, or more relevantly, to what extent our values are served by the creation of U-maximizers. Concerns may be divided into a few natural categories:

- Even if the process works as intended, the ultimate intentions of the simulations within the process do not reflect our values, so we should not expect them to output a U which reflects our values.
- The process has some chance of failing to work as intended. If it fails completely, then value will certainly be lost. If it fails with some small probability (over the stochasticity within the process, or over our uncertainty about its behavior) then the resulting utility function may be substantially altered and desirable outcomes may no longer be achieved.
- Any real U-maximizer will be unable to actually carry out the simulation described in the definition of U, and so even if this

process would produce a U reflecting our values, it is not clear how a real U-maximizer will behave.

- Because we have started from the standpoint of bounded expected utility maximization we have ruled out the vast majority of possible value systems. It may be that there is no way to "shorehorn" our values into this framework, so that no matter what bounded U we choose, the resulting agent doesn't satisfy our values very well.
- Are there possible negative consequences to "passing the buck" as in this proposal?
- Could the process itself be morally abhorrent?

We respond to each of these objections in turn.

If it Works as Intended, Will This Process Reflect our Values?

If the process works as intended, we will reach a stage in which a large community of humans reflects on their values, undergoes a process of discovery and potentially self-modification, and then outputs its result. We may be concerned that this dynamic does not adequately capture what we value.

For example, we may believe that some other extrapolation dynamic captures our values, or that it is morally desirable to act on the basis of our current beliefs without further reflection, or that the presence of realistic disruptions, such as the threat of catastrophe, has an important role in shaping our moral deliberation.

The important observation, in the defense of our proposal, is that whatever objections we could think of today, we could think of within the simulation. If, upon reflection, we decide that too much reflection is undesirable, we can simply change our plans appropriately. If we decide that realistic interference is important for moral deliberation, we can construct a simulation in which such interference occurs, or determine our moral principles by observing moral judgments in our own world's possible futures.

There is some chance that this proposal is inadequate for some reason which won't be apparent upon reflection, but then by definition this is a fact which we cannot possibly hope to learn by

deliberating now. It therefore seems quite difficult to maintain objections to the proposal along these lines.

One aspect of the proposal does get "locked in," however, after being considered by only one human rather than by a large civilization: the distribution of authority amongst different humans, and the nature of mechanisms for resolving differing value judgments.

Here we have two possible defenses. One is that the mechanism for resolving such disagreements can be reflected on at length by the individual simulated in H. This individual can spend generations of subjective time, and greatly expand her own cognitive capacities, while attempting to determine the appropriate way to resolve such disagreements. However, this defense is not completely satisfactory: we may be able to rely on this individual to produce a very technically sound and generally efficient proposal, but the proposal itself is quite value laden and relying on one individual to make such a judgment is in some sense begging the question.

A second, more compelling, defense, is that the structure of our world has already provided a mechanism for resolving value disagreements. By assigning decision-making weight in a way that depends on current influence (for example, as determined by the simulated ability of various coalitions to achieve various goals), we can generate a class of proposals which are at a minimum no worse than the status quo. Of course, these considerations will also be shaped by the conditions surrounding the creation or maintenance of systems which will be guided by U–for example, if a nation were to create a U-maximizer, they might first adopt an internal policy for assigning influence on U. By performing this decision making in an idealized environment, we can also reduce the likelihood of destructive conflict and increase the opportunities for mutually beneficial bargaining. We may have moral objections to codifying this sort of "might makes right" policy, favoring a more democratic proposal or something else entirely, but as a matter of empirical fact a more 'cosmopolitan' proposal will be adopted only if it is supported by those with the appropriate forms of influence, a situation which is unchanged by precisely codifying existing power structure.

Finally, the values of the simulations in this process may diverge from the values of the original human models, for one reaosn or another. For example, the simulated humans may predictably disagree with the original models about ethical questions by virtue of (probably) having no physical instantiation. That is, the output of this process is

defined in terms of what a particular human would do, in a situation which that human knows will never come to pass. If I ask "What would I do, if I were to wake up in a featureless room and told that the future of humanity depended on my actions?" the answer might begin with "become distressed that I am clearly inhabiting a hypothetical situation, and adjust my ethical views to take into account the fact that people in hypothetical situations apparently have relevant first-person experience." Setting aside the question of whether such adjustments are justified, they at least raise the possibility that our values may diverge from those of the simulations in this process.

These changes might be minimized, by understanding their nature in advance and treating them on a case-by-case basis (if we can become convinced that our understanding is exhaustive). For example, we could try and use humans who robustly employ updateless decision theories which never undergo such predictable changes, or we could attempt to engineer a situation in which all of the humans being emulated do have physical instantiations, and naive self-interest for those emulations aligns roughly with the desired behavior (for example, by allowing the early emulations to "write themselves into" our world).

Will This Process Work as Intended?

We can imagine many ways in which this process can fail to work as intended–the original brain emulations may accurately model human behavior, the original subject may deviate from the intended plans, or simulated humans can make an error when interacting with their virtual environment which causes the process to get hijacked by some unintended dynamic.

Robustness

We can argue that the proposal is likely to succeed, and can bolster the argument in various ways (by reducing the number of assumptions necessary for succees, building in fault-tolerance, justifying each assumption more rigorously, and so on). However, we are unlikely to eliminate the possibility of error. Therefore we need to argue that if the process fails with some small probability, the resulting values will only be slightly disturbed.

This is the reason for requiring U to lie in the interval [0, 1]–we will see that this restriction bounds the damage which may be done by an unlikely failure.

If the process fails with some small probability ε, then we can represent the resulting utility function as $U = (1-\varepsilon)\, U_1 + \varepsilon\, U_2$, where U_1 is the intended utility function and U_2 is a utility function produced by some arbitrary error process. Now consider two possible states of affairs A and B such that $U_1(A) > U_1(B) + \varepsilon/(1-\varepsilon) \approx U_1(B) + \varepsilon$. Then since $0 \le U_2 \le 1$, we have:

$$U(A) = (1-\varepsilon)\, U_1(A) + \varepsilon\, U_2(A) > (1-\varepsilon)\, U_1(B) + \varepsilon \ge (1-\varepsilon)\, U_1(B) + \varepsilon\, U_2(B) = U(B)$$

Thus if A is substantially better than B according to U_1, then A is better than B according to U. This shows that a small probability of error, whether coming from the stochasticity of our process or an agent's uncertainty about the process' output, has only a small effect on the resulting values.

Moreover, the process contains a humans who have access to a simulation of our world. This implies, in particular, that they have access to a simulation of whatever U-maximizing agents exist in the world, and they have knowledge of those agents' beliefs about U. This allows them to choose U with perfect knowledge of the effects of error in these agents' judgments.

In some cases this will allow them to completely negate the effect of error terms. For example, if the randomness in our process causes a perfectly cooperate community of simulated humans to "control" U with probability ⅔, and causes an arbitrary adversary to control it with probability ⅓, then the simulated humans can spend half of their mass outputting a utility function which exactly counters the effect of the adversary.

In general, the situation is not quite so simple: the fraction of mass controlled by any particular coalition will vary as the system's uncertainty about U varies, and so it will be impossible to counteract the effect of an error term in a way which is time-independent. Instead, we will argue later that an appropriate choice of a bounded and noisy U can be used to achieve a very wide variety of effective behaviors of U-maximizers, overcoming the limitations both of bounded utility maximization and of noisy specification of utility functions.

Other errors

Many possible problems with this scheme were described or implicitly addressed above. But that discussion was not exhaustive,

and there are some classes of errors that fall through the cracks.

One interesting class of failures concerns changes in the values of the hypothetical human H. This human is in a very strange situation, and it seems quite possible that the physical universe we know contains extremely few instances of that situation (especially as the process unfolds and becomes more exotic). So H's first-person experience of this situation may lead to significant changes in H's views.

For example, our intuition that our own universe is valuable seems to be derived substantially from our judgment that our own first-person experiences are valuable. If hypothetically we found ourselves in a very alien universe, it seems quite plausible that we would judge the experiences within that universe to be morally valuable as well (depending perhaps on our initial philosophical inclinations).

Another example concerns our self-interest: much of individual humans' values seem to depend on their own anticipations about what will happen to them, especially when faced with the prospect of very negative outcomes. If hypothetically we woke up in a completely non-physical situation, it is not exactly clear what we would anticipate, and this may distort our behavior. Would we anticipate the planned thought experiment occurring as planned? Would we focus our attention on those locations in the universe where a simulation of the thought experiment might be occurring? This possibility is particularly troubling in light of the incentives our scheme creates—anyone who can manipulate H's behavior can have a significant effect on the future of our world, and so many may be motivated to create simulations of H.

How Will a U-Maximizer Behave In Light of Uncertainty About U?

A realistic U-maximizer will not be able to carry out the process described in the definition of U—in fact, this process probably requires immensely more computing resources than are available in the universe. (It may even involve the reaction of a simulated human to watching a simulation of the universe!) To what extent can we make robust guarantees about the behavior of such an agent?

We have already touched on this difficulty when discussing the maxim "A state of affairs is valuable to the extent I would judge it valuable after a century of reflection." We cannot generally predict our own judgments in a hundred years' time, but we can have well-founded beliefs about those judgments and act on the basis of those

beliefs. We can also have beliefs about the value of further deliberation, and can strike a balance between such deliberation and acting on our current best guess.

A U-maximizer faces a similar set of problems: it cannot understand the exact form of U, but it can still have well-founded beliefs about U, and about what sorts of actions are good according to U. For example, if we suppose that the U-maximizer can carry out any reasoning that we can carry out, then the U-maximizer knows to avoid anything which we suspect would be bad according to U (for example, torturing humans). Even if the U-maximizer cannot carry out this reasoning, as long as it can recognize that humans have powerful predictive models for other humans, it can simply appropriate those models (either by carrying out reasoning inspired by human models, or by simply asking).

Moreover, the community of humans being simulated in our process has access to a simulation of whatever U-maximizer is operating under this uncertainty, and has a detailed understanding of that uncertainty. This allows the community to shape their actions in a way with predictable (to the U-maximizer) consequences.

Can Our Values be Expressed in This Framework?

It is easily conceivable that our values cannot be captured by a bounded utility function. Easiest to imagine is the possibility that some states of the world are much better than others, in a way that requires unbounded utility functions. But it is also conceivable that the framework of utility maximization is fundamentally not an appropriate one for guiding such an agent's action, or that the notion of utility maximization hides subtleties which we do not yet appreciate.

We will argue that it is possible to transform bounded utility maximization into an arbitrary alternative system of decision-making, by designing a utility function which rewards worlds in which the U-maximizer replaced itself with an alternative decision-maker.

It is straightforward to design a utility function which is maximized in worlds where any particular U-maximizer converted itself into a non-U-maximizer–even if no simple characterization can be found for the desired act, we can simply instantiate many communities of humans

to look over a world history and decide whether or not they judge the U-maximizer to have acted appropriately.

The more complicated question is whether a realistic U-maximizer can be made to convert itself into a non-U-maximizer, given that it is logically uncertain about the nature of U. It is at least conceivable that it couldn't: if the desirability of some other behavior is only revealed by philosophical considerations which are too complex to ever be discovered by physically limited agents, then we should not expect any physically limited U-maximizer to respond to those considerations. Of course, in this case we could also not expect normal human deliberation to correctly capture our values. The relevant question is whether a U-maximizer could switch to a different normative framework, if an ordinary investment of effort by human society revealed that a different normative framework was more appropriate.

If a U-maximizer does not spend any time investigating this possibility, than it may not be expected to act on it. But to the extent that we assign a significant probability to the simulated humans deciding that a different normative framework is more appropriate, and to the extent that the U-maximizer is able to either emulate or accept our reasoning, it will also assign a significant probability to this possibility (unless it is able to rule it out by more sophisticated reasoning). If we (and the U-maximizer) expect the simulations to output a U which rewards a switch to a different normative framework, and this possibility is considered seriously, then U-maximization entails exploring this possibility. If these explorations suggest that the simulated humans probably do recommend some particular alternative framework, and will output a U which assigns high value to worlds in which this framework is adopted and low value to worlds in which it isn't, then a U-maximizer will change frameworks.

Such a "change of frameworks" may involve sweeping action in the world. For example, the U-maximizer may have created many other agents which are pursuing activities instrumentally useful to maximizing U. These agents may then need to be destroyed or altered; anticipating this possibility, the U-maximizer is likely to take actions to ensure that its current "best guess" about U does not get locked in.

This argument suggests that a U-maximizer could adopt an arbitrary alternative framework, if it were feasible to conclude that humans

would endorse that framework upon reflection.

Is "Passing the Buck" Problematic?

Our proposal appears to be something of a cop out, in that it declines to directly take a stance on any ethical issues. Indeed, not only do we fail to specify a utility function ourselves, but we expect the simulations to which we have delegated the problem to in turn delegate it at least a few more times. Clearly at some point this process must bottom out with actual value judgments, and we may be concerned that this sort of "passing the buck" is just obscuring deeper problems which will arise when the process does bottom out.

As observed above, whatever such concerns we might have can also be discovered by the simulations we create. If there is some fundamental difficulty which always arises when trying to assign values, then we certainly have not exacerbated this problem by delegation. Nevertheless, there are at least two coherent objections one might raise:

- Even if the simulated humans can uncover any objections we could raise now, this does not guarantee that we can ignore all objections. After all, the objection "Isn't passing the buck problematic?" could be raised at every stage, and always countered by the same response: "If it is problematic, then this will be realized by the people to whom we have passed the buck." If we do not take this objection seriously, then it may be that none of the delegates take it seriously either, and this call and response could be repeated indefinitely.
- This proposal could fail in many (potentially unexpected) ways. If fundamentally resolving ethics requires overcoming some difficulties which we are evading by passing the buck, then we may be adding additional risk without buying much benefit.

Both of these objections can be met with a single response. In the current world, we face a broad range of difficult and often urgent problems. By passing the buck the first time, we delegate resolution of ethical challenges to a civilization which does not have to deal with some of these difficulties–in particular, it faces no urgent existential threats. This allows us to divert as much energy as possible to dealing with practical problems today, while still capturing most of the benefits of nearly arbitrarily extensive ethical deliberation.

Does This Process Have Moral (Dis)Value?

This process is defined in terms of the behavior of unthinkably many hypothetical brain emulations. It is conceivable that the moral status of these emulations may be significant.

We must make a distinction between two possible sources of moral value: it could be the case that a U-maximizer carries out simulations on physical hardware in order to better understand U, and these simulations have moral value, or it could be the case that the hypothetical emulations themselves have moral value.

In the first case, we can remark that the moral value of such simulations is itself incorporated into the definition of U. Therefore a U-maximizer will be sensitive to the possible suffering of simulations it runs while trying to learn about U—as long as it believes that we may might be concerned about the simulations' welfare, upon reflection, it can rely as much as possible on approaches which do not involve running simulations, which deprive simulations of the first-person experience of discomfort, or which estimate outcomes by running simulations in more pleasant circumstances. If the U-maximizer is able to foresee that we will consider certain sacrifices in simulation welfare worthwhile, then it will make those sacrifices. In general, in the same way that we can argue that estimates of U reflect our values over states of affairs, we can argue that estimates of U reflects our values over processes for learning about U.

In the second case, a U-maximizer in our world may have little ability to influence the welfare of hypothetical simulations invoked in the definition of U. However, the possible disvalue of these simulations' experiences are probably seriously diminished.

In general the moral value of such hypothetical simulations' experiences is somewhat dubious. If we simply write down the definition of U, these simulations seem to have no more reality than story-book characters whose activities we describe.

The best arguments for their moral relevance comes from the great causal significance of their decisions: if the actions of a powerful U-maximizer depend on its beliefs about what a particular simulation would do in a particular situation, including for example that simulation's awareness of discomfort or fear, or confusion at the absurdity of the hypothetical situation in which they find themselves, then it may be the case that those emotional responses are granted moral significance. However, although we may define astronomical numbers of hypothetical simulations, the detailed emotional

responses of very view of these simulations will play an important role in the definition of U.

Moreover, for the most part the existences of the hypothetical simulations we define are extremely well-controlled by those simulations themselves, and may be expected to be counted as unusually happy by the lights of the simulations themselves. The early simulations (who have less such control) are created from an individual who has provided consent and is selected to find such situations particularly non-distressing.

Finally, we observe that U can exert control over the experiences of even hypothetical simulations. If the early simulations would experience morally relevant suffering because of their causal significance, but the later simulations they generate robustly disvalue this suffering, the later simulations can simulate each other and ensure that they all take the same actions, eliminating the causal significance of the earlier simulations.

. . .

Originally published at ordinaryideas.wordpress.com on April 21, 2012.