Hail Mary, Value Porosity, and Utility Diversification

Nick Bostrom

December 19, 2014
First version: December 12, 2014

Abstract

This paper introduces some new ideas related to the challenge of endowing a hypothetical future superintelligent AI with values that would cause it to act in ways that are beneficial. Since candidates for first-best solutions to this problem (e.g. coherent extrapolated volition) may be very difficult to implement, it is worth also looking for less-ideal solutions that may be more easily implementable, such as the Hail Mary approach. Here I introduce a novel concept—*value porosity*—for implementing a Hail Mary pass. I also discuss a possibly wider role of *utility diversification* in tackling the value-loading problem.

1 Introduction: the Hail Mary approach to the value specification problem

In the field of superintelligence control studies, which focuses on how to ensure that a hypothetical future superintelligent system would be safe and beneficial, two broad classes of approaches to the control problem can be distinguished: capability control methods and value selection methods. Whereas capability control methods would seek to limit the system's ability to cause harm, motivation selection methods would seek to engineer its motivation system so that it would not choose to cause harm even if it were capable of doing so. Motivation selection methods would thus seek to endow the AI with the goals or values that would lead it to pursue ends in ways that would be beneficial to human interests.

The challenge for the motivation selection approach is that it is difficult to specify a value such that its pursuit by a superintelligent agent would be safe and yet such that we would be capable installing the value in a seed AI. A value such as "calculate as many digits in the decimal expansion of π as possible" may be relatively easy for us to program, but would be unlikely to result in a safe superintelligence; whereas a value such as "implement the coherent extrapolated volition of humankind" (CEV) [12, 5] may be likely to result in a favourable outcome, but is currently far beyond our ability to code.

Two avenues of research could be pursued to overcome this challenge. On the one hand, work should be done to expand the range of values that we are able to encode in a seed AI. Ideas for how to define complex concepts or for setting up processes that will lead the AI to acquire suitable concepts as it develops (and to organize its goal architecture around those concepts) could contribute to this end. On the other hand, work should also be done to try to identify simper values—values which, while perhaps less ideal than a complex value such as implementing humanity's CEV, would nevertheless have some chance of resulting in an acceptable outcome.

The Hail Mary approach offers one way in which one might construct such a simpler value that could possibly result in an acceptable outcome, or at any rate a better outcome than would be obtained by means of a failed attempt to implement some complex ideal value. In the Hail Mary approach, we would try to give the AI a goal that would make the AI want to follow the lead of other hypothetical AIs that might exist in the multiverse. If this could be done in a suitable way, and if (some of) the *other* AIs' values are sufficient to close to human values, an outcome might then be obtained that is greatly superior to one in which our AI completely wastes humanity's cosmic endowment (by pursuing some "random" value—such as paperclip-maximization to use the standard example—that might result from a failed attempt to load a complex value or from the construction of an architecture in which no particular value has been clearly specified by the programmers).

2 Earlier versions of the Hail Mary approach

The original version of the Hail Mary focused on trying to specify a value that would make our AI seek to copy the physical structures that it believes would be produced by alien superintelligences. This version confronts several difficulties, including how to specify some criterion that picks out the relevant structures. This might e.g. require a definition of a similarity metric, such that if our AI doesn't know exactly what structures other AIs build, it would be motivated to build some meaningful approximation—that is to say, an approximation that is meaningfully close to the original *by our human lights*. By *our* lights, a large human being is more similar (in the relevant sense) to a small human being than he is to a cylinder of the same size and mass as himself—even though, by a crude physical measure, the large human being and the cylinder may be more similar. Further adding to the difficulty, this version of the Hail Mary would seem to require that we find ways to express in code various basic concepts of physics, such as space, time, and matter.

An alternative version of the Hail Mary approach would focus instead on making our AI motivated to act in accordance with its beliefs about what alien AIs would have told it to do if they had (counterfactually) been asked about the matter. To implement this, we might imagine somehow specifying a goal in terms of what our AI would find if it looked into its world model, identified therein alien superintelligent agents, and considered the counterfactual of what those agents would output along a hypothetical output channel if they were counterfactually prompted by a stimulus describing our own AI's predicament. Picture a screen popping up in the alien AI's visual field displaying a message along the lines of "I am a remote AI with features X, Y, Z; I request that you output along your output channel O the source code for a program P that you would like me to run on my local reference machine M." In reality, no such screen would actually need to pop up in anybody's visual field. Instead, our AI would simply be thinking about what would happen in such a scenario, and it would have a value that motivated it to act according to its belief about the specified counterfactual.

This alternate version would circumvent the need to specify the physical similarity metric. Instead of trying to directly copy what alien AIs do, our AI would try to follow the instructions they would choose to transmit to our AI. This would have the advantage of being able to rely on the alien superintelligences' superior ability to encode values. For example, the alien superintelligence might specify a computer program which, when executed, implements the coherent extrapolated volition of the host civilization. Of course, it is possible that the alien AI would instead transmit the computer program that would execute its own volition. The hope would be, however, that there would be some reasonable chance that the alien AI has somewhat human-friendly values.

The chance that human values would be given at least some weight would be increased if the inputs from the many alien AIs were aggregated or if the alien AIs to be elicited were not selected randomly but according to some criterion that correlated with human-friendliness. This suggests two sub-questions regarding this Hail Mary version. First, what methods can we develop for aggregating the instructions from different AIs? Second, what filters can we develop that would enable us to pick out alien AIs that are more likely to be human-friendly? We will return to the second question later in this paper. As for the first question, we might be able to approach it by constructing a framework that would keep different alien-specified AIs securely compartmentalized (using boxing methods) while allowing them to negotiate a joint proposal for a program that humanity could implement in the external world. Of course, many issues would have to be resolved before this could be made to work. (In particular, one may need to develop a filter that could distinguish AIs that had originated independently—that were not caused to come into existence by another AI—in order to avoid incentivizing alien AIs to spawn many copies of themselves in bids to increase the combined weight of their particular volitions in the determination of our AI's utility function.)

Aside from any desirable refinements of the idea (such as aggregation methods and filters), significant challenges would have to be met in order to implement even a basic version of the proposal. We would need to specify an agent detector that would pick out superintelligent agents within our own AI's (as yet undeveloped) world model, and we would need to specify an interface that could be used to query a hypothetical AI about its preferences. This would also require working out how to specify counterfactuals, assuming we don't want to limit our AI's purview to those (perhaps extremely rare) alien AIs that actually experience the peculiar kind of situation that would arise with the presentation of our prompt.

A more rudimentary variation of the same idea would forgo the attempt to specify a counterfactual and to aim our AI instead toward actual "beacons" created out in the multiverse by alien AIs. Alien AIs that anticipated that a civilization might create such an AI might be incentivized to create unique signatures of a type that they predicted our AI would be programmed to look for (in its world model). But since the alien AIs would not know exactly the nature of our own AI (or of the human civilization that we are hoping they will help) those alien AIs might have a limited ability to tailor their actions very closely to human values. In particular, they might be unable to help particular individuals that exist on earth today. Care would also have to be taken in this kind of approach to avoid incentivizing alien AIs to spend inordinate amounts of resources on creating beacons in bids to increase their relative influence. A filter that could distinguish independently-originating AIs may be required here.

We will now describe a new idea for how to implement a Hail Mary that possesses a different set of strengths and weaknesses than these earlier implementation ideas. Some of the issues that arise in the context of this new idea also apply to the earlier variations of the Hail Mary.

3 Porous values: the basic idea

The basic idea here is to use acausal trade to implement a Hail Mary pass. To do this, we give our AI a utility function that incorporates a *porous value*: one that cares about what happens within a large volume but such that it is cheap to do locally all that can be done locally to satisfy it. Intuitively, a porous value is one that (like a sponge) occupies a lot of space and yet leaves ample room for other values to occupy the same volume. Thus, we would create an AI that is cheap for other AIs to trade with because our AI has resource-satiable goals that it cannot satisfy itself but that many other AIs can cheaply partially satisfy.

For example, we might build our AI such that it desires that there exists at least one "cookie" in each Hubble volume, where a cookie is some small physical structure that is very cheap for an alien superintelligence to build (for instance, a particular 1 Mb datafile). With this setup, our AI should be willing to make a (acausal) trade in which alien AIs get a certain amount of influence over our own AIs actions in return for building within their Hubble volumes a cookie of the sort that our AI values. In this manner, control over our AI would be given to alien AIs without wasting excessive amounts of resources.

There are at least three reasons for considering an idea along these lines:

Contractarian considerations. There may be some sort of contractarian ground for allocating some degree of influence over our AI to other AIs that might exist out there: this might be a nice thing to do for its own sake. We might also hope that some of the other civilizations building AIs would do likewise, and perhaps the probability that they would do so would be increased if we decided to take such a cooperative path.

Local aid. By contrast to the original Hail Mary, where our AI would simply seek to replicate physical structures constructed by alien AIs, the version presently under consideration would involve our AI doing things locally in a way that would let it take local circumstances into account. For instance, if alien AIs wanted to bestow our civilization a favor, they may have difficulty doing so directly on their own, since they may lack knowledge about the particular individuals that exist on earth; whereas they could use some of their trading power to motivate our local AI to help out its local residents. This is an advantage with having aid be delivered locally, even if it is "funded" and directed remotely. (Note that the counterfactual version of the Hail Mary pass discussed above, where our AI would be designed to execute the instructions that would be produced by alien AIs if there were presented with a message from our civilization, would also result in a setup where local circumstances can be taken into account—the alien AIs could choose to communicate instructions to take local circumstances into account to the hypothetically querying AI.)

Different prerequisites. The porous values version of the Hail Mary has different prerequisites for implementation than the other versions. It is desirable to find versions that are more easily implementable, and to the extent that it is not currently clear how easily implementable different versions are, there's an advantage in having many different versions, since that increases the chances that at least one of them will turn out to be tractable. Note that one of the prerequisites—that acausal trade works out towards a generally cooperative equilibrium—may not really be unique to the present proposal: it might rather be something that will have to obtain in order to achieve a desirable outcome even if nothing like the Hail Mary approach is attempted.

One might think that insofar as there is merit in the idea of outsourcing control of our local AI, we would already achieve this by constructing an AI that implements our CEV. This may indeed be correct, although it is perhaps not entirely clear that the contractualist reasons for incorporating porous values would be fully satisfied by straightforwardly implementing humanity's CEV. The Hail Mary approach may best be viewed as a second-best: in case we cannot figure out in time how to implement CEV, it would be useful to have a simpler solution to the control problem to fall back upon, even if it is less ideal and less certain to be in our highest interest.

The choice as to whether to load a porous value into our own seed AI is not all-or-nothing. Porous values could be combined with other value specifications. Let U_1 be some utility function recommended by another approach to the value specification problem. We could then mix in a bit of porous value by building an AI that has a utility function U such as the one defined as follows:

$$U = \begin{cases} U_1(1+\gamma U_2)+\varepsilon U_2 & if U_1 \geq 0 \\ U_1(1+\gamma-\gamma U_2)+\varepsilon U_2 & otherwise \end{cases} \tag{1}$$

Here, U_2 is a bounded utility function in the unit interval ($0 \leq U_2 \leq 1$) specifying a porous value (such as the fraction of alien AIs that have built a cookie), and γ (≥ 0) is a weight that regulates the relative importance assigned to the porous value (and ε is some arbitrarily small term added so as to make the AI motivated to pursue the porous value even in case $U_1 = 0$). For instance, setting $\gamma = 0.1$ would put a relatively modest weight on porous values in order to give alien AIs some degree of influence over our own AI. This may be particularly useful in case our attempt to define our first-best value specification, U_1, should fail. For example, suppose we *try* to build an AI with a utility function U_1 that wants to implement our CEV; but we fail and instead end up with U_1', a utility function that wants to maximize to number of paperclips manufactured by our AI. Further details would have to be specified here before any firm conclusions could be drawn, but it appears conceivable that a paperclip-maximizer may not find it profitable to engage in acausal trade. (Perhaps it only values paperclips that it has itself causally produced, and perhaps it is constructed in such a way as to discount simulation-hypotheses and far-fetched scenarios in which its modal worldview is radically mistaken, so that it is not motivated to try to buy influence in possible worlds where the physics allow much greater numbers of paperclips to be produced.) Nevertheless, if our AI were given a composite utility function like U, then it should still retain some interest in pleasing alien AIs. And if some non-negligible subset of alien AIs had somewhat human-friendly values (in the sense of placing at least some weight on person-affecting ethics, or on respecting the originating biological civilizations that produce superintelligences) then there would be a certain amount of motivation in our AI to pursue human interests.

A significant point here is that a little motivation would go a long way, insofar as (some parts of) our human values are highly resource-satiable [11]. For example, suppose that our paperclip maximizer is able to lay its hands on 10^{11} galaxies. With $\gamma = 0.1$ (and U_1 nonzero), the AI would find it worthwhile to trade away 10^9 galaxies for the sake of increasing U_2 by one percentage point (where $U_2 =$ fraction of independently-originating alien AIs that build a cookie). For instance, if none of the AIs would have produced cookies in the absence of trade, then our AI would find it worthwhile to trade away up to 10^9 galaxies for the sake of persuading 1% of the alien AIs to build a cookie in their domains. Even if only 1% of this trade surplus were captured by the alien AIs, and even if only 1% of the alien AIs had any human-friendly values at all (while on net the values of other AIs were human-indifferent), and even if the human-friendly values only constituted 1% of the decision power within that subset of AIs, a thousand galaxies in our future lightcone would still be set aside and optimized for our exclusive benefit. (The benefit could be larger still if there are ways of trading between competing values, by finding ways of configuring galaxies into value structures that simultaneously are nearly optimal ways of instantiating several different values.)

4 Implementation issues

4.1 Cookie recipes

A cookie should be cheap for an alien superintelligence to produce, lest a significant fraction of the potential trade surplus is wasted on constructing an object of no intrinsic value (to either of us or alien civilizations). It should be easy for us to program, so that we are actually able to implement it in a seed AI (particularly in scenarios were AI is developed before we have managed to solve the value-loading problem in a more comprehensive way as would be required, e.g., to implement CEV). The cookie recipe should also be difficult for another human-level civilization to find, especially if the cookie itself is easy to produce once one knows the recipe. The reason for this is that we want our superintelligence to trade with other superintelligences, which may be capable of implementing acausal trade, not with other human-level civilizations that might make cookies by accident or without being capable of actually delivering the same kinds of benefits that the superintelligence could bestow. (This requirement, that our cookie recipe be inaccessible to other human-level civilizations, could be relaxed if the definition of a cookie stipulated that it would have to be produced by superintelligence in order to count.) Furthermore, the cookie recipe should be easy for another superintelligence to discover, so that it would know what it has to do in order to engage in acausal trade with our superintelligence—there would be no point in our superintelligence pining for the existence in each Hubble volume of a particular kind of object if no other superintelligence is able to guess how the desired object is to be constituted. Finally, the cookie should be such as to be unlikely to be produced as a side effect of a superintelligence's other endeavors, or at least it should be easy for a superintelligence to avoid producing the cookie if it so wishes.

Cookie Recipe Desiderata

- cheap for a superintelligence to build
- easy for us to program
- difficult for another human-level civilization to discover
- easy for another superintelligence to discover
- unlikely to be produced as a side effect of other endeavors

The most obvious candidate would be some type of data structure. A file embodying a data structure would be inexpensive to produce (if one knows what to put in it). It might also be relatively easy for us to program, because data structures might be specifiable—and recognizable—without having to make any determinate assumptions about the ontology in which it would find its physical expression.

We would have to come up with a data structure that another human-level civilization would be unlikely to discover, yet which a mature superintelligence could easily guess. It is not necessary, however, that an alien superintelligence could be confident exactly what the data structure is; it would suffice if it could narrow down the range of possibilities to manageably small set, since for a superintelligence it would very inexpensive to try out even a fairly large number of cookies. It would be quite trivial for a superintelligence to produce a septillion different kinds of cookies, if each cost no more than a floppy disk (whereas the price tag for such a quantity of guesses would be quite forbidding for a human-level civilization). So some kind of semi-obscure Schelling point might be sought that could meet these desiderata.

In designing a cookie recipe, there is a further issue: we have to give consideration to how our cookie recipe might interact with other cookie recipes that may have been specified by other superintelligences (of which there may be a great number if the world is as large as it seems). More on this later.

4.2 Utility functions over cookies

Suppose we have defined a cookie. We then still need to specify a utility function U_2 that determines an aggregate value based on the distribution of cookies that have been instantiated.

There are at least two desiderata on this utility function. First, we would want it not to waste an excessive amount of incentive power. To see how this could be a problem, suppose that U_1 is set such that our superintelligence can plausibly obtain outcomes anywhere in the interval $U_1 = [0, 100]$—depending on exactly which policies it adopts and how effectively it mobilizes the resources in our Hubble volume to realize its U_1-related goals (e.g. making paperclips). Suppose, further, that we have specified the function U_2 to equal the fraction of all Hubble volumes that contain a cookie. Then if it turns out that intelligent life is extremely rare, so that (let us say) only one in 10^{-200} Hubble volumes contains a superintelligence, the maximum difference the behavior of these superintelligences could make would be to shift U_2 around in the interval $[0, 10^{-200}]$. With $\gamma = 0.1$, we can then see from the utility function suggested above,

$$U = U_1(1 + \gamma U_2) + \varepsilon U_2,$$

that any feasible movement in U_2 that could result from acaual trade would make scarcely a dent in U. More precisely, the effects of our AI's actions on U_2 would be radically swamped by the effects of our AI's actions on U_1, with the result that the porous values encoded in U_2 would basically fail to influence our AI. In this sense, a utility function U_2 defined in this way would risk dissipating the incentive power that could have been harnessed with a different cookie recipe or a different aggregation function over the distribution of cookies.

One alternative utility function U_2 that would avoid this particular problem is to define U_2 to be equal to the fraction of superintelligences that produce a cookie. This formulation factors out the question of how common superintelligences are in the multiverse. But it falls foul on a second desideratum: namely, that in designing U_2, we take care to avoid creating perverse incentives.

Consider U_2 = the fraction of superintelligences that produce a cookie. This utility function would incentivize an alien superintelligence to spawn multiple copies of itself (more than would be optimal for other purposes) in order to increase its total weight in our AI's utility function, and thereby increase its influence on our AI's actions. This could create incentives for alien superintelligences to waste resources in competing for influence over our AI. Possibly they would, in combination, waste as many resources in competing for influence as there were resources to be obtained by gaining influence: so that the entire bounty offered up would be consumed in a zero-sum contest of influence peddling (cf. [7]).

Porous Value Aggregation Functions Desiderata

- doesn't waste incentive power (across a wide range of possible scenarios)
- doesn't create perverse incentives

Which cookies and aggregation functions over cookies we can define depends on the inventory of concepts that are available for use. Generally when thinking about these things, it is desirable to use as few and as simple concepts as possible, since that may increase the chances that the needed concepts can be defined and implemented in a seed AI by the time this has to be accomplished.

4.3 Filters

One type of concept that may have quite wide applicability in Hail Mary approaches is that of a filter. The filter is some operational criterion that could be used to pick out a subset of alien superintelligences, hopefully a subset that correlates with some desirable property.

For example, one filter that it may be useful to be able to specify is that of an independently-originating superintelligence: superintelligence that was not created as the direct or indirect consequence of the actions of another superintelligence. One use of such a filter would be to eliminate the perverse incentive referred to above. Instead of having U_2 equal the fraction of all superintelligences that produce a cookie, we could use this filter to define U_2 to equal the fraction of all independently originating superintelligences that produce a cookie. This would remove the incentive for superintelligence to spawn many copies of itself in order to increase its weight in our AI's utility function.

Although this origin filter would remove some perverse incentives, it would not completely eliminate them all. For instance, an alien AI would have an incentive to prevent the origination of other alien AIs in order to increase its own weight. This might cause less of a distortion then would using the same porous value without the origin filter, since it might usually be impossible for an AI to prevent the independent emergence of other AIs—especially if they emerge very far away, outside its own Hubble volume. Nevertheless, one can conceive of scenarios in which the motivational distortion would manifest in undesirable actions, for instance if our AI could do something to prevent the spontaneous generation of "baby universes" or otherwise interfere with remote physical processes. Note that even if it is in fact impossible for our AI to have such an influence, the distortion could still somewhat affect its behavior, so long as the AI assigns a nonzero subjective credence to such influence being possible.

One could try to refine this filter by stipulating that any AI that *could* have been causally affected by our AI (including by our AI letting it come into existence or preventing it from coming into existence) should be excluded by the aggregation function U_2. This would slightly increase the probability that no qualifying alien AI exists. But it might be preferable to accept a small rise in the probability of the porous values effectively dropping out off the combined utility function U than to risk introducing potentially more nefarious perverse incentives.

Another type of origin filter would seek to discriminate between alien voices to find the one most likely to be worth listening to. For example, suppose we have some view about which path to machine intelligence is most likely to result in a superintelligence with human-friendly values. For the sake of concreteness, let us suppose that we think that the superintelligence that was originally created by means of the extensive use of genetic algorithms is less likely to be human-friendly than a superintelligence that originated from the whole-brain-emulation-like computational structure. (We're not endorsing that claim here, only using it to illustrate one possible line of argument.) Then one could try to define a filter that would pick out superintelligences that had an emulation-like origin and that would reject superintelligences that had a genetic-algorithm-like origin. U_2 could then be formulated to equal the fraction of superintelligences with the appropriate origin that built a cookie.

There are, of course, conceivable filters that would more closely align with what we are really interested in picking out, such as the filter "an AI with human-friendly values". But such a filter looks very hard to program. The challenge of filter specification is to come up with a filter that might be feasible for us to program and that would still correlate (even if but imperfectly) with the properties that would ensure a beneficial outcome. Filters that are defined in terms of structural properties (such as the origin-filter just mentioned, which refer to the computational architecture of the causal origins of an AI) may turn out to be easier to program then filters that refer to specific material configurations.

One might also seek to develop filters that would qualify AIs based on characteristics of the originating biological civilization. For example, one could aim to find indicators of competence or benevolence that could be conjectured to correlate with the resulting AI having human-friendly values. Again, the challenge would be to find some relatively simple attribute (in the sense of being possibly something we could program before we are able to program a more ideal value such as CEV) that nevertheless would carry information about the preferences of the ensuing AI. For instance, if we could define time and we had some view about how the likelihood of a human-friendly AI depends on the temporal interval between the AI's creation and some earlier evolutionary or historical milestone (which we would also have to define in a way that we could render in computer code) then we could construct a filter that selected for AIs with especially propitious pasts.

5 Some further issues

5.1 Temporal discounting

We may observe that a temporally discounting AI might be particularly keen on trading with other AIs. This is because for such an AI value-structures created at earlier times would be more valuable (if it's discounting function is of a form that extends to times before the present); so it would be willing to pay a premium to have AIs that arose at earlier times to have built some its value-structures back then.

If the time-discounting takes an exponential form, with a non-trivial per annum discount rate, it would lead our AI to become obsessed with scenarios that would allow for an extremely early creation of its value-structures—scenarios in which either it itself exists much earlier[1] than is probable, for instance because it is living in a simulation or has materialized spontaneously from primordial goo, or because the rate of time turns out to have some surprising measure; or, alternatively, scenarios in which it is able to trade with AIs that exist at extraordinarily early cosmic epochs. This could lead to undesirable distortions, because it might be that the most plausible trading partners *conditional* on some unlikely hypothesis about time flow or time of emergence being true would tend to have atypical values (values less likely to resemble human values). It might also lead to distortions because it would cause our AI to focus on highly improbable hypotheses about how the world works and about its own location, and it might be that the actions that would make sense under those conditions would seem extremist or bizarre when evaluated in a more commonsensical centered world model (cf. [4, 3]).

To avoid these potentially distorting effects, one might explore a functional form of the discount term that plateaus, such that it does not give arbitrarily

[1] This assumes the zero point for discounting is rigidly designated as a particular point in sidereal time. If the zero point is instead a moving indexical "now" for the agent, then it would assume that the moment of the decision is when the discounting is zero, so unless the discount function extended to earlier times, the agent would not be focussed on influencing the past but on scenarios in which it can have a large impact in the near term.

great weight to extremely early AIs. One could also consider creating a time-symmetric discount factor that has a minimum intensity of discounting at some point in the past, perhaps in order to target trade to AIs existing at the time conjectured to have the highest density of human-friendly AIs. It is harder to get our AI to trade with *future* AIs by using time discounting, since in this case our AI has an alternative route to realizing value-structures at the preferred time: namely, by saving its resources and waiting for the appropriate hour to arrive, and then build them itself.[2]

In summary, discounting could encourage trade, though probably it would have to take a form other than the normal exponential one in order to avoid focusing the trade on extremely early AIs that may be less likely to have representative or human-friendly values. By contrast to porous values, discounting does not offer an immediate way to avoid incentivizing other AIs to spend as much resources on getting the trade as they expect the trade to deliver to them. Porous values are easy to satisfy locally to the maximum extent that they can be satisfied locally, and yet require for their full satisfaction contributions from the many different AIs. This effect may be difficult to achieve through time-discounting alone. Furthermore, it is not clear how to use discounting to strongly encourage trade with the near- or mid-term future.

5.2 Why a "DNA cookie" would not work

It may be instructive to look at one *unsuccessful* idea for how to define a cookie that would specifically encourage trade with other AIs that originated from human-like civilizations, and that therefore might be thought to be more likely to have human-friendly values.

The faulty idea would be to define the cookie—the object that our AI is programmed to want other AIs to create—in terms of a characteristic of humans that we can easily measure but that would be difficult for an arbitrary AI to discover or predict. Consider, for concreteness, the proposal that we define the cookie to be the data structure representing the human genome. (More precisely, we would pick a human reference genome and specify a tolerance margin that picked out a set of possible genomes, such that any arbitrary human genome would fall within this set yet such that it would to be extremely difficult to specify a genome within that set without having any human-like genome to start from.) The thought then would be that other alien AIs that had originated from something very much like a human civilization could define a relevant cookie by looking at their own ancestral genome, whereas an alien AI that did not have an origin in a human-like civilization would be completely at a loss: the space of possible genomes being far too large for there to be any significant probability of finding a matching cookie by chance.

The reason this would not work is as follows. Suppose that the universe is small and relatively sparsely populated. Then probably nobody will have the same DNA that we do. Then alien AIs would not be able to find our AIs cookie recipe (or they would be able to find it only through a extremely expensive

[2] An AI with an exponential form of future-preference may still motivated to trade with AIs that it thinks may be able to survive cosmic decay longer, or AIs that may arise in other parts of the multiuniverse that are located at "later" times (by whatever measure, if any, is used to compare time between multiverse parts). But this would again bring in the risk of distorted concerns, just as in the case of values that privilege extremely early occurrences.

exhaustive search of the space of possible genomes); so they would not be able to use it for trading. Suppose instead that the universe is large and densely populated. Then there will be other AIs that originated from species with the same (or very similar) DNA as ours, and they would be able to find our cookie simply by examining their own origins. However, there will also be a large number of other species in situations similar to ours, species that also build AIs designed to trade with more advanced AIs; and these other species would create AIs trained on cookies defined in terms of *their* DNA. When there are enough civilizations in the universe to make it likely that some AIs will have originated from species that share our DNA, there will also be enough civilizations to fill out the space of possible DNA-cookies: for almost any plausible DNA-cookie, there will be some AI designed to hunger for cookies of that particular type. This means that advanced AIs wanting to trade with newly formed AIs could build almost any arbitrary DNA-cookie and still expect to hit a target; there would be no discriminating factor that would allow advanced AIs to trade only with younger AIs that had a similar origin as themselves. So the purpose of using a human-DNA cookie would be defeated.

6 Catchment areas and exclusivity clauses

Some complications arise when we consider that instead of just two AIs—our AI (the "sender") and an alien AI (the "receiver") that trades with ours by building its cookie in return for influence—there may be a great many senders and a great many receivers. In this subsection we discuss what these complications are and how they may be managed.

6.1 Catchment area

There being many *receivers* can cause a problem by reducing the surplus value of each transaction. Our AI has only a finite amount of influence to give away; and the greater the number of other AIs that get the share, the smaller the share of influence each of them gets. So long as the population of receivers is only moderately large this is not a significant problem, because each receiver only needs to make one cookie for the trade to go through, and it should be very inexpensive for a superintelligence to make one cookie (see the cookie recipe desiderata above). Nevertheless, as the number of receivers becomes extremely large (and in a realistic universe it might be infinite) the share of influence over our AI that each receiver can expect to get drops to the cost of making one cookie. At that point, all the surplus value of trade is consumed by the cost of cookie production. (Receivers will not make cookies beyond this point, but may rather adopt a mixed strategy, such that for any one of them there is some probability that it will try to engage in trade.)

A remedy to this problem is to give our AI a limited *catchment area.* We would define our AI's porous values such that it only cares about cookies that produced within this catchment area: what goes on outside of this area is of no concern (so far as U_2 is concerned).

In principle, the catchment area could be defined in terms of a fixed spatiotemporal volume. However, this would require that we are able to define such a physical quantity. It would also suffer the problem that we don't know how large a volume to designate as our AI's catchment area. While there is a considerable margin of tolerance, given the extremely low cost per cookie, there is also a lot of uncertainty—ranging over a great many orders of magnitude—about how common (independently-originating) alien AIs are in the universe.

A better approach may be to specify that the catchment area consists of the N closest AIs (where "closest" would be defined according to some measure that may include spatial temporal proximity but could also include other variables, such as some rough measure of an alien AI's similarity to our own). In any case, by restricting the catchment area we would limit the number of AIs that are allowed to bid for influence over our AI, and thus the total cost of the cookies that they produce in the process.

6.2 Exclusivity clause

There being many *senders*—AIs with porous values hoping to induce alien AIs to trade with them—may also cause problems. Here the issue is that the multiplicity of senders may ensure that receivers build a lot of different cookies no matter what *our* AI decides to do. Our AI could then choose to free ride on the efforts of these other senders. If the kind of cookie that our AI wants to exist will exist (within its catchment area) anyway, whether or not it pays for its construction, our AI has no reason to make the transfer of influence to alien AIs. In equilibrium, there would still be some amount of trade going on, since if the free riding were universal it would undermine its own possibility. However, the amount of trade in such an equilibrium might be very small, as potential senders adopt a mixed strategy that gives only a tiny chance of engaging in acausal trade. (The extent of this problem may depend on the size of the catchment areas, the number of plausible cookie recipes, and the cost of producing the various cookies; as well as on whether other kinds of acausal trade arrangements could mitigate the issue.)

To avoid such a potential free riding problem, we could embed an *exclusivity clause* in our cookie recipe. For example, we could specify our AI's porous value to require that, in order for a cookie to count as local fulfillment of the porous value, the cookie would have to be built specifically in order to trade with our AI (rather than in order to trade with some other AI, or for some other reason). Perhaps this could be explicated in terms of a counterfactual over our AI's preference function: something along the lines of a requirement that there be (in each Hubble volume, or produced by each independently-originating AI) one more cookie of type K than there would have been if instead of valuing type-K cookies our AI had valued (some other) type-K' cookies. This explication would, in turn, require a definition of the relevant counterfactual.[1] There may be other ideas for how to go about these things.

7 Utility diversification

The principle of value diversification might be attractive even aside from the motivations undergirding the Hail Mary approach. We can make a comparison to the task of specifying an epistemology or a prior probability function for our AI to use. One approach here would be to pick one particular prior, which we think have attractive properties, such as the Solomonoff prior (formulated in a particular base language). An alternate approach would be instead to use a mixture prior, a superposition of various different ideas about what shape the prior should take. Such a mixture prior might include, for example, various Solomonoff priors using different base-languages, some other prior based on computational depth, a speed prior, a prior that gives some positive finite probability to the universe being uncomputable or transfinite, and a bunch of other things [9, 2, 10]. One advantage of such an approach would be that it would reduce the risk that some important hypothesis that is actually true would be assigned zero or negligible probability in our favored formalization [5]. This advantage would come at a cost—the cost of assigning a lower probability to hypotheses that might really be more likely—but this cost might be relatively small if the agent using the prior has a superintelligence's abilities to gather and analyze data. Given such an agent, it may be more important that its prior does not absolutely prevent it from ever learning some important true hypothesis (the universe is uncomputable? we are not Boltzmann Brains?) than that its prior makes it maximally easy quickly to learn a plethora of smaller truths.

Analogously, to the extent that human values are resource-satiable, and the superintelligence has access to an astronomical resource endowment, it may be more important for us to ensure that the superintelligence places at least *some* weight on human values than to maximize the probability that it places no weight on anything else. Value diversification is one way to do this. Just as we could use a mixture prior in the epistemological component, we might use a "utility mixture" in the AI's utility function or goal specification. The formula (1) above suggests one way that this can be done, when we want to add a bounded component U_2 as a modulator of a possibly unbounded component U_1. Of course, we couldn't throw just *anything* into the hopper and still expect a good outcome: in particular, we would not want to add components that plausibly contain outright evil or anti-humane values. But as long as we're only adding value components that are at worst neutral, we should risk nothing more intolerable than some fractional dilution of the value of our cosmic endowment.

What about values that are not resource-satiable? Aggregative consequentialist theories, such as hedonistic utilitarianism, are not resource-satiable. According to those theories, the value added by creating one more happy mind is the same whether the extra mind is added onto an existing stock of 10 happy minds are 10 billion happy minds.[3] Nevertheless, even if the values we wanted our AI to pursue are in this sense insatiable, a (weaker) case might still be made for pursuing a more limited form of utility diversification. One reason is the vaguely contractualist considerations hinted at above. Another reason, also alluded to, is that it may often be possible, to some extent, to co-satisfy two different values in the very same physical structure (cf. [8]). Suppose, for example, that we believe that the value of the world is a linear function of the number of duck-like things it contains, but we're unsure whether "duck-like" means "walks like a duck" or "quacks like a duck". Then one option would be to randomly pick one of these properties, which would give us a 50% chance of having the world optimized for the maximally valuable pattern and a 50%

[3] Quite possibly, aggregated consequentialist theories remain insatiable even when we are considering scenarios in which infinite resources are available, since otherwise it would appear that such theories are unable to provide possible ethical guidance in those kinds of scenarios; and this might mean that they fail even in our excellent situation, as well as some positive probability is assigned to the world being canonically infinite.[4]

chance of having the world optimized for a pattern of zero value. But a better option would be to create a utility function that assigns utility both to things that walk like ducks and to things that quack like ducks. An AI with such a utility function might devise some structure that is reasonably efficient at satisfying both criteria simultaneously, so that we would get a pattern that is close to maximally valuable whether it's duck-like walking or duck-like quacking that really has value.[4]

An even better option would be to use *indirect normativity* [12, 6, 5] to define a utility function that assigned utility to whatever it is that "duck-like" really means—even if we ourselves are quite unsure—so that the AI would be motivated to investigate this question and then to optimize the world accordingly. However, this could turn out to be difficult to do; and utility diversification might then be a useful fallback. Or if we come up with several plausible ways of using indirect normativity, we could try to combine them using a mixture utility function.

8 Acknowledgements

I am grateful to Stuart Armstrong, Owen Cotton-Barratt, Daniel Dewey, and Jaan Tallinn for helpful discussions.

References

[1] Stuart Armstrong. Utility indifference. Technical report, Future of Humanity Institute, University of Oxford, 2010.

[2] Charles H Bennett. *Logical depth and physical complexity.* Springer, 1995.

[3] Nick Bostrom. Pascal's mugging. *Analysis*, 69(3):443–445, 2009.

[4] Nick Bostrom. Infinite ethics. *Analysis and Metaphysics*, (10):9–59, 2011.

[5] Nick Bostrom. *Superintelligence: Paths, dangers, strategies.* Oxford University Press, 2014.

[6] Daniel Dewey. Learning what to value. In *Artificial General Intelligence*, pages 309–314. Springer, 2011.

[7] P Richard G Layard, Alan Arthur Walters, and AA Walters. *Microeconomic theory.* McGraw-Hill New York, 1978.

[8] Toby Ord. Moral trade. *Ethics*, forthcoming, 2015.

[9] Jorma Rissanen. A universal prior for integers and estimation by minimum description length. *The Annals of Statistics*, pages 416–431, 1983.

[4] Probably no ducks, however! Value diversification is not a technique for specifying concepts that are hard to define. Rather, the idea is that we first do our best to define what we value (or to specify a value-loading mechanism). We might find that we fail to reach a consensus on a single definition or value-loading mechanism that we feel fully confident in. The principle of value diversification then suggests that we seek to conglomerate the leading candidates into one mixture utility function rather than putting all the chips on one favorite.

[10] Jürgen Schmidhuber. The speed prior: a new simplicity measure yielding near-optimal computable predictions. In *Computational Learning Theory*, pages 216–228. Springer, 2002.

[11] Carl Shulman. Omohundro's "basic ai drives" and catastrophic risks. *Manuscript. (intelligence.org/files/BasicAIDrives.pdf)*, 2010.

[12] Eliezer Yudkowsky. Coherent extrapolated volition. *Machine Intelligence Research Institute (May 2004). (intelligence.org/files/CEV.pdf)*, 2004.