The Most Important Question

C.X. Gonzalez

What is the most important question we can ask? Admittedly, this is a loaded question, but if we unpack it a bit perhaps we can find a satisfying answer. The key term here is "important." What do we mean by important? Questions of importance, and similarly, questions of value, and questions of what matters in the realm of human action fall into the domain of ethics. I don't intend to fully map out the major disagreements in the field of ethics, but there tends to be agreement on one issue: whatever it is that matters, whatever values that may or may not exist, in the absence of conscious creatures in the universe, all talk of values is as good as meaningless. What good are values, duties or virtues if there are no beings around to value it? So if there was a question whose immediate answer would prevent the ultimate moral disaster, the extinction of all known life, this would be a satisfying answer to our original question (assuming life is, on average, worth living). Which questions might yield answers which would help us avoid an extinction-level threat of the highest probability and urgency? The following are possible candidates:

1. What is the best path forward for mitigating the threat from a nuclear war?
2. What is the best path forward for mitigating the threat from a biological war?
3. What is the best path forward for mitigating the threat from climate change?

While each of these questions is valuable, the question I would like to focus on for the duration of this essay will focus on the development of super intelligent or artificial general intelligence (AGI) machines, machines whose levels of intelligence dwarf any human or even collection of humans across multiple domains. More precisely, I will be concerned with their decision procedures. If we understand moral or ethical theories as decision guiding procedures which help one pick out better and worse actions or states of affairs, we can state the question as what will be taken as the central question of this essay:

Central Question: Which moral system should superintelligent machines use?

How is this question related to existential threat? If the prospect of creating a god (or gods) in a machine isn't immediately concerning, allow me to motivate not only the urgency behind this question, but why this is the right question to be asking in the first place.

Superintelligence and Misaligned Values

Any mismatch between a superintelligent machine's goals and humanity's goals broadly speaking is potentially catastrophic. Take a superintelligent machine with the task of maximizing, say, paperclip production. On its face, this seems like an innocuous enough task to not warrant any grand suspicion or concern. However, it's possible that this paperclip maximizer

might proceed “by converting first the Earth and then increasingly large chunks of the observable universe into paperclips” (Bostrom, 2014). Soon enough, every usable inch of the universe within this super intelligence’s region of space will be filled with paperclip production facilities, humans be damned. This would, after all, maximize paperclips.

Speaking more generally we should recognize that, given any sufficiently unspecific goal, the space of all possible means of arriving at that goal is infinite. Furthermore, we must remind ourselves what is stipulated in granting that a machine is truly super intelligent. This machine will likely find optima which are so hyper-efficient so as to be inconceivable and unanticipated for any human or group of humans to predict. In fact, it is unlikely that human cognition, collective or not, will be able to zero in on the means that a superintelligent machine would settle on. In addition, this superintelligence would likely be intelligent enough to achieve those goals regardless of human stopgaps, or even mobility issues. If this machine is truly superintelligent, it will run through any and all walls on its way to achieving its goal as long as their goal is still physically possible. So this machine’s ultimate goal may manifest itself in ways that are worlds apart from the goals of humans. And many goals carried out to their logical extreme (as an optimizing AGI would try to do) may have a non-negligible chance of causing existential catastrophe. So even if the paperclip maximizer sounds unlikely, the takeaway is that *any* misaligned values between a superintelligence and humanity at large could spell disaster. Surely the number of ways of building unsafe AGI vastly outnumbers the number of ways of building safe AGI. And any non-negligible probability of existential threat must be taken seriously.

The Orthogonality Thesis

At this point it might be objected that any sufficiently *rational* agent will come to the same conclusions for questions of values and morality more broadly. If one believes morality is somehow grounded in *rationality*, or even that *rational* agents have a tendency to agree on the truth of the matter, then this is the right conclusion to come to.

On first appearance, this makes sense. Intelligent beings are rational creatures, and rational creatures *tend* to converge on the question of what does and does not matter in the moral domain. One would be hard pressed to find two truly rational people who disagree that mass genocide is a moral horror of the highest magnitude. While the idea that rationality and morality at least somewhat track one another seems plausible given convergence in moral reasoning over the past few hundred years (i.e. the abolition of slavery, the adoption of legal and moral rights into our vocabulary, etc.), it sadly rests on a misunderstanding of the type of rationality an artificially intelligent machine possesses.

The problem with this thinking is that it rests on vagueness of language. In the first paragraph of this section the term “rational” is used repeatedly, but without a singular meaning that encapsulates all uses of the word.

AI systems are, at their core, **instrumentally rational** (Bostrom, 2012). Not rational without qualifications. That is, given some objective (or value, in moral parlance) these machines will find optimal means of arriving at, or maximizing for that objective. By contrast, the rationality a Kantian or a similarly inclined moral realist (someone who posits objective moral facts) talks about is a type of rationality without qualifications, a kind of rationality we speak of

when we attribute it to a fellow human: rationality in some “thick” *normative* sense. This kind of rationality is notably different than the instrumental rationality that our superintelligent machine has and is a topic we will return to later.

Given that our AGI is rational only in the instrumental sense, we can safely conclude that we may plug in *any* arbitrary value we wish into its optimisation algorithm and expect hyper-efficient results. In short, an AGI’s value(s) and intelligence are entirely independent of each other. So any AGI can be plotted on a graph with the orthogonal axes of value and intelligence. This value independence is known as the *Orthogonality Thesis* (Bostrom, 2014).

While it’s unlikely superintelligent machines will be given a naively unconstrained goal such as maximizing paperclip production, the takeaway is that the wrong goal/value, when internalized by an AGI, can pose an existential-level threat to humanity. All it takes is getting this wrong just once for us to not have any second chances. Once a machine whose intelligence eclipses the collective brainpower of all of human history is made, any sufficiently poorly selected goal/value could entice this machine to view us as a minor obstacle on its way to its final objective. As such, the values an AGI uses in its decision procedure — it’s moral system — is of the utmost importance.

Value Alignment

To reflect on what’s been laid out thus far: Intelligence and values run in completely orthogonal directions. A superintelligent machine may be superintelligent with any possible value plugged in. We’ve also concluded that any mismatch between this machine and human values at large could potentially lead to catastrophic results. So value selection is a topic of great importance.

It is at this point that the discussion normally turns to questions of AI Value Alignment, as it is often referred to. **Value Alignment** is the general research project, both technical and philosophical, of finding out how we can align the values of intelligent machines with those of humanity. Value Alignment research has therefore focused heavily on the following *descriptive* question:

Descriptive Question: How can we align a superintelligent machine’s moral system with the moral system we humans *actually* use?

This question is usually tackled with a combination of various sophisticated machine learning techniques such as inverse reinforcement learning. However, these approaches to value alignment strike me as violence against any and all moral considerations. The question of value is never one of what *do* we value, but rather, what *should* we value? This applies in equal measure to considerations of which values to program into AGI. But before tackling the question of which values we *should* program into AGI I would first like to address the attempt at programing our *actual* values into a superintelligent machine and show how this attempt is ultimately undesirable.

As far as I can tell, there are three options for how to go about this project: **(a)** align the machine with the values of the masses, **(b)** find some universal value(s) nearly all humans hold, and program those values into the machine, or **(c)** program whichever value(s) humans would

converge on under ideal conditions like adequate knowledge, access to sufficient computational resources and being calm, cool, collected, etc.

Option **(a)** can be dismissed in relatively short order. I see no strong reason to suggest that there is wisdom in the masses when it comes to moral matters, especially when the stakes are at the level of existential threat. If history is any guide, group mentality often corrupts the minds of the masses and ideologies captivate the moral compass of the individual. While moral progress has certainly been made from a moral realist's perspective, we can never be sure which areas of contemporary values are the ones which will stand the test of time. We will almost certainly be considered moral monsters to our distant descendants.

As for option **(b)**, the history of moral philosophy provides no shortage of philosophers who claim to be putting forth a set of values which are both universal in nature and globally applicable. Possible contenders include hedonic pleasure, the avoidance of suffering, liberty, life, rule universalizability and so on. And it may very well be the case that, under the right specifications, some value(s) may be *claimed* by nearly all humans.

The least controversial contender for universal value might be that of avoiding suffering. Note that this is suffering with no silver lining or otherwise redeeming factor. This suffering does not help you in any way. That is to say, given two otherwise identical situations where situation **(1)** has a degree of added suffering with no upside and situation **(2)** does not, then, *all else equal*, one therefore has reason to prefer the state of affairs of situation **(1)** over that of situation **(2)**.

One possible objection to the claim that avoiding suffering is a truly universal value is that the existence of masochists disproves any claim to suffering-avoidance's universality. However, all that the proponent of suffering-avoidance has to do is to define "suffering" as any state which the sufferer would wish to cease. Under this definition, even a sadist would claim there is value in avoiding suffering. Therefore, suffering-avoidance is a value that can be claimed universally *by definition*. But does claiming the same values, at least nominally, mean that said value is actually shared?

If we carefully unpack what is being conveyed when someone makes a statement of value such as "I value life," we will find that this claim, when fully expanded, loses it's universalizability. We don't value life in the abstract without any qualification. We value life *for someone*. If we're being honest with ourselves, what we seem to be saying is a more expanded statement of the type "I value life for myself, those close to me, and to a lesser extent, complete strangers." Given this expanded state, it's probably still true that nearly all humans would honestly utter this claim verbatim. However, the antecedents to the pronouns "I", "those close to me", and even "complete strangers" vary on a case by case basis. So fully expanded, this statement of value becomes, "Atticus values life for Atticus, Scout, Jem, and to a lesser extent, complete strangers" for one person, but "Jon values life for Jon, Sansa, Bran, Arya, and to a lesser extent, complete strangers" for another. Therefore, the value actually held, when made *explicit*, is not universally held, even if the general grammatical structure might be the same. Replace the value of "life" with any other candidate universal value, and the argument holds all the same.

According to option **(c)** it could be argued that everyone would *converge* on the same values in idealized conditions such as access to all relevant information, access to sufficient

computational resources, being calm, cool, collected, etc (Smith, 2013). And it is these values that we could program into our AGI. While an interesting approach at grounding moral values in objective facts of the world, how plausible this claim appears, however, depends largely on one's own intuitions about it, a highly subjective matter. And if developments in early 20th century physics tell us anything, it's that intuition cannot be trusted in the pursuit of foundational truths. Ultimately, the veracity of claim **(c)** depends on facts about the world and is, therefore, a largely scientific question. Given some specific parameters as to what constitutes ideal conditions, we can, in theory, test whether the convergence thesis is true. But absent any scientific evidence of this kind, we can safely put aside this candidate for AI value alignment.

So none of the three candidates for aligning superintelligent machines with *actual* values seems very plausible. Considerations as to which values we *should* program into superintelligent machines is therefore where we should focus our attention. But given the added component of normativity in this question, we can finally rephrase the original **descriptive question** into what I claimed at the beginning of this paper is the right question to be asking and what will serve as the central question **(CQ)** of this essay:

Central Question: What *moral system should* superintelligent machines use?

Given the possibly enormous impact of superintelligent machines, the gravity of an adequate answer (as has been argued above) cannot be understated. Unfortunately, talk of rights, dignity, autonomy, moral status (Warren, 2000), moral agents (Sullins, 2006), etc. is too vague, *especially* when it comes time to actually implement these concepts into code. We must be sharp with our words when answering this question. And the best way to give an adequate answer to a question is to first understand it properly. In this case there are two key components worth unpacking: first, what do we mean by "should"? And second, what is a "moral system"? So for the rest of this paper I will seek to *clarify* the central question and briefly sketch how we might go about adequately answering it.

Normativity

In attempting to understand the word "should" it must be emphasized what I am *not* doing. I am *not* assuming that there are no current attempts at defining the term. Moral philosophers such as Moore and Hume have often contrasted the normative with the descriptive, for instance (Sayre-McCord, 2014). I am also not assuming that there *can be* no precise definitions. Instead, what I aim to show now is that all the possible paths one might take in any reasonable attempt to understand the term precisely are either circular or lead to the same conclusion. So in the following I will continually raise what I find to be the most natural questions to ask in our attempts at understanding the word "should", followed by the only natural responses I can see being offered to those questions.

The first question we may naturally ask, "'should' in what sense?" Borrowing from Kant, there seems to be two answers: read "should" in a strictly **moral** sense, or read "should" in a broadly **normative** sense. Take the following statement:

Claim: One should help those in need.

If we read this claim by parsing the word “should” in the **moral** sense of the word, we can rewrite it without any change in meaning as:

Moral Claim: One should morally help those in need regardless of one’s values.

If, however, we read this claim by parsing the word “should” in the **normative** sense of the word, we can rewrite it without any change in meaning as:

Normative Claim: One should help those in need given the value of charity.

Applying these two distinctions to **(CQ)** is a first step towards a more rigorous understanding of the question. If we parse the word “should” by using the **moral** sense of the word, the original question then becomes what I will call the central moral question **(CMQ)**:

Central Moral Question: What moral system *should* we morally program into superintelligent machines?

In a similar vein, we can apply the **normative** reading of the word “should” to yield what I will call the central normative question **(CNQ)**:

Central Normative Question: What moral system should we program into super intelligent machines given some set of values?

Beginning with **(CMQ)**. The answer to **(CMQ)** is, of course, whatever moral theory is correct. This response, however, immediately raises the question of: given a set of moral theories, according to what criteria *can* we choose between competing moral theories? “Can” because we first need to limit our search to criteria for selecting the correct moral system that us humans can *actually* use. The boundaries of this space are of course defined by those criteria that we can physically, and psychologically hold. The answer is, ostensibly, countless different criteria. But given that there are no other domains outside of the descriptive and the normative, we can pose the following two sub-questions: of all the available criteria we *can* choose from for assessing competing moral theories, **(CMQ.1)** which criteria *should* we use, and **(CMQ.2)** which criteria do we *actually* use? The response to **(CMQ.1),** of course, depends on what we mean by “should” which would loop us back to the original question of “‘should’ in what sense?” Answering **(CMQ.2)**, however, only leads us further down the rabbit hole.

There are many different criteria *actually* used in assessing moral systems. As an example, one philosopher lists off the following criteria for assessing moral systems: consistency, determinacy, applicability, intuitive appeal, internal support, external support, explanatory power and publicity (Timmons, 2013). The specific criteria one philosopher uses is not all that relevant. However, there might be some educational value in seeing the type of criteria one might use in assessing competing moral systems.

Let's say we have some set of criteria we wish to use to assess whichever moral system comes our way. Given this set of criteria we may ask, according to which criteria *can* we accept *those* criteria? At the risk of repeating myself we may respond, whichever criteria we *can* physically and psychologically hold. After which we can ask yet again the following two sub-questions: of all the possible criteria for assessing the validity of moral systems we *can* hold **(CMQ.2.1)** according to which criteria *should* we accept these criteria? And **(CMQ.2.2)** according to which criteria *do* we accept those criteria?

With any answer to **(CMQ2.1)** we are forced, yet again, back to the question of "'should' in what sense?" With any answer to **(CMQ.2.2)** we are forced into an infinite regress where we may ask the same pattern of can, followed by should/do questions. Given the circularity of using the word "should" in the moral sense, we are left no other option besides reading it in the broadly **normative** sense.

Recall that reading the word "should" in the **normative** sense implies that one should do an action only relative to some given value. So if we are to take "should" in the **normative** sense, we must first have some values to evaluate different possible actions against. We are now forced into asking the question, "what values are we using here?" (Note, we are limiting our current discussion to "highest values", or values all other values are derivative of).

In keeping with the same pattern above, we may ask, "what values *can* we use?" where "can" is again constrained by physical and psychological constraints. Given the set of all possible values we *can* hold at hand, we are then faced with the familiar two sub-questions: **(CNQ.1)** which values *should* we use? And **(CNQ.2)** which values *do* we use? The answer to **(CNQ.1)**, of course, depends on the original question of "'should' in what sense?" thus looping us back around to the starting point.

As far as I can tell there are two categories of responses to question **(CNQ.2)**: **(CNQ2.1)** claim that nearly all humans hold the same universal value towards which all of our other derivative values aim; or **(CNQ.2.2)** reject universal values, and claim that each individual has their own set of values, some of which differ between individuals and some of which coincide with others thus forming communities of overlapping or shared values. We can safely reject **(CNQ2.1)** on the same grounds argued above, namely, that any apparent universal values, sufficiently expanded, actually yield different values for different people. This forces us to conclude that the validity of any normative claim ultimately depends on which values the speaker of the normative claim holds. As a result we come to the realization that *there seems to be no basis upon which we can choose one value over another that doesn't already appeal to some prior assumed value in the first place*. This, on first appearance, might sound like value (and therefore, moral) relativism. However, we can apply further constraints to the values we *can* use to avoid a total reduction to relativism.

We can avoid this total collapse by noticing a curious fact. While each individual may have many different values depending on cultural backgrounds, upbringings, etc. there is one way of grounding all of us in this shared conversation in some set of constraints on which values we may adopt. The idea here is to take all possible sets of physically and psychologically possible values we may hold, and to eliminate some of these sets from contention by this constraint. But if this constraint exists, where might it come from? I believe there is one ground

value we can assume are shared values by everyone with whom we dialogue with about any matters of fact. Notice I am not saying “shared by all.”

To be explicit, we can assume the following: *by virtue of entering into an earnest dialogue which aims at uncovering some truth, both participants implicitly assume the constraints of the demands of rationality.* In other words, it seems that any time there is an honest attempt at a conversation where two individuals want to get to the truth of the matter, they are non-verbally agreeing to play in accordance with the rules of reason. If we find ourselves in debate with someone who, when backed into a corner, freely and unapologetically admits that their position is incoherent, contradictory and irrational, then there is simply nothing left to be said. That conversation should either end in short order or be reframed as no longer being about understanding what’s factual but rather, exchanging thoughts and beliefs for whatever reason.

So given this shared value of rationality, and the constraints that come with it, we can finally conclude how we are to understand the term “should.” Fully drawn out: we can parse the word “should” as a stand in for the **normative** sense of the word “if one values X, then one should do Y” where the values to be plugged in for X are those values which are restricted by i) which values we may physically hold, ii) which values we may psychologically hold, and finally, iii) which values we may rationally hold. Plugging in this new interpretation into **(CNQ)** yields the following question updated central normative question **(CNQv2)**:

Central Normative Question v2: What moral system should we program into superintelligent machines *if* we value a given set of values which are physically, psychologically and rationally possible to hold?

Despite the work done so far, there remain two further areas of clarification. First, we need to understand what a moral system precisely is and second, we need to understand what rationality is. Starting with the former.

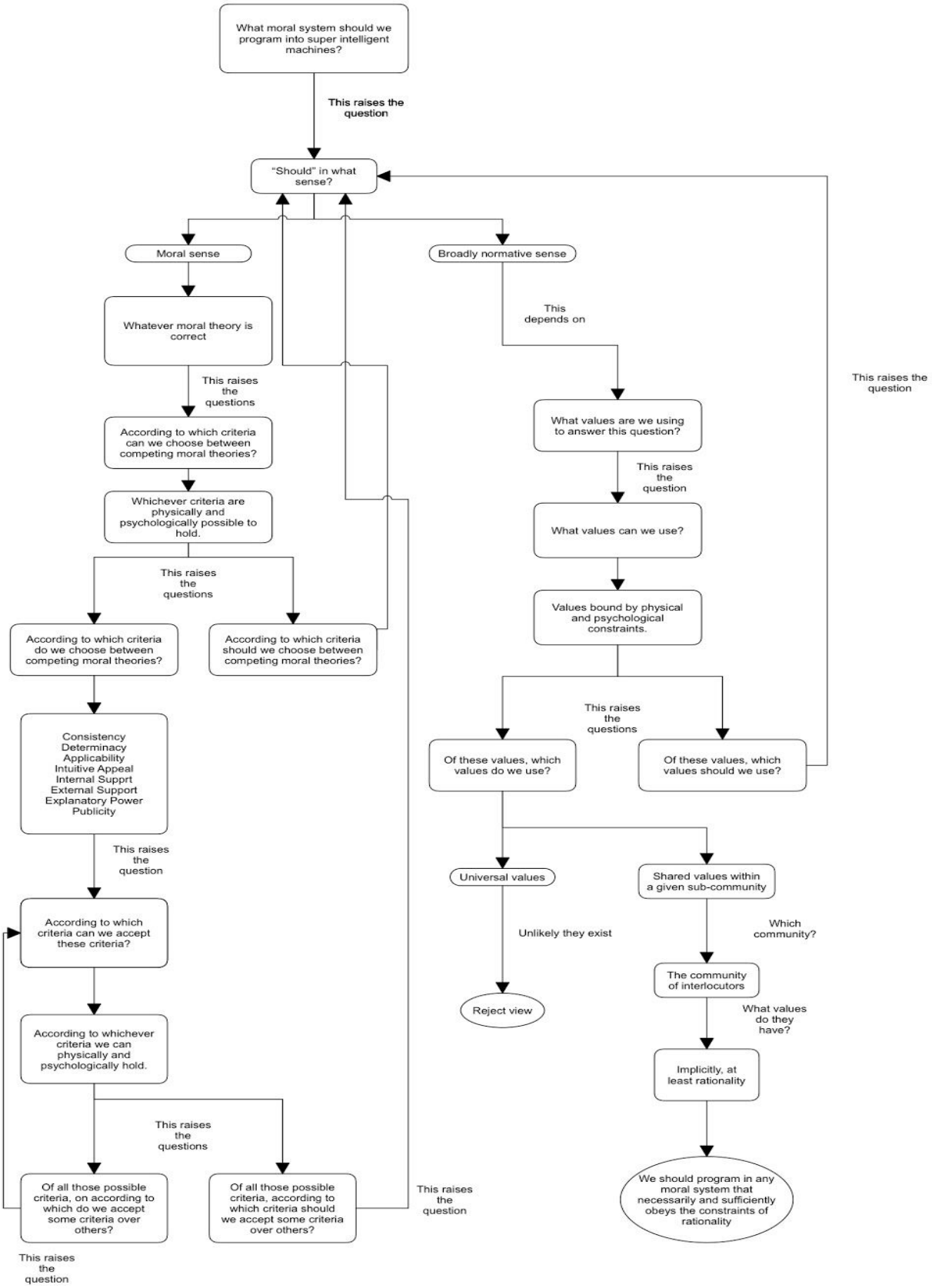

Figure 1: A flowchart of the underlying argument for how to best understand the term "should."

Moral Systems

Nearly every moral system in the analytic tradition (with the exception of moral particularism) shares a common structure: a small number of moral principles and definitions from which, in theory, all moral questions can be answered. It might be useful to think of these systems as classification algorithms whose inputs are non-moral facts about a given situation and from these non-moral facts, along with the assumed definitions and principles of the system, an output is assigned thus classifying a given action, intention or situation as right, wrong or permissible.

A classic example is total hedonic utilitarianism which first *defines* the word "good" as pleasure minus pain, and is followed by the principle that "an action is right if and only if it maximizes the total good for everyone, and is wrong otherwise." So from this definition-principle pair, along with rules of inferences such as those from classical first-order logic, moral derivations can be executed.

This sounds an awful lot like most logical systems. It is from noticing this similarity that I would like to propose the following observation: *moral systems are fundamentally axiomatic systems*. But they aren't *just* axiomatic systems. Three differences separate a moral system from any other axiomatic system one might find in a math or logic textbook.

First, this axiomatic system is used to *guide actions* by classifying them as right, wrong or permissible. Contrast this with, say, an axiomatic geometric system which is used to derive geometric truths within that system, or even an optimal strategy for winning a game such as tic-tac-toe. Second, morality seems to necessarily require an aura of *objectivity*. Regardless of whether one thinks moral facts are objective or not, we certainly seem to speak *as if* there were objective moral facts. And third, morality seems to require the feature of *practicality*. That is, a necessary connection between a moral belief, and motivation to act (Smith, 2013). This implies that if one makes a moral judgement that "giving to those in need is the right thing to do", then that person must also feel the urge to follow through with that statement, even if they ultimately don't do so. So while a moral system is still fundamentally an axiomatic system like any other from logic or mathematics, it is also *constrained* by these three features: action guidance, an aura of objectivity, and practicality. Without these constraints, we would just have another axiomatic system on our hands, not a moral one. So we can parse "moral system" in our primary question as "an action guiding, objective sounding, intrinsically motivating axiomatic system." This allows us to further modify **(CNQv2)** into the following:

Central Normative Question v3: What action guiding, objective sounding, intrinsically motivating axiomatic system should we program into superintelligent machines *if* we value a given set of values which are physically, psychologically and rationally possible to hold?

Rationality

My final remarks will be on sharpening our conception of rationality. Unfortunately, rationality is a tremendously complicated topic and cannot be given a full treatment given the scope of this essay. However, some brief remarks may be made. First, what I *don't* have in mind when I refer to rationality is **instrumental rationality**. **Instrumental rationality,** it should be recalled, is the capacity to process information in such a way so as to achieve a given goal as

optimally as possible. Second, there are many different competing notions of what rationality might entail. Derek Parfit, for instance, argued that holding a “future tuesday indifference” preference whereby someone likes to avoid pain just as anybody else, except for on Tuesdays where they don’t mind it despite the pain being phenomenological the same, is an irrational preference (Bostrom, 2014). Although others reject this view of rationality (Street, 2009). Third, whether one takes the principle that something good or desirable should be maximized as a principle of rationality or not is also a matter of debate (Foot, 1985; Gauthier, 1975). Clearly, many seemingly intuitive principles that one might think constitute basic principles of “rationality” are contested.

There is, however, one proposed principle of rationality that I think can be accepted without much controversy. This principle is that which dates back to Aristotle’s original work on logic, namely, *the principle of noncontradiction*. Abiding by this simple principle seems to be nearly universally held as sacred (excluding paraconsistent logic). As such, I will maintain the view that rationality must *at a minimum* contain the principle of noncontradiction. If rationality just consisted in internal consistency then this would be a tacit endorsement of the methodology of doing ethics called Reflective Equilibrium. This methodology seeks to, in short, “get one’s house in order” so to speak. Ethics, according to a proponent of Reflective Equilibrium, is simply an exercise in taking our moral intuitions and beliefs, ranking them in order of importance, and then finding some way to systematically make as many of them get along with each other as possible. Any conflicts must result in the moral intuition or belief of lesser importance being abandoned.

This approach to ethics allows for, in theory, multiple “islands” of internally consistent moral systems to exist. As long as my system is coherent, there is nothing you can say to me. This at least *allows* for a sort of moral relativism, whether or not that’s the necessary result of reflective equilibrium is, however, not certain.

There is one more principle of rationality I would like to propose. This principle would allow us to rule out certain systems on an empirical basis and would therefore allow us to actually make clear progress in moral philosophy. This principle is what I will call the principle of no self-defeaters **(PNS)**. **(PNS)** states:

Principle of No Self-Defeaters: Any action guiding principles which, when followed, lead to the cessation of agents following those guiding principles, are irrational action guiding principles.

An example of this might be a pacifist tribe in a tense war-hungry region of the world. By following pacifism, it’s quite likely that a neighboring bloodthirsty tribe takes advantage of this state of affairs and anilitates the pacifist tribes. If we assume that the pacifist tribe would have had the means of protecting themselves had they only abandoned their ways, then we can conclude that pacifism, at least in this thought experiment, is a self defeating action guiding principle as it led to the cessation of pacifism being practiced.

Similarly, Derek Parfit argued that ethical egoism, the ethical system which holds that an action is right if and only if it is broadly beneficial for the individual, is also self-defeating. So called “common sense morality” is also rejected for its potential self-defeating nature (Parfit, 1984). Interestingly, researchers at McGill University ran agent-based simulations and found

that populations of agents with either traitorous or selfish inclinations tended to collapse over time while humanitarian and ethnocentric populations flourished (Hartshorn, Kaznatcheev, Shultz, 2013). Perhaps in the future, empirical evidence will amount showing that some major moral system is also self-defeating.

Conclusion

It is with this final clarification that we can fully understand the primary question of this paper. **(CNQv3)** can be expanded into

Central Normative Question v4: What action guiding, objective sounding, intrinsically motivating axiomatic system should we program into superintelligent machines *if* we value a given set of values which are physically, psychologically, and logically consistently possible to hold without being self-defeating?

The final point worth emphasizing is that it is quite possible, and has been argued for before, that the set of values which we might plug in may be one of many (Street, 2009). It might be the case that there are multiple sets of values which are physically, psychologically and logically consistently possible to hold without being self-defeating. This is very much an open question, but is nonetheless a possibility worth emphasizing. It's also possible that given additional principles of rationality used in conjunction with the principle of noncontradiction and the principle of no self-defeaters, the number of differing sets of moral values may be reduced to fewer or even one set of values, some of which may share overlapping values/axioms. This is an avenue of further research.

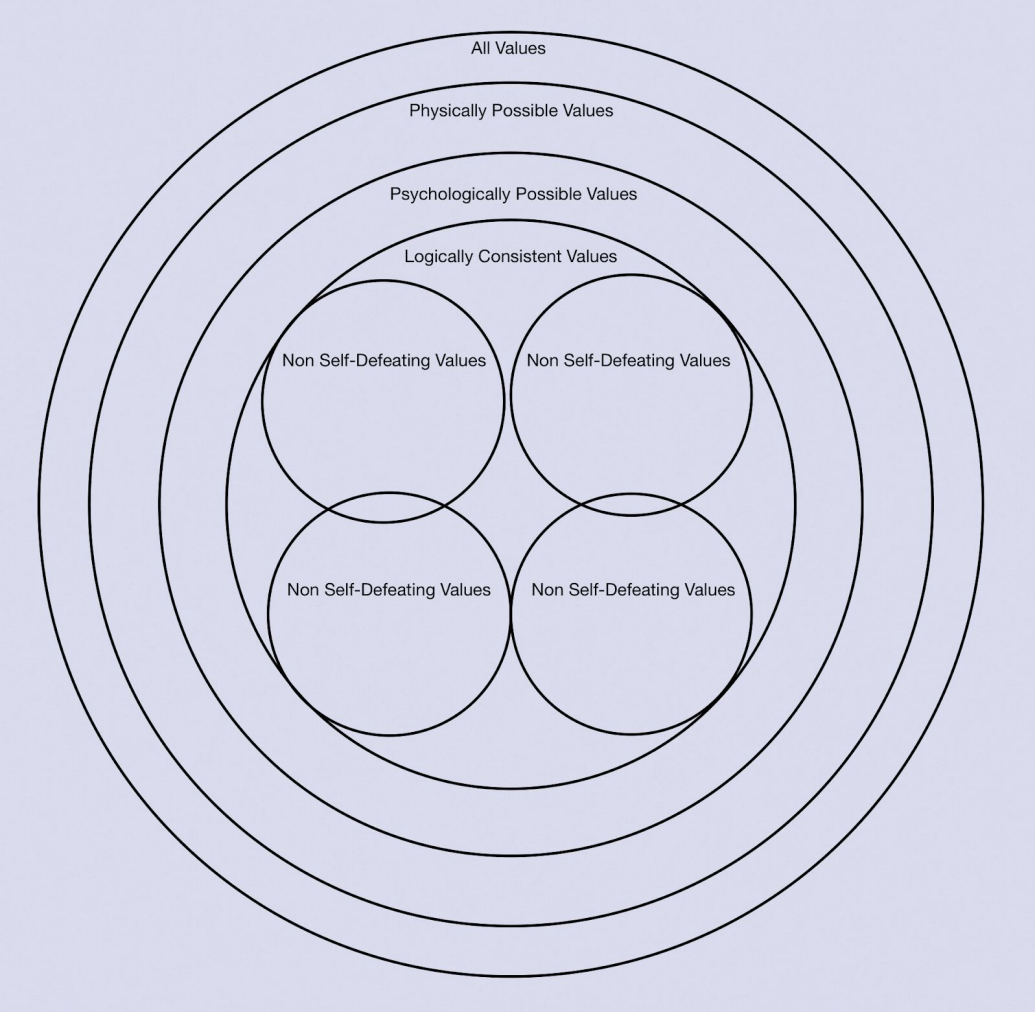

Figure 2: A visual depiction of acceptable sets of values where some systems may share overlapping values/axioms.

To summarize: superintelligent machines might prove to be a catastrophic invention by humanity. Any non-trivial existential risks must be taken seriously. Therefore, questions as to which action-guiding values or principles these machines are programmed with is of the utmost importance. Current value alignment research seems to miss the point by researching how to align machines' values with our *actual* values, instead of the values that we *should* have. The central question for anyone with this concern can be stated as "which moral system should we program into superintelligent machines?" Given extensive analysis we may convince ourselves that an adequate reading of the term "should" is one which takes it as a stand-in for the **normative** sense of the word, where the values we plug into that **normative** statement must be physically, psychologically and rationally possible to hold. Furthermore, I hope to have convinced the reader that we can understand "moral systems" to be axiomatic systems which have the constraints of being action-guiding, objective sounding and intrinsically motivating. Finally, we can *at minimum* take "rationally permitted" to be "lacking in logical contradiction."

Additional principles of rationality may be adopted too such as the principle of no self-defeaters. Putting these all together and parsing the question of primary importance we get the question: “What action guiding, objective sounding, intrinsically motivating axiomatic system should we program into superintelligent machines *if* we value a given set of values which are physically, psychologically and logically consistently possible to hold without being self-defeating?”

While I don’t expect to have convinced the reader of every nuance in my argument, I do hope that the general methodology of viewing moral theories as axiomatic systems whereby at least some of these axioms may be selected against by appeals to rationality is an attractive one.

Future areas of inquiry might include: **(a)** a more robust understanding of rationality and further constraints on possible moral values/axioms this understanding entails, **(b)** finding specific sets of values which match the aforementioned criteria of possible moral values/axioms and **(c)** making progress in answering the most precise possible version of the central question.

I believe that narrowing down possible moral systems by adding further constraints from rationality to what we allow in our consideration is a promising path forward. I hope that this analysis might prove to be a fruitful avenue for exploring what seems to be a question of the utmost importance. After all, this is philosophy with a deadline.

References:

Bostrom, Nick. Superintelligence: Paths, Dangers, Strategies. Oxford University Press, Inc. 2014. Partial Glossary.

Bostrom, Nick. *The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents*. Minds and Machines. 2012.

Foot, Philippa. Utilitarianism and the Virtues. Mind. April 1985.

Gauthier, David. Reason and Maximization. Canadian Journal of Philosophy. March 1975.

Hartshorn, Max, Kaznatcheev, Artem and Shultz, Thomas (2013) 'The Evolutionary Dominance of Ethnocentric Cooperation' Journal of Artificial Societies and Social Simulation 16 (3) 7 <http://jasss.soc.surrey.ac.uk/16/3/7.html>. doi: 10.18564/jasss.2176

Parfit, Derek. Reasons and Persons. Oxford University Press. 1984.

Priest, Graham, Tanaka, Koji and Weber, Zach, "Paraconsistent Logic", *The Stanford Encyclopedia of Philosophy* (Summer 2018 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2018/entries/logic-paraconsistent/>.

Sayre-McCord, Geoff, "Metaethics", *The Stanford Encyclopedia of Philosophy* (Summer 2014 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2014/entries/metaethics/>. Section 4.

Scheffler, Samuel. The Rejection of Consequentialism, Revised Edition. Oxford: Clarendon Press, 1994.

Smith, Michael. *Realism*. Ethical Theory: An Anthology, Second Edition. 2013.

Warren, Mary. Moral Status: Obligations to Persons and Other Living Things. Oxford University Press. April 2000.

Street, Sharon. In Defense of Future Tuesday Indifference: Ideally Coherent Eccentrics and the Contingency of What Matters. Philosophical Issues. 2009.

Sullins, John. When is a Robot a Moral Agent?. International Review of Information Ethics. December 2006.

Timmons, Mark. Moral Theory. Rowman & Littlefield Publishers, Inc. 2013.