Global Catastrophic Risk
INSTITUTE

http://gcrinstitute.org

On the Promotion of Safe and Socially Beneficial Artificial Intelligence

Seth D. Baum
Global Catastrophic Risk Institute

Global Catastrophic Risk Institute Working Paper 16-1

Cite as: Seth D. Baum, 2016, On the Promotion Of Safe and Socially Beneficial Artificial Intelligence, Global Catastrophic Risk Institute Working Paper 16-1

Note from the author: Comments welcome. Revisions expected. This version 29 July 2016.

On the Promotion of Safe and Socially Beneficial Artificial Intelligence

Seth D. Baum
Global Catastrophic Risk Institute

Abstract

This paper discusses means for promoting artificial intelligence (AI) that is designed to be safe and beneficial for society (or simply "beneficial AI"). The promotion of beneficial AI is a social challenge because it seeks to motivate AI developers to choose beneficial AI designs. Currently, the AI field is focused mainly on building AIs that are more capable, with little regard to social impacts. Two types of measures are available for encouraging the AI field to shift more towards building beneficial AI. Extrinsic measures impose constraints or incentives on AI researchers to induce them to pursue beneficial AI even if they do not want to. Intrinsic measures encourage AI researchers to want to pursue beneficial AI. Prior research focuses on extrinsic measures, but intrinsic measures are at least as important. Indeed, intrinsic factors can determine the success of extrinsic measures. Efforts to promote beneficial AI must consider intrinsic factors by studying the social psychology of AI research communities.

1. Introduction

The challenge of building technologies that are safe and beneficial for society is really two challenges in one. There is the technical challenge of developing safe and beneficial technology designs, and there is the social challenge of ensuring that such designs are used. The two challenges are interrelated. Motivating technologists to pursue safe and beneficial designs is itself a social challenge. Furthermore, motivating people to use safe and beneficial designs is made easier when the designs also have other attractive features such as low cost and ease of use; creating these features is a technical challenge.

This paper is concerned with the social challenge. Specifically, the paper examines a range of approaches to motivating technologists to pursue safe and beneficial technology designs. The paper focuses on artificial intelligence (AI) technologies, including both near-term AI and the proposed future "strong" or "superintelligent" AI that some posit could bring extreme social benefits or harms depending on its design. Much of the paper's discussion also applies to other technologies.

That AI has significant social impacts is now beyond question. AI is now being used in finance, medicine, military, transportation, and a range of other critical sectors. The impact is likely to grow over time as new technologies are adopted, such as autonomous vehicles and lethal autonomous weapons (unless the latter are banned or heavily restricted). The prospects for strong AI are controversial; this paper takes the position that the stakes are sufficiently high that it warrants careful attention even if the probability of achieving it appears to be low. Regardless, while the paper is motivated in part by the risk of strong AI, the insights are more general.[1]

For brevity, the paper uses the term "beneficial AI" to refer to AI that is safe and beneficial for society. It also uses the term "promoting beneficial AI" to refer to efforts to encourage technologists to design and build beneficial AI, or to have them avoid designing and building AI that is not beneficial. The technologists include AI researchers/designers/developers (the paper uses these terms more or less interchangeably) as well as adjacent personnel in management, business development, etc.

[1] For perspectives on near-term AI impacts, see e.g. Lin et al. (2011). For strong AI, see e.g. Eden et al. (2013).

The paper's implicit value judgment is that AI should be built so as to have net benefits for the whole of society—or, in the face of uncertainty, net expected benefits. This is to say that AI should not be built just for the sake of making it more capable or more intellectually interesting. Also, AI should not be built for the benefit of its builders if this comes at the expense of society as a whole. These positions may seem to cut against ideals of academic freedom, intellectual progress, and capitalist entrepreneurship. The paper takes the position that these ideals are only worth pursuing to the extent that doing so benefits society.

Were it the case that the field of AI was already focused on beneficial design, efforts to promote it would be unnecessary. Unfortunately, this is not the case. The field is largely focused on building systems that are more capable, regardless of whether this capability is used for social good. This tendency and the need to shift it is articulated, for example, by distinguished AI researcher Stuart Russell:

> I think the right approach is to build the issue [beneficial AI] directly into how practitioners define what they do. No one in civil engineering talks about "building bridges that don't fall down." They just call it "building bridges." Essentially all fusion researchers work on containment as a matter of course; uncontained fusion reactions just aren't useful. Right now we have to say "AI that is probably beneficial," but eventually that will just be called "AI." [We must] redirect the field away from its current goal of building pure intelligence for its own sake, regardless of the associated objectives and their consequences (Bohannon 2015:252).

This paper takes on the challenge of how to shift the AI field towards greater emphasis on social impacts. The paper reviews and critiques existing proposals for promoting beneficial AI and lays out a wider portfolio of techniques. A core criticism is that existing proposals neglect human psychology: they seek to influence AI researchers without thinking carefully about how AI researchers are influenced. Neglect of human psychology limits the portfolio of techniques that get considered for promoting beneficial AI and reduces the effectiveness of those techniques that are considered. In some cases, measures taken in ignorance of human psychology can even backfire, resulting in less beneficial AI than would have existed without any measures taken.

Broadly speaking, there are two types of measures for promoting beneficial AI. *Extrinsic measures* are imposed on AI designers from the outside so that they adopt beneficial designs even if they do not want to. These measures include constraints that require or forbid certain designs, incentives to encourage or discourage certain designs, and compliance measures to make sure that constraints or incentives are being followed. *Intrinsic measures* are cultivated within AI designers so that they want to adopt beneficial designs. These measures include the cultivation of social norms and the framing of communications. There can also be intrinsic effects of extrinsic measures, such as when a technology ban sparks backlash, making designers less interested in adopting beneficial designs. Extrinsic and intrinsic measures are discussed in Sections 2 and 3, respectively.[2]

Prior discussions of the promotion of beneficial AI focus almost exclusively on extrinsic measures.[3] However, both types of measures can help. Indeed, strategies based purely on extrinsic measures run a significant risk of having no net effect or even being counterproductive.

[2] On the extrinsic/intrinsic distinction, see e.g. Markowitz and Sharif (2012:246) and references therein.

[3] See Sotala and Yampolskiy (2015, section 3) for a review in the context of strong AI. Russell et al. (2015) also discuss a range of predominantly extrinsic measures. A notable exception to the focus on extrinsic measures is Russell's emphasis on shifting "how practitioners define what they do" (Bohannon 2015:252).

As this paper discusses, the success of extrinsic measures often depends heavily on intrinsic factors. Meanwhile, pure intrinsic strategies can be quite effective, as can hybrid extrinsic-intrinsic strategies. The bottom line is that the promotion of beneficial AI demands attention to human psychology.

2. Extrinsic Measures

2.1 Constraints

Constraints are perhaps the simplest means of promoting beneficial AI, and the most simplistic. The logic is direct: if a design feature is beneficial, require it; if it is harmful, ban it. A ban on dangerous AI technologies is implicit in Joy's (2000) call for relinquishment of dangerous AI, and it is explicit in other work (Posner 2004; Wilson 2013; Yampolskiy and Fox 2013). Requirements for beneficial AI designs are less common in discussions of AI. Requirements could be used to insist that AI developers adopt certain beneficial designs such as: verification, validity, security, and control (Russell et al. 2015) and avoiding negative side effects, avoiding reward hacking, scalable oversight, safe exploration, and robustness to distributional shift (Amodei et al. 2016).

When constraints work, they guarantee that AI designs are beneficial. However, they also limit the freedom and flexibility of AI designers. This can provoke backlash by AI designers, which is one example of an intrinsic effect of extrinsic measures (Section 2.4). Even without backlash, enacting constraints can require extensive institutional and political changes, which makes them difficult to implement.

Constraints pose other challenges as well. One is that unless they are carefully designed, they can unwittingly constrain the wrong features, resulting in AI that is less beneficial. Designing successful AI constraints can thus require close interaction between AI experts and policy makers. A related issue is that constraints may need to be constantly updated as AI technology evolves. An AI design attribute that was harmful in early AI may be beneficial in later AI, and vice versa. New design attributes will also emerge; these could merit new constraints. One potential solution is to phrase constraints in more general terms (Moses 2007); for AI, this could mean requiring AI designers to select the most beneficial available design. Such an approach makes constraints more durable as AI technology evolves, but it comes at the expense of making it more difficult to verify compliance.

2.2 Incentives

Incentives are the primary extrinsic alternative to constraints. Unlike constraints, incentives let AI developers keep the freedom to pursue whatever designs they desire. Incentives act by changing the rewards or penalties for specific designs, so as to push developers in different design directions. The AI literature has focused mainly on monetary incentives, such as by offering funding for beneficial AI research (McGinnis 2010) or by making AI companies pay compensation when found liable for the consequences of harmful AI (Gurney 2012). However, incentives can also take on other forms, such as social praise/scorn or professional advancement/sanction for developing beneficial/harmful AI.

Incentives hold several advantages over constraints. By giving AI designers more freedom, they are less likely to provoke backlash, which can make them easier to implement.[4] Policy

[4] Incentives can nonetheless provoke significant backlash. For example, in the United States, environmentalists have long pursued incentive-based policies such as taxes on pollution in order to appeal to industry interests that do not

makers can also avoid the need to identify beneficial design attributes by applying incentives to completed technologies, as in liability schemes (which impose penalties for AIs that turn out to be harmful) and in prize competitions (which can offer rewards for AIs that turn out to be beneficial).[5]

The core disadvantage of incentives is that they do not guarantee that beneficial AI designs would be chosen. An AI developer could simply choose to forgo the reward or pay the penalty and continue to develop harmful AI. The logical response to this is to strengthen the incentive, though this can provoke more backlash and can even erode the distinction between incentives and constraints. Indeed, a constraint could be defined as an incentive with an infinite or maximal reward/penalty.

2.3 Compliance

Constraints and incentives are generally built on the premise that AI designers do not want to choose beneficial designs. Otherwise, they would not need to be constrained or incentivized. AI designers thus have reason to avoid complying with the constraint or the incentive. When this is the case, mechanisms for achieving compliance are needed, including mechanisms for monitoring for noncompliance and mechanisms for enforcing penalties for noncompliance.

A simple approach to monitoring is to require AI groups to submit research proposals to review boards prior to conducting the research. AI review boards could be analogous to the review boards that already exist at many universities and other institutions for reviewing medical and social science research (Yampolskiy and Fox 2013). Existing review boards are focused mainly on harms that could be caused by the conduct of the research, in particular through abuse of human research subjects. AI review boards would need an expanded scope that includes the societal impacts of the products of research. Such an expansion would be in line with a more general expansion of research ethics to include ethical assumptions embedded within the research (such as ethical positions implicit in AI objective functions) and ethical aspects of the societal impacts of research (Schienke et al. 2009; 2011).

One challenge for the review boards proposal is that some AI groups may not be at institutions that have review boards and thus could go undetected. Hard-to-monitor groups can include private companies, especially startup companies, and groups in unregulated countries. Indeed, there is some concern that national AI regulations could simply push AI research to unregulated countries. This problem can be addressed via international AI treaties (Posner 2004; Wilson 2013), though this is easier said than done. Another approach in some AI monitoring proposals is to implement a draconian mass surveillance regime in order to find any harmful AI group wherever they are (e.g., Shulman 2009). Suffice to say, such surveillance poses extreme problems for privacy, intellectual property, and trustful geopolitical relations. It is a downside of extrinsic measures that such problematic surveillance mechanisms would even be considered.

If monitoring succeeds and harmful AI groups are identified, the next step is to enforce whatever penalty is to be applied.[6] Enforcement should in general be less of a challenge than monitoring because AI groups have limited means of resisting penalties. Government penalties can be imposed through the threat or application of force. Institutional penalties can be imposed via the threat or application of measures such as firing non-compliant personnel. These sorts of

want constraints, yet industry has been largely successful at avoiding these incentive-based policies.

[5] Incentives for completed technologies are less relevant for AIs that could be catastrophic because there may be no penalty that could adequately compensate for damages and, in the extreme case, no one alive to process the penalty.

[6] Conversely, when beneficial AI groups are identified, rewards are to be applied, though this is less of a challenge because AI groups are likely to seek rewards, not dodge them.

actions could succeed at achieving compliance to extrinsic measures, but, in addition to being intrinsically regrettable (i.e., one should not want AI developers to lose their jobs or suffer physical harm), they can also alienate AI developers, provoke backlash, and motivate them to relocate to unregulated places. The net effect of the typical extrinsic measure is to unwittingly create an antagonistic relationship between AI developers and those who seek beneficial AI, which makes beneficial AI more difficult to achieve. The essential solution to this predicament is to consider intrinsic factors, i.e. the psychology of AI developers.

2.4 Intrinsic Aspects of Extrinsic Measures

Consider two different extrinsic measures: a ban on flag burning and a requirement that dog owners clean up after their dogs. Flag burning is legal in several countries, including the United States. The US has repeatedly considered a constitutional amendment to ban flag burning. Such a ban has never passed, but here is one analysis of what would happen if it did:

> Few people have burned the American flag in recent years, and it is reasonable to suppose that a constitutional amendment making it possible to criminalize flag burning would have among its principal consequences a dramatic increase in annual acts of flag burning. In fact, adopting a constitutional amendment may be the best possible way to promote the incidence of flag burning (Sunstein 1996:2023).

Why would a ban on flag burning increase the rate of flag burning? One potential mechanism is that the ban would draw attention to flag burning, which is otherwise something that not many people think about. Some portion of people who think about it may then go on to do it. Another potential mechanism is that the ban changes the social meaning of flag burning. Without the ban, flag burning is seen as distasteful and anti-patriotic, whereas with the ban, flag burning becomes a patriotic rebellion against a bad law.

The story of dog cleanup is exactly the opposite. The following describes the effect of dog clean-up laws in Berkeley. Similar effects have been observed in other locations, such as New York City (Krantz et al. 2008).

> After the Berkeley town council enacted an ordinance requiring owners to clean up after their dogs the sidewalks became much cleaner, even though officials never issued citations for breaking the law. The law apparently tipped the balance in favor of informal enforcement. Citizens became more aggressive about complaining to inconsiderate dog owners, and, anticipating this fact, dog owners became more considerate (Cooter 2000:11).

The dog cleanup story is notable because *it achieved positive outcomes without enforcing compliance*. There was no draconian surveillance, and no need to worry about dog owners relocating to places that lacked cleanup laws. Instead, the law prompted dog owners and their neighbors to police themselves.

The point of the comparison between flag burning and dog cleanup is that people can react in different ways to different extrinsic measures, and that this can significantly affect outcomes. Indeed, how people react can be the difference between negative outcomes (i.e., the extrinsic measure is counterproductive) and positive outcomes. These two cases are directly applicable to extrinsic measures for promoting beneficial AI. In short, extrinsic measures for promoting

beneficial AI should strive to be like dog cleanup, not like flag burning. This means that extrinsic measures should aim to be considered desirable to AI developers—measures should be something that AI developers would want to comply with, not something they would want to push back against. In order to figure out what the effect of an extrinsic measure would be, it is necessary to consider not just the extrinsic measure itself, but also the intrinsic factors, i.e. the social psychology of AI communities.

A fuller accounting of intrinsic factors is presented in Section 3, but here some intrinsic factors that are specific to extrinsic measures. One recurrent finding is that monetary incentives can reduce intrinsic motivation (Deci 1971; Vohs et al. 2006). This means that once the money is gone, people become less motivated to perform some task than they would have been if there never was any money. In contrast, social praise and encouragement can increase intrinsic motivation (Deci 1971). Finally, carefully designed extrinsic measures can use a psychological phenomenon called cognitive dissonance to increase people's intrinsic motivation for a given type of activity (Dickerson et al. 1992; Section 3.6).

3. Intrinsic Factors and Intrinsic Measures

This section presents a variety of intrinsic phenomena of relevance to promoting beneficial AI. Some of the phenomena are dedicated intrinsic measures for promoting beneficial AI, and some of the phenomena are factors that can play a role in a range of intrinsic and extrinsic measures. All of the phenomena are oriented towards motivating AI developers to want to pursue beneficial designs.

3.1 Social Context and Social Meaning

People often behave differently depending on the social setting or context that they are in. For example, some people go to the library to study because the presence of other studious people compels them to focus more, whereas if they stayed at home, they would let themselves get distracted.

Efforts to promote beneficial AI can be more effective in certain social contexts. For example, some beneficial AI measures benefit from cooperation across AI groups in order to avoid some groups choosing harmful designs that give them competitive advantage. Research in other contexts has found that it is often easier to achieve cooperation when people are together in a group than when they are in isolation (Krantz et al. 2008). Therefore, workshops, conferences, and other meetings, or even group phone calls could all be more effective at achieving cooperation among AI groups than private conversations or written statements that would be read in private.

Second, promoting beneficial AI can be more effective in social contexts where beneficial AI is considered desirable. For any given AI researcher, he or she is more likely to support beneficial AI if he or she is in a group of people who openly support beneficial AI than if he or she is alone or in a group of people who do not openly support beneficial AI. Support for beneficial AI can thus be expanded by creating and expanding groups of open beneficial AI supporters. Openness is important: other people will not be influenced by the group's support for beneficial AI if they are not in any way aware of this support.

Related to social context is the concept of social meaning. An act or idea can have a different meaning depending on the social context. For example, the act of flag burning can have an anti-patriotic meaning if it is not banned or a patriotic meaning if it is banned (Section 2.4). Efforts to promote beneficial AI should aim for it to have a positive social meaning. This can be

accomplished, for example, by testing prospective measures using focus groups of AI researchers in order to gauge their reactions.

3.2 Social norms

Norms are that which is considered normal. Social norms are norms held by groups of people about the normal behaviors of people in that group, including which behaviors normally are practiced ("descriptive norms") and which behaviors normally should be practiced ("injunctive norms") (Lapinski and Rimal 2005). Norms can vary from group to group and place to place. For example, in some places, it is normal for pedestrians to cross the street whenever it is safe, whereas in other places, it is normal to wait for the streetlight to turn green. Norms can also change over time. For example, slavery was once considered normal throughout much of the world, but this norm has reversed in many places. Specific social contexts can also activate certain social norms. For example, the same group of people may show different norms in a library than in a nightclub.

Beneficial AI can be a norm, i.e. concern for social impacts can be considered normal among AI developers. Beneficial AI as a social norm is implicitly at the heart of Stuart Russell's call for the AI community to abandon "its current goal of building pure intelligence for its own sake, regardless of the associated objectives and their consequences" and instead "build the issue [beneficial AI] directly into how practitioners define what they do" (Bohannon 2015:252). It is difficult to overstate the importance of social norms for beneficial AI. If beneficial AI is considered normal, then it will be easier to achieve compliance on extrinsic measures and easier to succeed with intrinsic measures, and there will be less need for either sort of measure in the first place because many AI researchers will already be pursuing beneficial designs.

How can one go about shifting social norms in AI? Answering this question would benefit from dedicated research on AI social norms, but some insights can be gained from other issues. For example, Posner (2000:1784-1785) lists several ways in which social norms for paying taxes can be strengthened, including showing that other people also pay taxes, creating social sanctions for not paying taxes,[7] and reminding people of their civic obligation to pay taxes. Schultz et al. (2007) find that descriptive norms messages can reduce deviance but can also have a counterproductive "boomerang" effect for people with better-than-normal behavior, and that this effect can be attenuated with injunctive norms messages of social approval for good behavior. These approaches could be adapted for promoting beneficial AI by showing that other AI researchers also support it, creating social sanctions for those who do not support it and social approval for those who do, and cultivating a sense of duty for AI researchers to attend to the social impacts of AI.

3.3 Messengers and Allies

In promoting beneficial AI, it is not just *what is said* that matters, but also *who says it*. This is because people interpret meanings differently depending on who is conveying a message. This holds for AI researchers just as much as it does for anyone else. The fact that this happens cuts against the scientific ideal of objectivity, but scientists are humans too, and try as we might to avoid it, the identity of the messenger still matters for how we react to messages.

One important class of messenger is the fellow AI researcher. Prior research has found that when conveying messages about ethics and social impacts to young scientists (e.g., graduate

[7] Social sanctions are an extrinsic measure, specifically an incentive using a social penalty, though they can also cultivate certain social norms.

students or post-docs), it is important that the messages be delivered by established researchers in that field instead of by outside ethics professionals (Schienke et al. 2009; 2011). Using in-field researchers shows that ethics and social impacts is something that "we" (i.e., people in the field) care about, and is not just something that "they" (i.e., people outside the field) want "us" to care about. This is reflective of a more general tendency for people to respond better to messages from other people in their "in group". Thus, to the extent possible, messages should be selected from respected AI researchers. The quotes in this paper from Stuart Russell offer one example of this.

Sometimes, using "out-group" messengers can also succeed. This can occur when the out-group is seen as having some sort of high status. For example, funders, institutional leadership, and policy makers can fit this role because they have some control over AI researchers' professional success and some credentials for setting social norms and research directions. Celebrities—including academic and business celebrities like Stephen Hawking and Bill Gates—can also fit this role because they can be perceived as successful, important, and influential. It is important that these people deliver thoughtful messages so as to not come off as "ignorant blowhards", but when they do deliver thoughtful messages, their influence can be substantial.

As AI becomes more widely used across society, new out-group allies will also emerge. For example, the automotive industry is currently applying AI to autonomous vehicles. Automobiles must be safe, otherwise they will not sell and manufacturers can face steep liability claims. The automotive industry likewise has a safety culture that is currently pushing back against the AI culture of rapid product development and post-launch debugging. As Ford CEO Mark Fields put it in a recent interview, "You can't hit control-alt-delete when you're going 70 miles an hour" (Griffith 2016). Insofar as AI researchers would like the business opportunities of autonomous vehicles, they may be motivated to listen to the safety messages from messengers like Fields.

Another important potential ally could be found in militaries. This may seem surprising, since militaries are associated with violence and potential misuse of AI. However, militaries can be influential to AI researchers because they provide extensive AI research funding. Furthermore, militaries can be more safety conscious than is commonly believed. One study of AI researcher opinion found that a slight majority viewed the United States military as the most likely institution to produce harmful AI, but that "experts who estimated that the US military scenario is relatively safe noted that the US military faces strong moral constraints, has experience handling security issues, and is very reluctant to develop technologies that may backfire (such as biological weapons)" (Baum et al. 2011:193). For these reasons, military officials will often be motivated to promote beneficial AI. For their messages to resonate with AI researchers, they must achieve trust, which could be difficult given AI researchers' negative perceptions of the military. If trust can be achieved, then the military could offer another powerful ally in efforts to promote beneficial AI.

3.4 Framing

To frame is to present a message in a certain way. Framing is a matter of not *what is said* but *how it is said*. Skillful communication frames messages to achieve certain effects for certain audiences. For example, climate change is commonly framed as an environmental issue, which resonates with liberals more than with conservatives. For conservative audiences, climate change is sometimes framed as a threat to national security or to the economy, or framed as an injustice that compels religious duty (Shome and Marx 2009). Using the wrong frame for the wrong audience can lead people to reject a cause that they might otherwise support.

AI technologies can be framed in a variety of ways as well. Unfortunately, existing messages about beneficial AI are not always framed well. One potentially counterproductive frame is the framing of strong AI as a powerful winner-takes-all technology. This frame is implicit (and sometimes explicit) in discussions of how different AI groups might race to be the first to build strong AI (e.g., Shulman 2009; Armstrong et al. 2016). The problem with this frame is that it makes a supposedly dangerous technology seem desirable. If strong AI is a winner-takes-all technologies race, then AI groups will want to join the race and rush to be the first to win. This is exactly the opposite of what the discussions of strong AI races generally advocate—they postulate (quite reasonably) that the rush to win the race could compel AI groups to skimp on safety measures, thereby increasing the probability of dangerous outcomes. Instead of framing strong AI as a winner-takes-all race, those who are concerned about this technology should frame it as a dangerous and reckless pursuit that would quite likely kill the people who make it. AI groups may have some desire for the power that might accrue to whoever builds strong AI, but they presumably also desire to not be killed in the process.

Another potentially counterproductive frame is the framing of AI researchers as people who do not want to pursue beneficial designs. This framing is implicit in the existing literature's emphasis on extrinsic measures: extrinsic measures are used because AI researchers would not want to pursue beneficial designs. In the worst case, heavy-handed extrinsic measures could counterproductively instill a social norm of AI researchers not pursuing beneficial measures. This would be a reaction like "I didn't think I was someone who ignores social impacts, but since you mention it, I guess I am." Light penalties, cooperative relationships, and positive framing of AI researchers could make them more inclined to pursue beneficial designs.

Finally, extreme proposals like draconian global surveillance can inadvertently frame efforts to promote beneficial AI as being the problem, not the solution. In other words, it could give the impression that the efforts are misguided and causing more harm than good. The potential for aggressive beneficial AI efforts to be perceived as conspiratorial should not be discounted. Conspiracy theories are already prominent in the perceptions of global warming held by policy makers and the public (Lewandowsky et al. 2015). If AI succumbs to similar conspiracy theories, this could make it more difficult to promote beneficial AI. And even without conspiracy theories, floating extreme proposals can make the beneficial AI cause seem at best out of touch, and at worst outright harmful.

3.5 Stigmatization

Stigmatization is a type of framing oriented towards making an object or an activity feel socially undesirable or even taboo. Stigmatization can be an effective technique for preventing the use of dangerous technologies. For example, stigmatization has been used repeatedly for international arms control, most notably to achieve the 1997 Ottawa Treaty banning landmines and the 2008 Convention on Cluster Munitions banning cluster bombs, and currently to promote nuclear disarmament (Borrie 2014). The experience with landmines and cluster munitions is notable in part because the treaties have wider compliance than they have ratification. That is, some countries (e.g., the United States) have not ratified the treaties, *yet they still act in compliance with them, even though they have no legal obligation to do so*. The effort to stigmatize landmines and cluster munitions was so successful that the legal requirements are not necessary to achieve the social goal.

The international community's experience with stigmatization could be applied to dangerous AI. Stigmatizing dangerous AI can help build support for extrinsic measures such as national

regulations and international treaties. However, even in the absence of any extrinsic measures, stigmatization can still lead people to avoid building dangerous AI. A successful stigmatization effort causes people to not want to do the stigmatized activity. Stigmatization thus complements extrinsic measures. Indeed, it is difficult to imagine bans on dangerous AI technologies succeeding without an effective stigmatization effort.

In order for stigmatization to work, it must be based on a convincing argument. The landmine and cluster munitions campaigns were so successful because there was a sound moral and legal argument against these weapons, specifically that they cause indiscriminate harm to civilian populations. This argument was crucial for convincing countries to reject the weapons even though they had previously been accustomed to using them. Likewise, any attempts to stigmatize specific AI technologies must be based on some compelling reason. The potential for an AI technology to cause a massive global catastrophe is an example of such a reason.

Another challenge of stigmatization is that it can be alienating to those who disagree with it and/or those who are involved in a stigmatized activity. People do not like to think of themselves as being involved in a stigmatized activity and can resent being accused. This is seen, for example, in current debates about nuclear weapons, in which the nuclear-armed states distance themselves from efforts to stigmatize nuclear weapons even while they share an underlying concern about the weapons' catastrophic impacts. Likewise, efforts to stigmatize harmful AI should distinguish between the harms of the AI and the character of the AI designer so that the AI designers know that they are not seen as bad people and that they would be embraced if they switch to beneficial designs.

3.6 Cognitive Dissonance

Cognitive dissonance occurs when a person holds conflicting beliefs in her or his mind. People typically seek to resolve the dissonance of conflicting beliefs by rejecting one of them. For example, people might reject reports that a seemingly good person committed a terrible harm on grounds that "He or she couldn't possibly have done that".

An example of relevance to beneficial AI is in the relation between economic activity and beliefs about climate change. Around 2008, public belief in the scientific evidence of climate change declined. One explanation for the decline is that the economic recession induced cognitive dissonance. In response to the recession, people want the economy to grow. However, economic activity typically increases greenhouse gas emissions, thereby worsening climate change. This creates dissonance between the belief that there should be more economic growth and the belief that climate change is a problem. Some data suggests that some people handled this dissonance by rejecting the scientific evidence of climate change, even though the evidence itself is about the natural environment, not the economy (Scruggs and Benegal 2012).

Similarly, cognitive dissonance could lead AI researchers to reject claims that AI could be harmful. The potential for harmful AI could imply that AI research should be restricted, bringing AI researchers diminished intellectual freedom and business opportunities and in some cases can even threatening their livelihoods. Just as people may reject the science of climate change when the economy is bad, AI researchers may reject evidence or argument about harmful AI when their welfare is at stake. Like all people, AI researchers can have "motivated reasoning" in which they are motivated not by a goal of accuracy but instead by other goals, such as the goal of believing that they are a good person (Kunda 1990). Therefore, in order to improve salience, messaging about beneficial AI should strive to be sympathetic to AI researchers' intellectual and

professional interests, and extrinsic and intrinsic measures alike should strive to minimize the intellectual and professional downsides that AI researchers could face.

Cognitive dissonance can also be used to promote certain beneficial activities. Dickerson et al. (1992) studied the combined effect of people making a public commitment to conserving water and people being told that they had taken long showers. People took shorter showers only when they made a public commitment and were told they had taken long showers. If only one of the two conditions were present, they took longer showers. The explanation is that people had a cognitive dissonance between their public commitment and their self-perception of taking long showers, which they resolved by taking shorter showers. Similar effects have been observed in other contexts, with the strongest effect coming when people make public commitments or advocacy of an action and then are privately reminded of their own failures to perform that action (Stone and Fernandez 2008). This model could be adapted for beneficial AI by having AI researchers make public commitments to beneficial AI (such as via professional societies or in classrooms) and then being privately informed that some of their designs are not beneficial. This technique is likely to be more effective if an injunctive social norm for beneficial AI is already in place, because then AI researchers would be motivated to resolve the cognitive dissonance in favor of more beneficial AI.

4. Conclusion

As the societal impacts of AI continue to increase, it becomes more and more important to promote the development of AI that is safe and beneficial to society—abbreviated throughout this paper as "beneficial AI".

Thus far, discussions of how to promote beneficial AI have focused mainly on extrinsic measures that are imposed on AI designers even if they do not want to pursue beneficial AI. These measures come in the form of constraints and incentives, and they are often accompanied by measures for monitoring and enforcing compliance. Extrinsic measures can be successful at promoting beneficial AI, but they can be difficult to implement and can be resisted by AI developers. The success of extrinsic measures can also depend heavily on intrinsic factors, i.e. on how AI developers react to the measures. If the reaction is favorable, AI developers could comply on their own without external monitoring and enforcement. Alternatively, if the reaction is unfavorable, AI developers could even pursue less beneficial designs than they would if there were no extrinsic measures in place.

Meanwhile, a range of dedicated intrinsic measures are available; these encourage AI developers to want to pursue beneficial designs. Social norms be shifted towards caring about beneficial design. Messengers can be selected and messages can be framed to resonate with AI developers and entice them to want to pursue beneficial designs. Harmful AI designs can be stigmatized such that AI developers want to avoid them. AI developers can make commitments to choosing beneficial designs that can, via cognitive dissonance, lead them to do so.

This paper draws heavily from research on other issues due to a lack of prior research on intrinsic aspects of beneficial AI. The paper makes especially heavy use of research on environmental issues because these issues have seen robust social and psychological research. Many insights from these other issues apply to beneficial AI, but beneficial AI will inevitably have its own unique characteristics. Therefore, dedicated research on the social psychology of AI research communities is needed to understand the effectiveness of both extrinsic and intrinsic measures. Such research should be included in broader research agendas for beneficial AI.

One potential objection to intrinsic measures is that they are unreliable because they depend on each AI researcher to cooperate. There is some truth to this. However, extrinsic measures can also be unreliable—hence the effectiveness of extrinsic measures can depend on intrinsic factors. Regardless, 100% success is an inappropriate goal. The aim of any measure should be to reduce the harms and increase the benefits of AI to society. A measure that does this should be pursued, even if it still leaves some potential for harm or for loss of benefit. Given the stakes involved in AI, all effective measures for promoting beneficial AI should be pursued.

Acknowledgments
This paper has benefited from conversations with many people including Dario Amodei, Miles Brundage, Sean Legassick, Richard Mallah, and Jaan Tallinn, and from time spent at the Columbia University Center for Research on Environmental Decisions. Tony Barrett, Steven Umbrello, and Peter Howe provided helpful comments on an earlier draft. Any errors or shortcomings in the paper are the author's alone. Work on this paper was funded in part by a grant from the Future of Life Institute. The views in this paper are the author's and are not necessarily the views of the Future of Life Institute or the Global Catastrophic Risk Institute.

References
Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D (2016) Concrete problems in AI safety, arXiv:1606.06565
Armstrong S, Bostrom N, Shulman C (2016) Racing to the precipice: a model of artificial intelligence development. AI & Society 31(2):201-206
Baum SD, Goertzel B, Goertzel TG, 2011. How long until human-level AI? results from an expert assessment. Technological Forecasting & Social Change 78(1):185-195
Bohannon J (2015) Fears of an AI pioneer. Science 349(6245):252
Borrie J (2014) Humanitarian reframing of nuclear weapons and the logic of a ban. International Affairs 90(3):625-646
Deci EL (1971) Effects of externally mediated rewards on intrinsic motivation. J. Pers. Soc. Psychol. 18:105-115
Dickerson CA, Thibodeau R, Aronson E, Miller D (1992) Using cognitive dissonance to encourage water conservation. J. Appl. Soc. Psychol. 22:841-854
Eden AH, Moor JH, Soraker JH, Steinhart E (2013) Singularity hypotheses: a scientific and philosophical assessment. Berlin, Springer
Griffith E (2016) Who will build the next great car company? Fortune Magazine
Gurney JK (2013) Sue my car not me: products liability and accidents involving autonomous vehicles. University of Illinois Journal of Law, Technology & Policy 2013(2):247-277
Hughes J (2007) Global technology regulation and potentially apocalyptic technological threats. In Allhoff F, Lin P, Moor J, Weckert J (eds) Nanoethics: the ethical and social implications of nanotechnology. John Wiley & Sons, Hoboken
Joy B (2000) Why the future doesn't need us. Wired
Krantz DH, Peterson N, Arora P, Milch K, Orlove B (2008) Individual values and social goals in environmental decision making. In Kugler T, Smith JC, Connolly T, Son YJ (eds) Decision modeling and behavior in complex and uncertain environments New York, Springer, 165-198
Kunda Z (1990) The case for motivated reasoning. Psychological Bulletin 108(3):480-498

Lapinski MK, Rimal RN (2005) An explication of social norms. Communication Theory 15(2):127-147
Lewandowsky S, Cook J, Oberauer K, Brophy S, Lloyd EA, Marriott M (2015) Recurrent fury: conspiratorial discourse in the blogosphere triggered by research on the role of conspiracist ideation in climate denial. Journal of Social and Political Psychology 3(1):142-178
Lin P, Abney K, Bekey GA (eds) (2011) Robot ethics: the ethical and social implications of robotics. MIT Press, Cambridge, MA
Markowitz EM, Shariff AF (2012) Climate change and moral judgement. Nature Climate Change 2(4):243-247
McGinnis JO (2010) Accelerating AI. Northwestern University Law Review 104:366-381
Moses LB (2007) Recurring dilemmas: the law's race to keep up with technological change. University of Illinois Journal of Law, Technology & Policy 2007(2):239-285
Oreskes N, Conway EM (2010) Merchants of doubt: how a handful of scientists obscured the truth on issues from tobacco smoke to global warming. New York, Bloomsbury
Posner EA (2000) Law and social norms: the case of tax compliance. Virginia Law Review 86:1781-1819
Russell S, Dewey D, Tegmark M (2015) Research priorities for robust and beneficial artificial intelligence. AI Magazine 36(4):105-114
Schienke EW, Baum SD, Tuana N, Davis KJ, Keller K (2011). Intrinsic ethics regarding integrated assessment models for climate management. Science and Engineering Ethics 17(3):503-523
Schienke EW, Tuana N, Brown DA, Davis KJ, Keller K, Shortle JS, Stickler M, Baum SD (2009). The role of the NSF Broader Impacts Criterion in enhancing research ethics pedagogy. Social Epistemology 23(3-4):317-336
Schultz PW, Nolan JM, Cialdini RB, Goldstein NJ, Griskevicius V (2007) The constructive, destructive, and reconstructive power of social norms. Psychological Science 18(5):429-434
Scruggs L, Benegal S (2012) Declining public concern about climate change: can we blame the great recession? Global Environmental Change 22(2):505-515
Shome D, Marx S (2009) The psychology of climate change communication: a guide for scientists, journalists, educators, political aides, and the interested public. Columbia University Center for Research on Environmental Decisions, New York
Shulman C (2009) Arms control and intelligence explosions. In 7th European Conference on Computing and Philosophy (ECAP), Bellaterra, Spain, July 2-4
Stone J, Fernandez NC (2008) To practice what we preach: the use of hypocrisy and cognitive dissonance to motivate behavior change. Social and Personality Psychology Compass 2(2):1024-1051
Sunstein CR (1996). On the expressive function of law. University of Pennsylvania Law Review 144(5):2021-2053
Vohs KD, Mead NL, Goode MR (2006). The psychological consequences of money. Science 314:1154-1156
Wilson G (2013). Minimizing global catastrophic and existential risks from emerging technologies through international law. Virginia Environmental Law Journal 31:307-364