information

MDPI

Article

Countering Superintelligence Misinformation

Seth D. Baum

Global Catastrophic Risk Institute, PO Box 40364, Washington, DC 20016, USA; seth@gcrinstitute.org

Received: 9 September 2018; Accepted: 26 September 2018; Published: 30 September 2018

Abstract: Superintelligence is a potential type of future artificial intelligence (AI) that is significantly more intelligent than humans in all major respects. If built, superintelligence could be a transformative event, with potential consequences that are massively beneficial or catastrophic. Meanwhile, the prospect of superintelligence is the subject of major ongoing debate, which includes a significant amount of misinformation. Superintelligence misinformation is potentially dangerous, ultimately leading bad decisions by the would-be developers of superintelligence and those who influence them. This paper surveys strategies to counter superintelligence misinformation. Two types of strategies are examined: strategies to prevent the spread of superintelligence misinformation and strategies to correct it after it has spread. In general, misinformation can be difficult to correct, suggesting a high value of strategies to prevent it. This paper is the first extended study of superintelligence misinformation. It draws heavily on the study of misinformation in psychology, political science, and related fields, especially misinformation about global warming. The strategies proposed can be applied to lay public attention to superintelligence, AI education programs, and efforts to build expert consensus.

Keywords: artificial intelligence; superintelligence; misinformation

1. Introduction

At present, there is an active scholarly and public debate regarding the future prospect of artificial superintelligence (henceforth just *superintelligence*), which is artificial intelligence (AI) that is significantly more intelligent than humans in all major respects. While much of the issue remains unsettled, some specific arguments are clearly incorrect, and as such can qualify as *misinformation*. (As is elaborated below, arguments can qualify as misinformation even when the issues are unsettled.) More generally, misinformation can be defined as "false or inaccurate information" [1], or as "information that is initially presented as true but later found to be false" [2] (p. 1). This paper addresses the question of what can be done to reduce the spread of and belief in superintelligence misinformation.

While any misinformation is problematic, superintelligence misinformation is especially worrisome due to the high stakes involved. If built, superintelligence could have transformative consequences, which could be either massively beneficial or catastrophic. Catastrophe is more likely to come from a superintelligence built based on the wrong ideas—and it could also come from *not* building a superintelligence that would have been based on the *right* ideas, because a well-designed superintelligence could prevent other types of catastrophe, such that abstaining from building such a superintelligence could result in catastrophe. Thus, the very survival of the human species could depend on avoiding or rejecting superintelligence misinformation. Furthermore, the high stakes of superintelligence have the potential to motivate major efforts to attempt to build it or to prevent others from doing so. Such efforts could include massive investments or restrictive regulations on research and development (R&D), or plausibly even international conflict. It is important for these sorts of efforts to be based on the best available understanding of superintelligence.

Superintelligence is also an issue that attracts a substantial amount of misinformation. The abundance of misinformation may be due to the many high-profile portrayals of superintelligence

in science fiction, the tendency for popular media to circulate casual comments about superintelligence made by various celebrities, and the relatively low profile of more careful scholarly analyses. Whatever the cause, experts and others often find themselves responding to some common misunderstandings [3–9].

There is also potential for superintelligence *disinformation*: misinformation with the intent to deceive. There is a decades-long history of private industry and anti-regulation ideologues promulgating falsehoods about socio-technological issues in order to avoid government regulations. This practice was pioneered by the tobacco industry in the 1950s and has since been adopted by other industries including fossil fuels and industrial chemicals [10,11]. AI is increasingly important for corporate profits and thus could be a new area of anti-regulatory disinformation [12]. The history of corporate disinformation and the massive amounts of profit potentially at stake suggest that superintelligence disinformation campaigns could be funded at a large scale and could be a major factor in the overall issue. Superintelligence disinformation could potentially come from other sources as well, such as governments or even concerned citizens seeking to steer superintelligence debates and practices in particular directions.

Finally, there is the subtler matter of the information that has not yet been established as misinformation, but is nonetheless incorrect. This misinformation is the subject of ongoing scholarly debates. Active superintelligence debates consider whether superintelligence will or will not be built, whether it will or will not be dangerous, and a number of other conflicting possibilities. Clearly, some of these positions are false and thus can qualify as misinformation. For example, claims that superintelligence will be built and that it will not be built cannot both be correct. However, it is not presently known which positions are false, and there is often no expert consensus on which positions are likely to be false. While the concept of misinformation is typically associated with information that is more obviously false, it nonetheless applies to these subtler cases, which can indeed be "information that is initially presented as true but later found to be false". Likewise, countering misinformation presents a similar challenge regardless of whether the misinformation is spread before or after expert consensus is reached (though, as discussed below, expert consensus can be an important factor).

In practical terms, the question then is what to do about it. There have been a number of attempts to reply to superintelligence misinformation in order to set the record straight [3–9]. However, to the best of the present author's knowledge, aside from a brief discussion in [12], there have been no efforts to examine the most effective ways of countering superintelligence misinformation. Given the potential importance of the matter, a more careful examination is warranted. That is the purpose of this paper. The paper's discussion is relevant to public debates about superintelligence, to AI education programs (e.g., in university computer science departments), and to efforts to build expert consensus about superintelligence.

In the absence of dedicated literature on superintelligence misinformation, this paper draws heavily on the more extensive research literature studying misinformation about other topics, especially global warming (e.g., [10,13,14]), as well as the general literature on misinformation in psychology, cognitive science, political science, sociology, and related fields (for reviews, see [2,15]). This paper synthesizes insights from these literatures and applies them to the particular circumstances of superintelligence. The paper is part of a broader effort to develop the social science of superintelligence by leveraging insights from other issues [12,16].

The paper is organized as follows. Section 2 presents some examples of superintelligence misinformation, in order to further motivate the overall discussion. Section 3 surveys the major actors and audiences (i.e., the senders and receivers) of superintelligence misinformation, in order to provide some strategic guidance. Section 4 presents several approaches for preventing the spread of superintelligence misinformation. Section 5 presents approaches for countering superintelligence misinformation that has already spread. Section 6 concludes.

2. Examples of Superintelligence Misinformation

It is often difficult to evaluate which information about superintelligence is false. This is because superintelligence is a possible future technology that may be substantially different from anything that currently exists, and because it is the subject of a relatively small amount of study. For comparison, other studies of misinformation have looked at such matters as whether Barack Obama was born in the United States, whether childhood vaccines cause autism, and whether Listerine prevents colds and sore throats [17]. In each of these cases, there is clear and compelling evidence pointing in one direction or the other (the evidence clearly indicates that Obama was born in the US, that vaccines do not cause autism, and that Listerine does not prevent colds or sore throats, despite many claims to the contrary in all three cases). Therefore, an extra degree of caution is warranted when considering whether a particular claim about superintelligence qualifies as misinformation.

That said, some statements about superintelligence are clearly false. For example, this statement from Steven Pinker: "As far as I know, there are no projects to build an AGI, not just because it would be commercially dubious, but also because the concept is barely coherent" [18]. The acronym AGI stands for artificial general intelligence, which is a form of AI closely associated with superintelligence. Essentially, AGI is AI that is capable of reasoning across a wide range of domains. AGI may be difficult to build, but the concept is very much coherent. Indeed, it has a substantial intellectual history and ongoing study [19], including a dedicated research journal (*Journal of Artificial General Intelligence*) and professional society (the Artificial General Intelligence Society). Furthermore, there are indeed projects to build AGI—one recent survey identifies 45, spread across many countries and institutions, including many for-profit corporations, the largest of which being DeepMind, acquired by Google in 2014 for £400 million, the Human Brain Project, an international project with $1 billion in funding from the European Commission, and OpenAI, a nonprofit with $1 billion in pledged funding [20]. (DeepMind and OpenAI explicitly identify as working on AGI. The Human Brain Project does not, but it is working on simulating the human brain, which is considered to be a subfield of AGI [19].) There is even an AGI project at Pinker's own university. (Pinker and the AGI project MicroPsi [21] are both at Harvard University.) Therefore, in the quoted statement, the "as far as I know" part may well be true, but the rest is clearly false. This particular point of misinformation is significant because it conveys the false impression that AGI (and superintelligence) is a nonissue, when in fact it is a very real and ongoing subject of R&D.

A more controversial matter is the debate on the importance of consciousness to superintelligence. Searle [22] argues that computers cannot be conscious and therefore, at least in a sense, cannot be intelligent, and likewise cannot have motivation to destroy humanity. Similar arguments have been made by Logan [23], for example. A counterargument is that the important part is not the consciousness a computer but its capacity to affect the world [4,24,25]. It has also been argued that AI could be harmful to humanity even if it is not specifically motivated to do so, because the AI could assess humanity as being in the way of it achieving some other goal [25,26]. The fact that AI has already shown the capacity to outperform humans in some domains is suggestive of the possibility for it to outperform humans in a wider range of domains, regardless of whether the AI is conscious. However, this is an ongoing area of debate, and indeed Chalmers [24] (p. 16) writes "I do not think the matter can be regarded as entirely settled". Regardless, there must be misinformation on one side or the other: computers either can be conscious or they cannot, and consciousness either matters for superintelligence or it does not. Additionally, many parties to the debate maintain that those who believe that consciousness or conscious motivation matter are misinformed [4,5,7–9], though it is not the purpose of this paper to referee this debate.

There are even subtler debates among experts who believe in the prospect of superintelligence. For example, Bostrom [25] worries that it would be difficult to test the safety of a superintelligence because it could trick its human safety testers into believing it is safe (the "treacherous turn"), while Goertzel [27] proposes that the safety testing for a superintelligence would not be so difficult because the AI could be tested before it becomes superintelligent (the "sordid stumble"; the term is

from [28]). Essentially, Bostrom argues that an AI would become capable of deceiving humans before humans realize it is unsafe, whereas Goertzel argues the opposite. Only one of these views can be correct; the other would qualify as misinformation. More precisely, only one of these views can be correct for a given AI system—it is possible that some AI systems could execute a treacherous turn while others would make a sordid stumble. Which view is more plausible is a matter of ongoing study [28,29]. This debate is important because it factors significantly into the riskiness of attempting to build a superintelligence.

Many more additional examples could be presented, such as on the dimensionality of intelligence [3], the rate of progress in AI [7,8], the structure of AI goals [6–8], and the relationship between human and AI styles of thinking [6,8]. However, this is not the space for a detailed survey. Instead, the focus of this paper is on what to do about the misinformation. Likewise, this paper does not wish to take positions on open debates about superintelligence. Some positions may be more compelling, but arguments for or against them are tangential to this paper's aim of reducing the preponderance of misinformation. In other words, this paper strives to be largely neutral on which information about superintelligence happens to be true or false. The above remarks by Pinker will occasionally be used as an example of superintelligence misinformation because they are so clearly false, whereas the falsity of other claims is more ambiguous.

The above examples suggest two types of superintelligence misinformation: information that is already clearly false and information that may later be found to be false. In practice, there may be more of a continuum of how clearly true or false a piece of information is. Nonetheless, this distinction can be a useful construct for efforts to address superintelligence misinformation. The clearly false information can be addressed with the same techniques that are used for standard cases of misinformation, such as Obama's place of birth. The not-yet-resolved information requires more careful analysis, including basic research about superintelligence, but it can nonetheless leverage some insights from the misinformation literature.

The fact that superintelligence is full of not-yet-resolved information is important in its own right, and it has broader implications for superintelligence misinformation. Specifically, the extent of expert consensus is an important factor in the wider salience of misinformation. This matter is discussed in more detail below. Therefore, while this paper is mainly concerned with the type of misinformation that is clearly false, it will consider both types. With that in mind, the paper now starts to examine strategies for countering superintelligence misinformation.

3. Actors and Audiences

Some purveyors of superintelligence misinformation can be more consequential than others. Ditto for the audiences for superintelligence misinformation. This is important to bear in mind because it provides strategic direction to any efforts to counter the misinformation. Therefore, this section reviews who the important actors and audiences may be.

Among the most important are the R&D groups that may be building superintelligence. While they can be influential sources of ideas about superintelligence, they may be especially important as audiences. For example, if they are misinformed regarding the treacherous turn vs. the sordid stumble, then they could fail to correctly assess the riskiness of their AI system.

Also important are the institutions that support the R&D. At present, most AGI R&D groups are based in either for-profit corporations or universities, and some also receive government funding [20]. Regulatory bodies within these institutions could ensure that R&D projects are proceeding safely, such as via university research review boards [30,31]. Successful regulation depends on being well-informed about the nature of AGI and superintelligence and its prospects and risks. The same applies to R&D funding decisions by institutional funders, private donors, and others. Additionally, while governments are not presently major developers of AGI, except indirectly as funders, they could become important developers should they later decide to do so, and they meanwhile can play important roles in regulation and in facilitating discussion across R&D groups.

Corporations are of particular note due to their long history of spreading misinformation about their own technologies, in particular to convey the impression that the technologies are safer than they actually are [10]. These corporate actors often wield enormous resources and have a correspondingly large effect on the overall issue, either directly or by sponsoring industry-aligned think tanks, writers, and other intermediaries. At this time, there are only hints of such behavior by AI corporations, but the profitability of AI and other factors suggest the potential for much more [12].

Thought leaders on superintelligence are another significant group. In addition to the aforementioned groups, this also includes people working on other aspects of superintelligence, such as safety and policy issues, as well as people working on other (non-superintelligence) forms of AI, and public intellectuals and celebrities. These are all people who can have outsized influence when they comment on superintelligence. That influence can be on the broader public, as well as in quieter conversations with AGI/superintelligence R&D groups, would-be regulators, and other major decision-makers.

Finally, there is the lay public. The role of the public in superintelligence may be reduced due to the issue being driven by technology R&D that (for now at least) occurs primarily in the private sector. However, the public can play roles as citizens of governments that might regulate the R&D and as consumers of products of the corporations that host the R&D. The significance of the public for superintelligence is not well established at this time.

While the above groups are presented in approximate order of importance, it would not be appropriate to formally rank them. What matters is not the importance of the group but the quality of the opportunity that one has to reduce misinformation. This will tend to vary heavily by the circumstances of whoever is seeking to reduce the extent of superintelligence misinformation.

With that in mind, the paper now turns to strategies for reducing superintelligence misinformation.

4. Preventing Superintelligence Misinformation

The cliché "an ounce of prevention is worth a pound of cure" may well be an understatement for misinformation. An extensive empirical literature finds that once misinformation enters into someone's mind, it can be very difficult to remove.

Early experiments showed that people can even make use of information that they acknowledge to be false. In these experiments, people were told a story and then were explained that some information in the story is false. When asked, subjects would correctly acknowledge the information to be false, but they would also use it in retelling the story as if it were true. For example, the story could be a fire caused by volatile chemicals, and then it is later explained that there were no volatile chemicals present. Subjects would acknowledge that the volatile chemicals were absent but then cite them as the cause of the fire. This is logically incoherent. The fact that people do this speaks to the cognitive durability that misinformation can have [32,33].

The root of the matter appears to be that human memory does not simply write and overwrite like computer memory. Corrected misinformation does not vanish. Ecker et al. [15] trace this to the conflicting needs for memory stability and flexibility:

> Human memory is faced with the conundrum of maintaining stable memory representations (which is the whole point of having a memory in the first place) while also allowing for flexible modulation of memory representations to keep up-to-date with reality. Memory has evolved to achieve both of these aims, and hence it does not work like a blackboard: Outdated things are rarely actually wiped out and over-written; instead, they tend to linger in the background, and access to them is only gradually lost. [15] (p. 15)

There are some techniques for reducing the cognitive salience of misinformation; these are discussed in detail below. However, in many cases, it would be highly desirable to simply avoid the misinformation in the first place. Therefore, this section presents some strategies for preventing superintelligence misinformation.

The ideas for preventing superintelligence misinformation are inevitably more speculative than those for correcting it. There are two reasons for this. One is that the correction of misinformation has been the subject of a relatively extensive literature, while the prevention of misinformation has received fairly little scholarly attention. (Rare examples of studies on preventing misinformation are [34,35].) The other reason is that the correction of misinformation is largely cognitive and thus conducive to simple laboratory experiments, whereas the prevention of misinformation is largely sociological and thus requires a more complex and case-specific analysis. Nonetheless, given the importance of preventing superintelligence misinformation, it is important to consider potential strategies for doing so.

4.1. Educate Prominent Voices about Superintelligence

Perhaps the most straightforward approach to preventing superintelligence misinformation is to educate people who have prominent voices in discussions about superintelligence. The aim here is to give them a more accurate understanding of superintelligence so that they can pass that along to their respective audiences. Prominent voices about superintelligence can include select scholars, celebrities, or journalists, among others.

Educating the prominent may be easier said than done. For starters, they can be difficult to access, due to busy schedules and multitudes of other voices competing for their attention. Additionally, some of them they may already believe superintelligence misinformation, especially those who are already spreading it. Misinformation is difficult to correct in general, and may be even more difficult to correct for busy people who lack the mental attention to revise their thinking. (See Section 5.4 for further discussion of this point.) People already spreading misinformation may seem to be ideal candidates for educational efforts, in order to persuade them to change their tune, but it may actually be more productive to engage with people who have not yet made up their minds. Regardless, there is no universal formula for this sort of engagement, and the best opportunities may often be a matter of particular circumstance.

One model that may be of some value is the effort to improve the understanding of global warming among broadcast meteorologists. Broadcast meteorologists are for many people the primary messenger of environmental science. Furthermore, as a group, meteorologists (broadcast and non-broadcast) have traditionally been more skeptical about global warming than most of their peers in other Earth sciences [36,37]. In light of this, several efforts have been made to provide broadcast meteorologists with a better understanding of climate science, in hopes that they would pass this on to their audiences (e.g., [38,39]).

The case of broadcast meteorologists has important parallels to the many AI computer scientists who do not specialize in AGI or superintelligence. Both groups have expertise on a topic that is closely related to, but not quite the same as, the topic at hand. Broadcast meteorologists' expertise is weather, whereas global warming is about climate. (Weather concerns the day-to-day fluctuations in meteorological conditions, whereas climate concerns the long-term trends. An important distinction is that while weather can only be forecast a few days in advance, climate can be forecasted years or decades in advance.) Similarly, most AI computer scientists focus on AI that has "narrow" intelligence (intelligence in a limited range of domains), not AGI. Additionally, broadcast meteorologists and narrow AI computer scientists are often asked to voice their views on climate change and AGI, respectively.

4.2. Create Reputational Costs for Misinformers

When prominent voices cannot be persuaded to change their minds, they can at least be punished for getting it wrong. Legal punishment is possible in select cases (Section 4.5). However, reputational punishment is almost always possible and has potential to be quite effective, especially for public intellectuals whose brands depend on a good intellectual reputation.

In an analysis of US healthcare policy debates, Nyhan [40] concludes that correcting misinformation is extremely difficult and that increasing reputational costs may be more effective.

Nyhan [40] identifies misinformation that was critical to two healthcare debates: in the 1990s, the false claim that the policy proposed by President Bill Clinton would prevent people from keeping their current doctors, and in the 2000s, the false claim that the policy proposed by President Barack Obama would have established government "death panels" to deny life-sustaining coverage to the elderly. Nyhan [40] traces this misinformation to Betsy McCaughey, a scholar and politician generally allied with US conservative politics and opposed to these healthcare policy proposals:

> "Until the media stops giving so much attention to misinformers, elites on both sides will often succeed in creating misperceptions, especially among sympathetic partisans. And once such beliefs take hold, few good options exist to counter them—correcting misperceptions is simply too difficult. The most effective approach may therefore be for concerned scholars, citizens, and journalists to (a) create negative publicity for the elites who are promoting misinformation, increasing the costs of making false claims in the public sphere, and (b) pressure the media to stop providing coverage to serial dissemblers". [40] (p. l6)

Nyhan [40] further notes that while McCaughey's false claims were widely praised in the 1990s, including with a National Magazine Award, they were heavily criticized in the 2000s, damaging her reputation and likely reducing the spread of the misinformation.

There is some evidence indicating the possibility that reputational threats can succeed at reducing misinformation. Nyhan and Reifler [34] sent a randomized group of US state legislators a series of letters warning them about the reputational and electoral harms that the legislators could face if an independent fact checker (specifically, PolitiFact) finds them to make false statements. The study found that the legislators receiving the warnings were significantly less likely to make false statements. This finding is especially applicable to superintelligence misinformation spread by politicians, whose statements are more likely to be evaluated by fact checker like PolitiFact. Conceivably, similar fact checking systems could be developed for other types of public figures, or even for more low-profile professional discourse such as occurs among scientists and other technical experts. Similarly, Tsipursky and Morford [41] and Tsipursky et al. [35] describe a Pro-Truth Pledge aimed at committing people to refrain from spreading misinformation and to ask other people to retract misinformation, which can serve as a reputational punishment for misinformers, as well as a reputational benefit for those who present accurate information. Initial evaluations provide at least anecdotal support for the pledge having a positive effect on the information landscape.

For superintelligence misinformation, creating reputational costs has potential to be highly effective. A significant portion of influential voices in the debate have scholarly backgrounds and reputations that they likely wish to protect. For example, many of Steven Pinker's remarks about superintelligence are clearly misinformed, including the one discussed in Section 2 and several in his recent book *Enlightenment Now* [42]. (For detailed analysis of *Enlightenment Now*, see Torres [9].) Given Pinker's scholarly reputation, it may be productive to spread a message such as 'Steven Pinker is unenlightened about AI'.

At the same time, it is important to recognize the potential downsides of imposing reputational costs. Criticizing a person can damage one's relationship with them, reducing other sorts of opportunities. For example, criticizing people who may be building superintelligence could make them less receptive to other efforts to make their work safer. (Or, it could make them more receptive—this can be highly specific to individual personalities and contexts.) Additionally, it can impose reputational costs on the critic, such as a reputation of negativity or of seeking to restrict free speech. Caution is especially warranted for cases in which the misinformation comes from a professional contrarian, who may actually benefit from and relish in the criticism. For example, Marshall [43] (p. 72–73) warns climate scientists against debating professional climate deniers, since the latter tend to be more skilled at debate, especially televised debate, even though the arguments of the former are more sound. The same could apply for superintelligence, if it is to ever have a similar class of professional debaters. Thus, the imposition of reputation costs is a strategy to pursue selectively in certain instances of superintelligence misinformation.

4.3. Mobilize against Institutional Misinformation

The most likely institutional sources of superintelligence misinformation are the corporations involved in AI R&D, especially R&D for AGI and superintelligence. These companies have a vested interest in cultivating the impression that their technologies are safe and good for the world.

For these companies, reputational costs can also be significant. Corporate reputation can be important for consumer interest in the companies' products, citizen and government interest in imposing regulations on the companies, investor expectations of future profits, and employee interest in working for the companies. Therefore, one potential strategy is to incentivize companies so as to align their reputation with accurate information about superintelligence.

A helpful point of comparison is to corporate messaging about environmental issues, in particular the distinction between "greenwashing" and "brownwashing" [44]. Greenwashing is when a company portrays itself as protecting the environment when it is actually causing much environmental harm. For example, a fossil fuel company may publicize the greenhouse gas emissions reductions from solar panels it installs on its headquarters building while downplaying the fact that its core business model is a major driver of greenhouse gas emissions. In contrast, brownwashing is when a company declines to publicize its efforts towards environmental protection, perhaps because they have customers who oppose environmental protection or investors who worry it reduces profitability. In short, greenwashing is aimed at audiences that value environmental protection, while brownwashing is aimed at audiences that disvalue it.

Greenwashing is often criticized for giving companies a better environmental reputation than they deserve. In many cases that criticism may be fair. However, from an environmental communication standpoint, greenwashing does have the benefit of promoting a pro-environmental message. At a minimum, audiences of greenwashing are told that environmental protection is important. Audiences may also be given accurate information about environmental issues—for example, an advertisement that touts a fossil fuel company's greenhouse gas emissions reductions may also correctly explain that global warming is real and is caused by human action.

Similarly, there may be value in motivating AI companies to present accurate messages about superintelligence. This could be accomplished by cultivating demand for accurate messages among the companies' audiences. For example, if the public wants to hear accurate messages about superintelligence, then corporate advertising may be designed accordingly. The advertising might overstate the company's positive role, which would be analogous to greenwashing and could likewise be harmful for reducing accountability for bad corporate actors, but even then it would at least be spreading an accurate message about superintelligence.

Another strategy is for the employees of AI companies to mobilize against the companies supporting superintelligence misinformation, or against misinformation in general. At present, this may be a particularly promising strategy. There is a notable recent precedent for this in the successful employee action against Google's participation in Project Maven, a defense application of AI [45]. While not specifically focused on misinformation, this incident demonstrates the potential for employee action to change the practices of AI companies, including when those practices would otherwise be profitable for the company.

4.4. Focus Media Attention on Constructive Debates

Public media can inadvertently spread misinformation via the journalistic norm of balance. For the sake of objectivity, journalists often aim to cover "both sides" of an issue. While this can be constructive for some issues, it can also spread misinformation. For example, media coverage has often presented "both sides" of the "debate" over whether tobacco causes cancer or whether human activity causes global warming, even when one side is clearly correct and the other side has a clear conflict of interest [10,13].

One potential response for this is to attempt to focus media attention on legitimate open questions about a given issue, questions for which there are two meaningful sides to cover. For global warming,

this could be a debate over the appropriate role of nuclear power or the merits of carbon taxes. For superintelligence, it could be a debate over the appropriate role of government regulations, or over the values that superintelligence (or AI in general) should be designed to promote. These sorts of debates satisfy the journalistic interest in covering two sides of an issue and provide a dramatic tension that can make for a better story, all while drawing attention to important open questions and affirming basic information about the topic.

4.5. Establish Legal Requirements

Finally, there may be some potential to legally require certain actors, especially corporations, to refrain from spreading misinformation. A notable precedent is the court decision of United States v. Philip Morris, in which nine tobacco companies and two tobacco trade organizations were found guilty of conspiring to deceive the public about the link between tobacco and cancer. Such legal decisions can have powerful effect.

However, legal requirements may be poorly suited to superintelligence misinformation. First, legal requirements can be slow to develop. The court case United States v. Philip Morris began in 1999, an initial ruling was reached in 2006, and that ruling was upheld in 2009. Furthermore, United States v. Philip Morris came only after several decades of tobacco industry misinformation. Given the evolving nature of AI technology, it could be difficult to pin down which information is correct over such long time periods. Second, superintelligence is a future technology for which much of the correct information cannot be established with the same degree of rigor. Furthermore, if and when superintelligence is built, it could be so transformative as to render current legal systems irrelevant. (For more general discussion of the applicability of legal mechanisms to superintelligence, see [46–48].) For these reasons, legal requirements are less likely to play a significant role in preventing superintelligence misinformation.

5. Correcting Superintelligence Misinformation

Correcting misinformation is sufficiently difficult that it will often be better to prevent it from spreading in the first place. However, when superintelligence misinformation cannot be prevented, there are strategies available for correcting it in the minds of those who are exposed to it. Correcting misinformation is the subject of a fairly extensive literature in psychology, political science, and related fields [2,15,33,49]. For readers unfamiliar with this literature, Cook et al. [2] provide an introductory overview accessible to an interdisciplinary readership, while Ecker et al. [15] provide a more detailed and technical survey. This section applies this literature to the correction of superintelligence misinformation.

5.1. Build Expert Consensus and the Perception Thereof

At present, there exists substantial expert disagreement about a wide range of aspects of superintelligence, from basic matters such as whether superintelligence is possible [50–52] and when it might occur if it does [53–55] to subtler matters such as the treacherous turn vs. the sordid stumble. The situation stands in contrast to the extensive expert consensus on other issues such as global warming [56]. (Experts lack consensus on some important details about global warming, such as how severe the damage is likely to be, but they have a high degree of consensus on the basic contours of the issue.).

The case of global warming shows that expert consensus on its own does not counteract misinformation. On the contrary, misinformation about global warming continues to thrive despite the existence of consensus. However, there is reason to believe that the consensus helps. For starters, much of the misinformation is specifically oriented towards creating the false perception that there is no consensus [10]. The scientific consensus is a target of misinformation because it is believed to be an important factor in people's overall beliefs. Indeed, several studies have documented a strong correlation among the lay public between rejection of the science of global warming and belief that there is no consensus [57,58]. Further studies find that presenting messages describing the consensus

increases belief in climate science and support for policy to reduce greenhouse gas emissions [14,59]. Notably, this effect is observed for people across the political spectrum, including those who would have political motivation to doubt the science. (Such motivations are discussed further in Section 5.2.) All of this indicates an important role for expert consensus in broader beliefs about global warming.

For superintelligence, at present there is no need to spread misinformation about the existence of consensus because there is rather little consensus. Therefore, a first step is to work towards consensus. (This of course should be consensus grounded on the best possible analysis, not consensus for the sake of consensus.) This may be difficult for superintelligence because of the inherent challenge of understanding future technologies and the complexity of advanced AI. Global warming has its own complexities, but the core science is relatively simple: increased atmospheric greenhouse gas concentrations trap sunlight and raise temperatures. However, at least some aspects of superintelligence should be easy enough to get consensus on, starting with the fact that there are a number of R&D groups attempting to build AGI. Other aspects may be more difficult to build consensus on, but this consensus is at least something that can be pursued via normal channels of expert communication: research articles, conference symposia, private correspondence, and so on.

Given the existence of consensus, it is also important to raise awareness about it. The consensus cannot counteract misinformation if nobody knows about it. The global warming literature provides good models for documenting expert consensus [56], and such findings of consensus can be likewise be publicized.

5.2. Address Pre-Existing Motivations for Believing Misinformation

The human mind tends to not process new information in isolation, but instead processes it in relation to wider beliefs and understandings of the world. This can be very valuable, enabling us to understand the context behind new information and relate it to existing knowledge. For example, people would typically react with surprise and confusion upon seeing an object rise up to the ceiling instead of fall down to the floor. This new information is related to a wider understanding of the fact that objects fall downwards. People may even struggle to believe their own eyes unless there is a compelling explanation. (For example, perhaps the object and the ceiling are both magnetized.). Additionally, if people did not see it with their own eyes, but instead heard it reported by someone else, they may be even less likely to believe it. In other words, they are motivated to believe that the story is false, even if it is true. This phenomenon is known as *motivated reasoning*.

While generally useful, motivated reasoning can be counterproductive in the context of misinformation, prompting people to selectively believe misinformation over correct information. This occurs in particular when the misinformation accords better with preexisting beliefs than the correct information. In the above example, misinformation could be that the object fell down to the floor instead of rising to the ceiling.

Motivated reasoning is a major factor in the belief of misinformation about politically contentious issues such as climate change. The climate science consensus is rejected mainly by people who believe that government regulation of industry is generally a bad thing [14,59]. In principle, belief that humans are warming the planet should have nothing to do with belief that government regulations are harmful. It is logically coherent to believe in global warming yet argue that carbon emissions should not be regulated. However, in practice, the science of global warming often threatens people's wider beliefs about regulations, and so they find themselves motivated to reject the science.

Motivated reasoning can also be a powerful factor for beliefs about superintelligence. A basic worldview is that humans are in control. Per this worldview, human technology is a tool; the idea that it could rise up against humanity is a trope for science fiction, not something to be taken seriously. The prospect of superintelligence threatens this worldview, predisposing people to not take superintelligence seriously. In this context, it may not help that media portrayals of the scholarly debate about superintelligence commonly include reference to science fiction, such as by using pictures of the Terminator. As one expert who is concerned about superintelligence states, "I think that at this

point all of us on all sides of this issue are annoyed with the journalists who insist on putting a picture of the Terminator on every single article they publish of this topic" [60].

Motivated reasoning has been found to be linked to people's sense of self-worth. As one study puts it, "the need for self-integrity—to see oneself as good, virtuous, and efficacious—is a basic human motivation" [61] (p. 415). When correct information threatens people's self-worth, they are more motivated to instead believe misinformation, so as to preserve their self-worth. Furthermore, motivated reasoning can be reduced by having people consciously reaffirm their own self-worth, such as by recalling to themselves ways in which they successfully live up to their personal values [61]. Essentially, with their sense of self-worth firmed up, they become more receptive to information that would otherwise threaten their self-worth.

As a technology that could outperform humans, superintelligence could pose an especially pronounced threat to people's sense of self-worth. It may be difficult for people to feel good and efficacious if they would soon be superseded by computers. For at least some people, this could be a significant reason to reject information about the prospect of superintelligence, even if that information is true. At the same time, it may still be valuable for messages about superintelligence to be paired with messages of affirmation.

Another important set of motivations comes from the people active in superintelligence debates. Many people in the broader computer science field of AI have been skeptical of claims about superintelligence. These people may be motivated by a desire to protect the reputation and funding of the field of AI, and in turn protect their self-worth as AI researchers. AI has a long history of boom-bust cycles in which hype about superintelligence and related advanced AI falls flat and contributes to an "AI winter". Peter Bentley, an AI computer scientist who has spoken out against contemporary claims about superintelligence, is explicit about this:

> "Large claims lead to big publicity, which leads to big investment, and new regulations. And then the inevitable reality hits home. AI does not live up to the hype. The investment dries up. The regulation stifles innovation. And AI becomes a dirty phrase that no-one dares speak. Another AI Winter destroys progress" [62] (p. 10). "Do not be fearful of AI—marvel at the persistence and skill of those human specialists who are dedicating their lives to help create it. And appreciate that AI is helping to improve our lives every day" (p. 11).

While someone's internal motivations can only be inferred from such text, the text is at least suggestive of motivations to protect self-worth and livelihood as an AI researcher, as well as a worldview in which AI is a positive force for society.

To take another example, Torres [9] proposes that Pinker's dismissal of AGI and superintelligence is motivated by Pinker's interest in promoting a narrative in which science and technology bring progress—a narrative that could be threatened by the potential catastrophic risk from superintelligence.

Conversely, some people involved in superintelligence debates may be motivated to believe in the prospect of superintelligence. For example, researcher Jürgen Schmidhuber writes on his website that "since age 15 or so, the main goal of professor Jürgen Schmidhuber has been to build a self-improving Artificial Intelligence (AI) smarter than himself, then retire." [63] Superintelligence is also sometimes considered the "grand dream" of AI [64]. Other common motivations include a deep interest in transformative future outcomes [65] and a deep concern about extreme catastrophic risks [4,66,67]. People with these worldviews may be predisposed to believe certain types of claims about superintelligence. If it turns out that superintelligence will not be built, or would not have transformative or catastrophic effects, then this can undercut people's deeply held beliefs in the importance of superintelligence, transformative futures, and/or catastrophic risks.

For each of these motivations for interest in superintelligence, there can be information that is rejected because it cuts against the motivations and misinformation that is accepted because it supports the motivations. Therefore, in order to advance superintelligence debates, it can be valuable to affirm people's motivations when presenting conflicting information. For example, one could affirm that

AI computer scientists are making impressive and important contributions to the world, and then explain reasons why superintelligence may nonetheless be a possibility worth considering. One could affirm that science and technology are bringing a great deal of progress, and then explain reasons why some technologies could nonetheless be dangerous. One could affirm that superintelligence is indeed a worthy dream, or that transformative futures are indeed important to pay attention to, and then explain reasons why superintelligence might not be built. Finally, one could affirm that extreme catastrophic risks are indeed an important priority for human society, and then explain reasons why superintelligence may not be such a large risk after all. These affirming messaging strategies could predispose participants in superintelligence debates to consider a wider range of possibilities and make more progress on the issue, including progress towards expert consensus.

Another strategy is to align motivations with accurate beliefs about superintelligence. For example, some AI computer scientists may worry that belief in the possibility of superintelligence could damage reputation and funding. However, if belief in the possibility of superintelligence would bring reputational and funding benefits, then the same people may be more comfortable expressing such belief. Reputational benefits could be created, for example, via slots in high-profile conferences and journals, or by association with a critical mass of reputable computer scientists who also believe in the possibility of superintelligence. Funding could likewise be made available. Noting that funding and space in conferences and journals are often scarce resources, it could be advantageous to target these resources at least in part toward shifting motivations of important actors in superintelligence debates. This example of course assumes that it is correct to believe in the possibility of superintelligence. The same general strategy of aligning motivations may likewise be feasible for other beliefs about superintelligence.

The above examples—concerning the reputations and funding of AI computer scientists, the possibility of building superintelligence, and the importance of transformative futures and catastrophic risks—all involve experts or other communities that are relatively attentive to the prospect of superintelligence. Other motivations could be significant for the lay public, policy makers, and other important actors. Research on the public understanding of science finds that cultural factors, such as political ideology, can factor significantly in the interpretation of scientific information [68,69]. Kahan et al. [69] (p. 79) propose to "shield" scientific evidence and related information "from antagonistic cultural information". For superintelligence, this could mean attempting to frame superintelligence (or, more generally, AI) as a nonpartisan social issue. At least in the US, if an issue becomes politically partisan, legislation typically becomes substantially more difficult to pass. Likewise, discussions of AI and superintelligence should, where reasonably feasible, attempt to avoid close association with polarizing ideologies and cultural divisions.

The fact that early US legislation on AI has been bipartisan is encouraging. For example, H.R.4625, FUTURE of Artificial Intelligence Act of 2017, sponsored by John Delaney (Democrat) and co-sponsored by Pete Olson (Republican), and H.R.5356, National Security Commission Artificial Intelligence Act of 2018, sponsored by Elise Stefanik (Republican) and co-sponsored by James Langevin (Democrat). This is a trend that should be praised and encouraged to continue.

5.3. Inoculate with Advance Warnings

The misinformation literature has developed the concept of *inoculation*, in which people are preemptively educated about a piece of misinformation so that they will not believe it if and when they later hear it. For example, someone might be told that there is a false rumor that vaccines cause autism, such that when they later hear the rumor, they know to recognize it as false. The aim is to get people to correctly understand the truth about a piece of misinformation from the beginning, so that their minds never falsely encode it. Inoculation has been found to work better than simply telling people the correct information [70].

Inoculation messages can include why a piece of misinformation is incorrect as well as why it is being spread [71]. For example, misinformation casting doubt on the idea that global

temperatures are rising could be inoculated with an explanation of how scientists have established that global temperatures are rising. The inoculation could also explain that industries are intentionally casting doubt about global temperature increases in order to avoid regulations and increase profits. Likewise, for superintelligence, misinformation claiming that there are no projects seeking to build AGI could be inoculated by explanations of the existence of AGI R&D projects, and perhaps also explanations of the motivations of people who claim that there are no such projects. For example, Torres [9] proposes that Pinker's dismissal of AGI and superintelligence is motivated by Pinker's interest in promoting a narrative in which science and technology bring progress—a narrative that could be threatened by the potential catastrophic risk from superintelligence.

5.4. Explain Misinformation and Corrections

When people are exposed to misinformation, it can be difficult to correct, as first explained in Section 4. This phenomenon has been studied in great depth, with the terms "continued influence" and "belief perseverance" used for cases in which debunked information continues to influence people's thinking [72,73]. There is also an "illusion of truth", in which information explained to be false is later misremembered as true—essentially, the mind remembers the information but forgets its falsity [74]. The difficulty of correcting misinformation is why this paper has emphasized strategies to prevent of misinformation from spreading in the first place.

Adding to the challenge is the fact that attempts to debunk misinformation can inadvertently reinforce it. This phenomenon is known as the "backfire effect" [74]. Essentially, when someone hears "X is false", it can strengthen their mental representation of X, thereby reinforcing the misinformation. This effect has been found to be especially pronounced among the elderly [74]. One explanation is that correcting the misinformation (i.e., successfully processing "X is false") requires the use of strategic memory, but strategic memory requires dedicated mental effort and is less efficient among the elderly [15]. Unless enough strategic memory is allocated to processing "X is false", the statement can end up reinforcing belief in X.

These findings about the backfire effect have important consequences for superintelligence misinformation. Fortunately, many important audiences for superintelligence misinformation are likely to have strong strategic memories. Among the prominent actors in superintelligence debates, relatively few are elderly, and many of them have intellectual pedigrees that may endow them with strong strategic memories. On the other hand, many of the prominent actors are busy people with limited mental energy available for processing corrections about superintelligence information. As a practical matter, people attempting to debunk superintelligence misinformation should generally avoid "X is false" messages, especially when their audience may be paying limited attention.

One technique that has been particularly successful at correcting misinformation is the use of refutational text, which provides detailed explanations of why the misinformation is incorrect, what the correct information is, and why it is correct. Refutational text has been used mainly as a classroom tool for helping students overcome false preexisting beliefs about course topics [75,76]. Refutational text has even been used to turn misinformation into a valuable teaching tool [77]. A meta-analysis found refutational text to be the most effective technique for correcting misinformation in the context of science education—that is, for enabling students to overcome preexisting misconceptions about science topics [78].

A drawback of refutational text is that it can require more effort and attention than simpler techniques. Refutational text may be a valuable option in classrooms or other settings in which one has an audience's extended attention. Such settings include many venues of scholarly communication, which can be important for superintelligence debates. However, refutational texts may be less viable in other settings, such as social media and television news program interviews, in which one can often only get in a short sound bite. Therefore, refutational text may be relatively well-suited for interactions with experts and other highly engaged participants in superintelligence debates, and relatively poorly suited for much of the lay public and others who may only hear occasional passing

comments about superintelligence. That said, it may still be worth producing and disseminating extended refutations for lay public audiences, such as in long-format videos and articles for television, magazines, and online. These may tend to only reach the most motivated segments of the lay public, but they can nonetheless be worthwhile.

6. Conclusions

Superintelligence is a high-stakes potential future technology as well as a highly contested socio-technological issue. It is also fertile terrain for misinformation. Making progress on the issue requires identifying and rejecting misinformation and accepting accurate information. Some progress will require technical research to clarify the nature of superintelligence. However, a lot of progress will likely also require the sorts of sociological and psychological strategies outlined in this paper. The most progress may come from interdisciplinary projects connecting computer science, social science, and other relevant fields. Computer science is a highly technical field, but as with all fields, it is ultimately composed of human beings. By appreciating the nuances of the human dimensions of the field, it may be possible to make better progress towards understanding superintelligence and acting responsibly about it.

As the first dedicated study of strategies for countering superintelligence misinformation, this paper has taken a broad view, surveying a range of options. Despite this breadth, there may still be additional options worth further attention. Indeed, this paper has only mined a portion of the insights contained within the existing literature on misinformation. There may also be compelling options that go beyond the literature. Likewise, because of this paper's breadth, it has given relatively shallow treatment to each of the options. More detailed attention to the various option would be another worthy focus of future research.

An especially valuable focus would be the proposed strategies for preventing superintelligence misinformation. Because misinformation can be so difficult to correct, preventing it may be the more effective strategy. There is also less prior research on the prevention of misinformation. For these reasons, there is likely to be an abundance of important research opportunities on the prevention of misinformation, certainly for superintelligence misinformation and perhaps also for misinformation in general.

For the prevention of superintelligence misinformation, a strategy that may be particularly important to study further is dissuading AI corporations from using their substantial resources to spread superintelligence misinformation. The long history of corporations engaging in such tactics, with a major impact on the surrounding debates, suggests that this could be a highly important factor for superintelligence [12]. It may be especially valuable to study this at an early stage, before such tactics are adopted.

For the correction of superintelligence misinformation, a particularly promising direction is on the motivations and worldviews of prominent actors and audiences in superintelligence debates. Essentially, what are people's motivations with respect to superintelligence? Are AI experts indeed motivated to protect their field? Are superintelligence developers motivated by the "grand dream"? Are others who believe in the prospect of superintelligence motivated by beliefs about transformative futures or catastrophic risks? Can attention to these sorts of motivations help them overcome their divergent worldviews and make progress towards consensus on the topic? Finally, are people in general motivated to retain their sense of self-worth in the face of a technology that could render them inferior?

Most important, however, is not the research on superintelligence misinformation, but the efforts to prevent and correct it. It can often be stressful and thankless work, especially amidst the heated debates, but it is essential to ensuring positive outcomes. This paper is one effort towards helping this work succeed. Given the exceptionally high potential stakes, it is vital that decisions about superintelligence be well-informed.

Funding: This research received no external funding.

Acknowledgments: Olle Häggström, Tony Barrett, Brendan Nyhan, Maurizio Tinnirello, Stephan Lewandowsky, Michael Laakasuo, Phil Torres, and three anonymous reviewers provided helpful feedback on earlier versions of this paper. All remaining errors are the author's alone. The views expressed in this paper are the author's and not necessarily the views of the Global Catastrophic Risk Institute.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Definition of Misinformation in English by Oxford Dictionaries. Available online: https://en.oxforddictionaries.com/definition/misinformation (accessed on 9 September 2018).
2. Cook, J.; Ecker, U.; Lewandowsky, S. Misinformation and how to correct it. In *Emerging Trends in the Social and Behavioral Sciences*; John Wiley & Sons: Hoboken, NJ, USA, 2015; pp. 1–17.
3. Kelly, K. The Myth of a Superhuman AI. Available online: https://www.wired.com/2017/04/the-myth-of-a-superhuman-ai (accessed on 24 September 2018).
4. Häggström, O. *Here Be Dragons: Science, Technology and the Future of Humanity*; Oxford University Press: Oxford, UK, 2016.
5. Häggström, O. Michael Shermer Fails in His Attempt to Argue That AI Is Not an Existential Threat. *Häggström Hävdar*. 19 September 2017. Available online: http://haggstrom.blogspot.com/2017/09/michael-shermer-fails-in-his-attempt-to.html (accessed on 24 September 2018).
6. Häggström, O. The AI meeting in Brussels Last Week. *Häggström Hävdar*. 23 October 2017. Available online: http://haggstrom.blogspot.com/2017/10/the-ai-meeting-in-brussels-last-week.html (accessed on 9 September 2018).
7. Muehlhauser, L. *Three Misconceptions in Edge.org's Conversation on "The Myth of AI"*; Machine Intelligence Research Institute: Berkeley, CA, USA, 18 November 2014; Available online: https://intelligence.org/2014/11/18/misconceptions-edge-orgs-conversation-myth-ai (accessed on 24 September 2018).
8. Torres, P. Why Superintelligence Is a Threat That Should Be Taken Seriously. *Bulletin of the Atomic Scientists*. 24 October 2017. Available online: https://thebulletin.org/why-superintelligence-threat-should-be-taken-seriously11219 (accessed on 24 September 2018).
9. Torres, P. A Detailed Critique of One Section of Steven Pinker's Chapter "Existential Threats" in Enlightenment Now. Project for Future Human Flourishing Technical Report 2, Version 1.2. 2018. Available online: https://docs.wixstatic.com/ugd/d9aaad_8b76c6c86f314d0288161ae8a47a9821.pdf (accessed on 9 September 2018).
10. Oreskes, N.; Conway, E.M. *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming*; Bloomsbury: New York, NY, USA, 2010.
11. Grandjean, P. *Only One Chance: How Environmental Pollution Impairs Brain Development—And How to Protect the Brains of the Next Generation*; Oxford University Press: Oxford, UK, 2013.
12. Baum, S.D. Superintelligence skepticism as a political tool. *Information* **2018**, *9*, 209. [CrossRef]
13. Boykoff, M.T.; Boykoff, J.M. Balance as bias: Global warming and the US prestige press. *Glob. Environ. Chang.* **2004**, *14*, 125–136. [CrossRef]
14. Lewandowsky, S.; Gignac, G.E.; Vaughan, S. The pivotal role of perceived scientific consensus in acceptance of science. *Nat. Clim. Chang.* **2013**, *3*, 399–404. [CrossRef]
15. Ecker, U.K.H.; Swire, B.; Lewandowsky, S. Correcting misinformation—A challenge for education and cognitive science. In *Processing Inaccurate Information: Theoretical and Applied Perspectives from Cognitive Science and the Educational Sciences*; Rapp, D.N., Braasch, J.L.G., Eds.; MIT Press: Cambridge, MA, USA, 2014; pp. 13–38.
16. Baum, S.D. On the promotion of safe and socially beneficial artificial intelligence. *AI Soc.* **2017**, *32*, 543–551. [CrossRef]
17. Lewandowsky, S.; Ecker, U.K.H.; Seifert, C.M.; Schwarz, N.; Cook, J. Misinformation and its correction: Continued influence and successful debiasing. *Psychol. Sci. Public Interest* **2012**, *13*, 106–131. [CrossRef] [PubMed]

18. Pinker, S. We're Told to Fear Robots. But Why Do We Think They'll Turn on Us?". *Popular Science*. 13 February 2018. Available online: https://www.popsci.com/robot-uprising-enlightenment-now (accessed on 9 September 2018).
19. Goertzel, B. Artificial general intelligence: Concept, state of the art, and future prospects. *J. Artif. Gen. Intell.* **2014**, *5*, 1–48. [CrossRef]
20. Baum, S.D. A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy. Global Catastrophic Risk Institute Working Paper 17-1. 2017. Available online: https://ssrn.com/abstract=3070741 (accessed on 9 September 2018).
21. Cognitive Artificial Intelligence: The MicroPsi Project. Available online: http://cognitive-ai.com (accessed on 9 September 2018).
22. Searle, J.R. What Your Computer Can't Know. *New York Review of Books*, 9 October 2014.
23. Logan, R.K. Can computers become conscious, an essential condition for the Singularity? *Information* **2017**, *8*, 161. [CrossRef]
24. Chalmers, D.J. The singularity: A philosophical analysis. *J. Conscious. Stud.* **2010**, *17*, 7–65.
25. Bostrom, N. *Superintelligence: Paths, Dangers, Strategies*; Oxford University Press: Oxford, UK, 2014.
26. Omohundro, S.M. The basic AI drives. In *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*; Wang, P., Goertzel, B., Franklin, S., Eds.; IOS: Amsterdam, The Netherlands, 2008; pp. 483–492.
27. Goertzel, B. Infusing advanced AGIs with human-like value systems: Two theses. *J. Evol. Technol.* **2016**, *26*, 50–72.
28. Baum, S.D.; Barrett, A.M.; Yampolskiy, R.V. Modeling and interpreting expert disagreement about artificial superintelligence. *Informatica* **2017**, *41*, 419–428.
29. Danaher, J. Why AI doomsayers are like sceptical theists and why it matters. *Minds Mach.* **2015**, *25*, 231–246. [CrossRef]
30. Hughes, J.J. Global technology regulation and potentially apocalyptic technological threats. In *Nanoethics: The Ethical and Social Implications of Nanotechnology*; Allhof, F., Ed.; Wiley: Hoboken, NJ, USA, 2007; pp. 201–214.
31. Yampolskiy, R.; Fox, J. Safety engineering for artificial general intelligence. *Topoi* **2013**, *32*, 217–226. [CrossRef]
32. Wilkes, A.L.; Leatherbarrow, M. Editing episodic memory following the identification of error. *Q. J. Exp. Psychol.* **1988**, *40A*, 361–387. [CrossRef]
33. Johnson, H.M.; Seifert, C.M. Sources of the continued influence effect: When misinformation in memory affects later inferences. *J. Exp. Psychol. Learn. Mem. Cognit.* **1994**, *20*, 1420–1436. [CrossRef]
34. Nyhan, B.; Reifler, J. The effect of fact-checking on elites: A field experiment on U.S. state legislators. *Am. J. Political Sci.* **2015**, *59*, 628–640. [CrossRef]
35. Tsipursky, G.; Votta, F.; Roose, K.M. Fighting fake news and post-truth politics with behavioral science: The pro-truth pledge. *Behav. Soc. Issues* **2018**, *27*, 47–70. [CrossRef]
36. Doran, P.T.; Zimmerman, M.K. Examining the scientific consensus on climate change. *Eos* **2009**, *90*, 22–23. [CrossRef]
37. Stenhouse, N.; Maibach, E.; Cobb, S.; Ban, R.; Bleistein, A.; Croft, P.; Bierly, E.; Seitter, K.; Rasmussen, G.; Leiserowitz, A. Meteorologists' views about global warming: A survey of American Meteorological Society professional members. *Bull. Am. Meteorol. Soc.* **2014**, *95*, 1029–1040. [CrossRef]
38. De La Harpe, J. TV Meteorologists, Weathercasters Briefed by Climate Experts at AMS Short Course. *Yale Climate Connnections*. 9 July 2009. Available online: https://www.yaleclimateconnections.org/2009/07/tv-meteorologists-weathercasters-briefedby-climate-experts-at-ams-short-course (accessed on 9 September 2018).
39. Ward, B. 15 Midwest TV Meteorologists, Weathercasters Weigh Climate Science at Chicago's Field Museum Climate Science for Meteorologists. *Yale Climate Connnections*. 5 May 2009. Available online: https://www.yaleclimateconnections.org/2009/05/meteorologists-weathercasters-weigh-climate-science-chicago (accessed on 9 September 2018).
40. Nyhan, B. Why the 'death panel' myth wouldn't die: Misinformation in the health care reform debate. *The Forum* **2010**, *8*. [CrossRef]
41. Tsipursky, G.; Morford, Z. Addressing behaviors that lead to sharing fake news. *Behav. Soc. Issues* **2018**, *27*, AA6–AA10.

42. Pinker, S. *Enlightenment Now: The Case for Reason, Science, Humanism, and Progress*; Penguin: New York, NY, USA, 2018.
43. Marshall, G. *Don't Even Think about It: Why Our Brains Are Wired to Ignore Climate Change*; Bloomsbury: New York, NY, USA, 2014.
44. Kim, E.-H.; Lyon, T.P. Greenwash vs. brownwash: Exaggeration and undue modesty in corporate sustainability disclosure. *Organ. Sci.* **2014**, *26*, 705–723. [CrossRef]
45. BBC. Google 'to end' Pentagon Artificial Intelligence Project. *BBC*. 2 June 2018. Available online: https://www.bbc.com/news/business-44341490 (accessed on 9 September 2018).
46. McGinnis, J.O. Accelerating AI. *Northwest. Univ. Law Rev.* **2010**, *104*, 366–381.
47. Wilson, G. Minimizing global catastrophic and existential risks from emerging technologies through international law. *Va. Environ. Law J.* **2013**, *31*, 307–364.
48. White, T.N.; Baum, S.D. Liability law for present and future robotics technology. In *Robot Ethics 2.0*; Lin, P., Abney, K., Jenkins, R., Eds.; Oxford University Press: Oxford, UK, 2017; pp. 66–79.
49. Ecker, U.K.H.; Lewandowsky, S.; Tang, D.T.W. Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Mem. Cognit.* **2010**, *38*, 1087–1100. [CrossRef] [PubMed]
50. Bringsjord, S. Belief in the singularity is logically brittle. *J. Conscious. Stud.* **2012**, *19*, 14–20.
51. McDermott, D. Response to the singularity by David Chalmers. *J. Conscious. Stud.* **2012**, *19*, 167–172.
52. Chalmers, D. The Singularity: A reply. *J. Conscious. Stud.* **2012**, *19*, 141–167.
53. Baum, S.D.; Goertzel, B.; Goertzel, T.G. How long until human-level AI? Results from an expert assessment. *Technol. Forecast. Soc. Chang.* **2011**, *78*, 185–195. [CrossRef]
54. Müller, V.C.; Bostrom, N. Future progress in artificial intelligence: A survey of expert opinion. In *Fundamental Issues of Artificial Intelligence*; Müller, V.C., Ed.; Springer: Cham, Switzerland, 2016; pp. 555–572.
55. Grace, K.; Salvatier, J.; Dafoe, A.; Zhang, B.; Evans, O. When will AI exceed human performance? Evidence from AI experts. *J. Artif. Intell. Res.* **2018**, *62*, 729–754. [CrossRef]
56. Oreskes, N. The scientific consensus on climate change. *Science* **2004**, *306*, 1686. [CrossRef] [PubMed]
57. Ding, D.; Maibach, E.W.; Zhao, X.; Roser-Renouf, C.; Leiserowitz, A. Support for climate policy and societal action are linked to perceptions about scientific agreement. *Nat. Clim. Chang.* **2011**, *1*, 462–466. [CrossRef]
58. McCright, A.M.; Dunlap, R.E.; Xiao, C. Perceived scientific agreement and support for government action on climate change in the USA. *Clim. Chang.* **2013**, *119*, 511–518. [CrossRef]
59. Van der Linden, S.L.; Leiserowitz, A.A.; Feinberg, G.D.; Maibach, E.W. The scientific consensus on climate change as a gateway belief: Experimental evidence. *PLoS ONE* **2015**, *10*, e0118489. [CrossRef] [PubMed]
60. Bensinger, R. Sam Harris and Eliezer Yudkowsky on 'AI: Racing toward the Brink'. *Machine Intelligence Research Institute*. 28 February 2018. Available online: https://intelligence.org/2018/02/28/sam-harris-and-eliezer-yudkowsky (accessed on 9 September 2018).
61. Cohen, G.L.; Sherman, D.K.; Bastardi, A.; Hsu, L.; McGoey, M.; Ross, L. Bridging the partisan divide: Self-affirmation reduces ideological closed-mindedness and inflexibility in negotiation. *J. Pers. Soc. Psychol.* **2007**, *93*, 415–430. [CrossRef] [PubMed]
62. Bentley, P.J. The three laws of artificial intelligence: Dispelling common myths. In *Should We Fear Artificial Intelligence? In-Depth Analysis*; Boucher, P., Ed.; European Parliamentary Research Service, Strategic Foresight Unit: Brussels, Belgium, 2018; pp. 6–12.
63. Jürgen Schmidhuber's Home Page. Available online: http://people.idsia.ch/~juergen (accessed on 9 September 2018).
64. Legg, S. Machine Super Intelligence. Doctoral's Thesis, University of Lugano, Lugano, Switzerland, 2008.
65. More, M.; Vita-More, N. (Eds.) *The Transhumanist Reader: Classical and Contemporary Essays on the Science, Technology, and Philosophy of the Human Future*; Wiley: New York, NY, USA, 2010.
66. Bostrom, N. Existential risk prevention as global priority. *Glob. Policy* **2013**, *4*, 15–31. [CrossRef]
67. Torres, P. *Morality, Foresight & Human Flourishing an Introduction to Existential Risks*; Pitchstone Publishing: Durham, NC, USA, 2017.
68. Kahan, D.M.; Jenkins-Smith, H.; Braman, D. Cultural cognition of scientific consensus. *J. Risk Res.* **2011**, *14*, 147–174. [CrossRef]
69. Kahan, D.M.; Peters, E.; Dawson, E.C.; Slovic, P. Motivated numeracy and enlightened self-government. *Behav. Public Policy* **2017**, *1*, 54–86. [CrossRef]

70. Banas, J.A.; Rains, S.A. A meta-analysis of research on inoculation theory. *Commun. Monogr.* **2010**, *77*, 281–311. [CrossRef]
71. Cook, J.; Lewandowsky, S.; Ecker, U.K.H. Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PLoS ONE* **2017**, *12*, e0175799. [CrossRef] [PubMed]
72. Cobb, M.D.; Nyhan, B.; Reifler, J. Beliefs don't always persevere: How political figures are punished when positive information about them is discredited. *Political Psychol.* **2013**, *34*, 307–326. [CrossRef]
73. Nyhan, B.; Reifler, J. Displacing misinformation about events: An experimental test of causal corrections. *J. Exp. Political Sci.* **2015**, *2*, 81–93. [CrossRef]
74. Skurnik, I.; Yoon, C.; Park, D.C.; Schwarz, N. How warnings about false claims become recommendations. *J. Consum. Res.* **2005**, *31*, 713–724. [CrossRef]
75. Kowalski, P.; Taylor, A.K. The effect of refuting misconceptions in the introductory psychology class. *Teach. Psychol.* **2009**, *36*, 153–159. [CrossRef]
76. Kuhn, D.; Crowell, A. Dialogic argumentation as a vehicle for developing young adolescents' thinking. *Psychol. Sci.* **2011**, *22*, 545–552. [CrossRef] [PubMed]
77. Bedford, D. Agnotology as a teaching tool: Learning climate science by studying misinformation. *J. Geogr.* **2010**, *109*, 159–165. [CrossRef]
78. Guzzetti, B.J.; Snyder, T.E.; Glass, G.V.; Gamas, W.S. Promoting conceptual change in science: A comparative meta-analysis of instructional interventions from reading education and science education. *Read. Res. Q.* **1993**, *28*, 117–159. [CrossRef]