
The Role of Cooperation in Responsible AI Development

Amanda Askell*

OpenAI

amanda@openai.com

Miles Brundage

OpenAI

miles@openai.com

Gillian Hadfield

OpenAI

gillian@openai.com

Abstract

In this paper, we argue that competitive pressures could incentivize AI companies to underinvest in ensuring their systems are safe, secure, and have a positive social impact. Ensuring that AI systems are developed responsibly may therefore require preventing and solving collective action problems between companies. We note that there are several key factors that improve the prospects for cooperation in collective action problems. We use this to identify strategies to improve the prospects for industry cooperation on the responsible development of AI.

Introduction

Machine learning (ML) is used to develop increasingly capable systems targeted at tasks like voice recognition, fraud detection, and the automation of vehicles. These systems are sometimes referred to as narrow artificial intelligence (AI) systems. Some companies are also using machine learning techniques to try to develop more general systems that can learn effectively across a variety of domains rather than in a single target domain. Although there is a great deal of uncertainty about the development path of future AI systems—whether they will remain specialized or grow increasingly general, for example—many agree that if the current rate of progress in these domains continues then it is likely that advanced artificial intelligence systems will have an increasingly large impact on society.

This paper focuses on the private development of AI systems that could have significant expected social or economic impact, and the incentives AI companies have to develop these systems responsibly. Responsible development involves ensuring that AI systems are safe, secure, and socially beneficial.

In most industries, private companies have incentives to invest in developing their products responsibly. These include market incentives, liability laws, and regulation. We argue that AI companies have the same incentives to develop AI systems responsibly, although they appear to be weaker than they are in other industries. Competition between AI companies could decrease the incentives of each company to develop responsibly by increasing their incentives to develop faster. As a result, if AI companies would prefer to develop AI systems with risk levels that are closer to what is socially optimal—as we believe many do—responsible AI development can be seen as a collective action problem.¹

We identify five key factors that make it more likely that companies will be able to overcome this collective action problem and cooperate—develop AI responsibly with the understanding that others will do likewise. These factors are: high trust between developers (*High Trust*), high shared gains from mutual cooperation (*Shared Upside*), limited exposure to potential losses in the event of unreciprocated cooperation (*Low Exposure*), limited gains from not reciprocating the cooperation of others (*Low Advantage*), and high shared losses from mutual defection (*Shared Downside*).

*Primary/corresponding author.

¹AI research companies increasingly have teams dedicated to the safe and ethical development of technology and many large technology companies participate in voluntary efforts to articulate and establish principles and guidelines, and in some cases call for government regulation, to address AI-related risks.

Using these five factors, we identify four strategies that AI companies and other relevant parties could use to increase the prospects for cooperation around responsible AI development. These include correcting harmful misconceptions about AI development, collaborating on shared research and engineering challenges, opening up more aspects of AI development to appropriate oversight, and incentivizing greater adherence to ethical and safety standards. This list is not intended to be exhaustive, but to show that it is possible to take useful steps towards more responsible AI development.

The paper is composed of three sections. In section 1, we outline responsible AI development and its associated costs and benefits. In section 2, we show that competitive pressures can generate incentives for AI companies to invest less in responsible development than they would in the absence of competition, and outline the five factors that can help solve such collective action problems. In section 3, we outline the strategies that can help companies realize the gains from cooperation. We close with some questions for further research.

1 The benefits and costs of responsible AI development

AI systems have the ability to harm or create value for the companies that develop them, the people that use them, and members of the public who are affected by their use. In order to have high expected value for users and society, AI systems must be safe—they must reliably work as intended—and secure—they must have limited potential for misuse or subversion. AI systems should also not introduce what Zwetsloot and Dafoe (2019) call “structural risks”, which involve shaping the broader environment in subtle but harmful ways.² The greater the harm that can result from safety failures, misuse, or structural risks, the more important it is that the system is safe and beneficial in a wide range of possible conditions (Dunn, 2003). This requires the responsible development of AI.

1.1 What is responsible AI development?

AI systems are increasingly used to accomplish a wide range of tasks, some of which are critical to users’ health and wellbeing. As the range of such tasks grows, the potential for accidents and misuse also grows, raising serious safety and security concerns (Amodei, Olah, et al., 2016; Brundage, Avin, et al., 2018). Harmful scenarios associated with insufficiently cautious AI development have already surfaced with, for example, biases learned from large datasets distorting decisions in credit markets and the criminal justice system, facial recognition technologies disrupting established expectations of privacy and autonomy, and auto-pilot functions in some automobiles causing new types of driving risk (while reducing others). Longer term, larger scale scenarios include dangers such as inadvertent escalation of military conflict involving autonomous weapon systems or widespread job displacement.

Responsible AI development involves taking steps to ensure that AI systems have an acceptably low risk of harming their users or society and, ideally, to increase their likelihood of being socially beneficial. This involves testing the safety and security of systems during development, evaluating the potential social impact of the systems prior to release, being willing to abandon research projects that fail to meet a high bar of safety, and being willing to delay the release of a system until it has been established that it does not pose a risk to consumers or the public. Responsible AI development comes in degrees but it will be useful to treat it as a binary concept for the purposes of this paper. We will say that an AI system has been developed responsibly if the risks of it causing harms are at levels most people would consider tolerable, taking into account their severity, and that the amount of evidence grounding these risk estimates would also be considered acceptable.³

Responsible AI development involves work on safety, security, and the structural risks associated with AI systems. Work on the safety of AI aims to mitigate accident risks (Amodei, Olah, et al., 2016) and ensure that AI systems function as intended (Ortega, Maini, et al., 2018) and behave in ways that people want (Irving et al., 2018). Work on the security of AI aims to prevent AI systems from being attacked,

²Zwetsloot and Dafoe (2019) argue that the “the accident-misuse dichotomy obscures how technologies, including AI, often create risk by shaping the environment and incentives”. We restrict accident risks to technical accidents and misuse risks to direct misapplications of a system. ‘Structural risks’, as we use the term, are intended to capture the broader impact of AI systems on society and social institutions.

³This will generally mean that if an AI system is developed responsibly, the risk of irreversible catastrophic harm from that system—whether through accident, misuse, or negative social impact—must be very low. This is consistent with what Sunstein (2005) calls the ‘Irreversible Harm Precautionary Principle’.

co-opted, or misused by bad actors (Brundage, Avin, et al., 2018).⁴ Work evaluating the structural impact of AI aims to identify and mitigate both the immediate and long term structural risks that AI systems pose to society: risks that don't quite fit under narrow definitions of accident and misuse. These include joblessness, military conflict, and threats to political and social institutions.⁵

1.2 The cost of responsible AI development

It is likely that responsible development will come at some cost to companies, and this cost may not be recouped in the long-term via increased sales or the avoidance of litigation. In order to build AI systems responsibly, companies will likely need to invest resources into data collection and curation, system testing, research into the possible social impacts of their system, and, in some cases, technical research to guarantee that the system is reliably safe. In general, the safer that a company wants a product to be, the more constraints there are on the kind of product the company can build and the more resources it will need to invest in research and testing during and after its development.

If the additional resources invested in ensuring that an AI system is safe and beneficial could have been put towards developing an AI system with fewer constraints more quickly, we should expect responsible AI development to require more time and money than incautious AI development. This means that responsible development is particularly costly to companies if the value of being the first to develop and deploy a given type of AI system is high (even if the first system developed and deployed is not demonstrably safe and beneficial).

There are generally several advantages that are conferred on the first company to develop a given technology (Lieberman and Montgomery, 1988). If innovations can be patented or kept secret, the company can gain a larger share of the market by continuing to produce a superior product and by creating switching costs for users. Being a first-mover also allows the company to acquire scarce resources ahead of competitors. If hardware, data, or research talent become scarce, for example, then gaining access to them early confers an advantage.⁶ And if late movers are not able to catch up quickly then first-mover advantages will be greater. In the context of AI development, having a lead in the development of a certain class of AI systems could confer a first mover advantage. This effect would be especially pronounced in the case of discontinuous changes in AI capabilities⁷, but such a discontinuity is not necessary in order for a first mover advantage to occur.

Responsible development may therefore be costly both in terms of immediate resources required, and in the potential loss of a first-mover advantage. Other potential costs of responsible AI development include performance costs and a loss of revenue from not building certain lucrative AI systems on the grounds of safety, security, or impact evaluation. An example of a performance cost is imposing a limit on the speed that self-driving vehicles can travel in order to make them safer. An example of revenue loss is refusing to build a certain kind of facial recognition system because it may undermine basic civil liberties (Smith, 2018b).

AI companies may not strongly value being the first to develop a particular AI system because first-mover advantages do not always exist. Indeed, there are often advantages to entering a market after the front-runner. These include being able to free-ride on the R&D of the front-runner, to act on more information about the relevant market, to act under more regulatory certainty, and having more flexible assets and structures that let a company respond more effectively to changes in the environment (Gilbert and Birnbaum-More, 1996). These can outweigh the advantages of being the first to enter that same market. And it has been argued that late mover advantages often do outweigh first mover advantages (Markides and Geroski, 2004; Querbes and Frenken, 2017). We therefore acknowledge that the assumption that there will be a first-mover advantage in AI development may not be true. If a first-mover advantage in AI is weak or non-existent then companies are less likely to engage in a race to the bottom on safety since speed is of lower value. Instead of offering predictions, this paper should be thought of as an analysis of more pessimistic scenarios that involve at least a moderate first mover advantage.

⁴Mitigating misuse risks is sometimes included under AI safety, broadly construed (Christiano, 2016)

⁵The literature on the societal impact of AI is vast. See Cummings (2017) on AI and warfare, for example. For a broader overview see Dafoe (2018).

⁶As Klepper (1996) notes, larger first-movers can also spread their prior investment in R&D over a larger number of applications.

⁷The possibility of discontinuous progress in AI is discussed by Good (1966), Chalmers (2009), Yudkowsky (2013), Shanahan (2015) and Bostrom (2017b). AI Impacts (2018) provide a critical overview of the arguments for discontinuity and Christiano (2018) presents arguments against the claim.

Much of the discussion of AI development races assumes that they have a definitive endpoint. Although some have hypothesized that if AI progress is discontinuous or sufficiently rapid then it could essentially have a definitive endpoint, the case for this remains speculative.⁸ It is therefore important to note that AI development may take the form of a perpetual R&D race: a race to stay technologically ahead of competitors rather than a race to reach some particular technological endpoint (Aoki, 1991; Breitmoser et al., 2010). If this is the case then AI companies would still have an incentive to speed up development in order to stay ahead of others, especially if the gap between companies was small. The present analysis is applicable to perpetual races in which there is at least a moderate first mover advantage, several companies are competing to stay ahead, and leadership is not yet entrenched.⁹

1.3 The benefits of responsible AI development

In the law and economics literature on product safety, it is generally accepted that market forces create incentives for companies to invest in making their products safe (Oi et al., 1973). Suppose that companies have accurate information about how safe the products they are developing are and that consumers have access to accurate information about how safe a company's product is, either prior to release or by observing the harms caused by a product after it is released (Ben-Shahar, 1998; Chen and Hua, 2017).¹⁰ If consumers have a preference for safer products and respond rationally to this preference, they will not buy products that are insufficiently safe, or will pay less for them than for safer alternatives (Polinsky and Shavell, 2009). Releasing unsafe products will also result in a costly loss of reputation for companies (Daughety and Reinganum, 1995).¹¹ Finally, releasing unsafe products could result in burdensome regulation of the industry or in litigation costs. Therefore companies that are concerned about a sufficiently long time-horizon involving repeated interaction with customers, regulators, and other stakeholders that incentivize safety should internalize the value of responsible development.

Market forces alone may not always incentivize companies to invest the appropriate amount into ensuring their products are safe. If consumers cannot get access to information about the safety of a product—how likely safety failures are or how costly they are—then companies have an incentive to under-invest in safety. And if companies have inaccurate information about the safety of the products they are developing, they will not invest in safety to the degree demanded by consumers. Finally, poor corporate governance can result in suboptimal decisions about risk (Cai et al., 2010). Product liability law and safety regulation are intended to correct such market failures by providing consumers with information about products, incentivizing companies to invest more in safety, and compensating consumers that are harmed by product safety failures (Hylton, 2012; Landes and Posner, 1985).¹²

We may expect companies to under-invest in safety if the costs to consumers don't result in commensurate costs for the company; either via a reduction in revenue, reputation loss, fines from regulators, or successful litigation by consumers. Safety failures can also affect those who do not consume the product, however. Consider a 2018 recall of over 8,000 Volkswagen vehicles potentially affected by a brake caliper issue that could result in increased stopping distances or loss of vehicle control (Consumer Reports, 2018). A safety failure resulting from this could harm not only the vehicle's occupants but also pedestrians and other drivers.¹³ Harms that safety failures inflict on non-consumers are negative externalities, and benefits that safer products produce for non-consumers are positive externalities. We should anticipate companies under-investing in reducing negative externalities and increasing

⁸If AI development is extremely rapid then gaps between each company would likely increase over time. This means that a company that is ahead of others may at some point be ahead of them by a great deal in strategically important areas, and could use this to undermine their competitors (Bostrom, 2017b, pp. 91-104)

⁹Breitmoser et al. (2010) note that perpetual R&D races tend to collapse into leadership monopolies. The larger the gap between the front-runner and the company in second place in a perpetual race, the less of an incentive the front-runner has to trade safety for speed.

¹⁰See Daughety et al. (2013), who make similar assumptions in their idealized model of markets.

¹¹Rhee and Haunschmid (2006) provide evidence that the relationship between safety failures and reputation loss may be more complex than this, however.

¹²See Stiglitz (2009) on government regulation as a response to market failures or inefficiencies, but note that actual motivations for government regulation are typically more complicated (Henson and Caswell, 1999). Calabresi (1970) provides comprehensive overview of the role of law in the minimization of costs from safety failures. The relationship between product liability law and safety regulations—in particular, whether it is efficient to use them jointly—is a matter of some debate (Shavell, 1984; Kolstad et al., 1990).

¹³There is currently uncertainty about who should be held liable for the harms that the safety failures of autonomous systems inflict on the public (Schellekens, 2015; The Atlantic, 2018).

positive externalities relative to their social value, since the costs and benefits this produces for society don't result in commensurate costs and benefits for the company (Dahlman, 1979).

To give a concrete example, consider facial recognition technology. Microsoft have argued that this technology could be used in ways that many would consider harmful: to violate individuals' privacy or suppress their political speech, for example (Smith, 2018a,b). Even if companies would prefer to build facial recognition systems that cannot be misused, either to avoid causing harm or to avoid the reputation costs of this harm, the cost of developing safeguards may not outweigh their benefits if companies cannot be held liable for these harms and there is no regulation preventing misuse. For this reason, Microsoft has called for regulation that would require that companies invest in measures that reduce the risks from facial recognition technology, and that could also mitigate potential misuse of the technology by commercial entities or by governments (Smith, 2018a).

The discussion thus far treats companies as though they were motivated only by profit, i.e. they only care about things like reputation and product safety insofar as they are a means to make more profit or avoid losses. This view is common in the literature on corporate social responsibility (Campbell, 2007; Devinney, 2009) but it is clearly an abstraction. Companies are run by, invested in, and composed of humans that care about the impact their products will have on the world and on other people. Employees at technology companies have already shown that they care a great deal about the social implications of the systems they are building (Minsberg, 2019).

The things that motivate AI companies other than profits, such as benefiting people rather than harming them, will generally push even more in favor of responsible development: they will rarely push against it. Assuming that companies are motivated solely by profit therefore lets us analyze a kind of 'worst case scenario' for responsible development. We will therefore often treat companies as though they were driven solely by profit, even though we do not find this plausible. It is important that the reader bear this in mind, since treating companies as profit-driven entities can be self-fulfilling, and can therefore contribute to the very problems we are attempting to solve.

1.4 Are existing incentives for responsible AI development enough?

If markets are functioning well and companies and consumers have perfect information about the expected harm of a product, companies should invest the socially optimal amount into product safety (Daughety et al., 2018). In real-world scenarios in which markets may not function perfectly and information asymmetries exist, incentives for companies to invest sufficiently in product safety typically come from three sources: market forces, liability law, and industry or government regulation.¹⁴ These three sources of incentives may not provide strong enough incentives for AI companies to engage in responsible AI development, however. We will briefly survey some reasons for this.

1.4.1 Limited consumer information

Consumers of AI systems include individuals, private companies, and public institutions. Although different consumers will have access to different levels of information about AI systems, information about the expected harm of AI systems is likely to be quite limited on average. As cutting-edge AI systems become more complex, it will be difficult for consumers not involved in the development of those systems to get accurate information about how safe the systems are. Consumers cannot directly evaluate the safety of aviation software, for example, and will face similar difficulties when it comes to directly evaluating the safety of complex machine learning models. This is compounded by the fact that it is notoriously difficult to explain the decisions made by neural networks (Doshi-Velez and Kim, 2017; Olah et al., 2018). If consumers cannot assess how risky a given AI system is, they cannot adjust their willingness to pay for it accordingly. They are also less able to identify and exert pressure on AI companies that are investing too little in safety (Anton et al., 2004).

Consumers could get information about how safe an AI system is by tracking safety failures after its release, but such a 'wait and see' strategy could leave both consumers and the public vulnerable to harmful safety failures. This is of particular concern if those safety failures could be irreversible or catastrophic. And the probability of irreversible or catastrophic safety failures is likely to increase as AI

¹⁴ Other mechanisms include no fault liability systems like mandatory insurance and increasing the information available to consumers (Cornell et al., 1976).

systems become more capable and general, since more advanced systems are more likely to be relied upon across a wider range of domains and in domains where failures are more harmful.¹⁵

1.4.2 Limited company and regulator information

Measuring the safety, security, and social impact of AI systems may turn out to be extremely difficult even for those who understand the technical details of the system. Neural networks are difficult to interpret and as such, failures may be difficult to predict. If AI companies are over-confident that their system is not risky, they may under-invest in important risk-reducing measures during development or release a system that causes unintended harm.

If regulators cannot assess how risky a given AI system is, they may be overly stringent or overly liberal when using regulatory controls (Shavell, 1984). The ability to get accurate information about AI systems therefore seems to be crucial for ex ante safety measures.¹⁶

Our current capacities to identify and measure the expected harms of particular AI systems are extremely limited. We still do not fully understand the decisions made by complex machine learning models (Olah et al., 2018; Hohman et al., 2018) and the high-dimensionality of the inputs to AI systems makes it such that exhaustive enumeration of all possible inputs and outputs is typically infeasible. There may therefore be little consensus about whether a particular system is likely to be unsafe, unsecure, or socially harmful at present. Given this, it is likely that additional capacity will need to be invested by companies or regulators or both in order to decrease these information asymmetries.

1.4.3 Negative externalities from AI

Harms caused by AI systems are likely to affect third parties. Biases in algorithmic pre-trial risk assessment are more likely to harm those accused of crimes than those that purchase the tools,¹⁷ those that benefit from AI automation may be quite distinct from the people who are displaced by automation,¹⁸ and a major AI disaster—such as an AI system with a faulty reward function¹⁹ being integrated into a critical system—could affect a large portion of society that is distinct from the AI company and its consumers. AI also has the potential to be a general purpose technology—a technology that radically affects many sectors of the economy—and if this is the case we should expect its impact to be systemic (Brynjolfsson et al., 2018; Cockburn et al., 2018).

The harms from AI systems may also be difficult to internalize. For example, the social harms that result from an increased use of AI systems—such as reduced trust in online sources—could be complex and diffuse, and it may be difficult to hold any one company strictly liable for them. If the harm is sufficiently large, it may also be too large for a company or insurer to cover all losses (see note 15). Finally, AI systems could create negative externalities for future generations that are not in a position to penalize companies or prevent them from occurring (Lazear, 1983). We should expect AI companies to under-invest in measures that could prevent these kinds of negative externalities.²⁰

1.4.4 The difficulty of constructing effective AI regulation

There is currently little in the way of AI-targeted regulation, including government regulation, industry self-regulation, international standards, and clarity on how existing laws will be applied to AI (see note

¹⁵Such safety failures could occur if AI systems have some critical function like controlling national power grids or nuclear weapons systems, or if they can be used to undermine these systems. The more consequential a given technology is, the higher the potential cost of releasing an insufficiently safe version of that technology to both the company and to society. But it is worth noting that if the expected cost of catastrophic safety failures is capped by a company's ability to pay then we might expect companies to under-weight these tail-end risks. For a taxonomy of AI risks, see Yampolskiy (2015).

¹⁶Leike et al. (2017) introduce simple environments for evaluating the safety of AI agents, and note that future versions of these environments could be used to benchmark the safety performance of AI agents.

¹⁷See Tsukayama and Williams (2018) on how bias in ML systems could harm those in the California criminal justice system. Pleiss et al. (2017) demonstrates the unique difficulties of designing bias-free ML systems.

¹⁸Segal (2018) notes that the jobs that have declined as a result of automation so far are intermediate-skill jobs like farming. Although it displaces some workers, automation has positive effects like increased productivity (Acemoglu and Restrepo, 2018). The overall effect that AI automation will have on the labor force is unclear.

¹⁹See Krakovna (2018) and Clark and Amodei (2016) for examples of faulty reward functions in ML systems.

²⁰Safety failures that affect a large portion of the population will not be treated as externalities by companies if they harm the company or its consumers, though they could still be given insufficient weight.

13). Well-designed regulatory mechanisms can incentivize companies to invest appropriate resources in safety, security, and impact evaluation when market failures or coordination failures have weakened the other incentives to do so. Poorly-designed regulation can be harmful rather than helpful, however. Such regulation can discourage innovation (Heyes, 2009) and even increase risks to the public (Latin, 1988).

AI regulation seems particularly tricky to get right, as it would require a detailed understanding of the technology on the part of regulators.²¹ The fact that private AI companies can generally relocate easily also means that any attempt to regulate AI nationally could result in international regulatory competition rather than an increase in responsible development.²² Regulation that is reactive and slow may also be insufficient to deal with the challenges raised by AI systems. AI systems can operate much faster than humans, which can lead to what Johnson et al. (2013) call ‘ultrafast extreme events’ (UEEs) such as flash crashes caused by algorithmic trading.²³

1.4.5 The potential for rapid AI development

Some have hypothesized that progress in AI development will be discontinuous (see note 7). On this view, there are some types of AI systems—typically advanced ‘general’ AI systems that are capable of learning effectively across a wide variety of domains—that, if developed, would represent a sudden shift from everything that came before them, and could produce the equivalent of many years of prior progress on some relevant metric.²⁴ If AI progress is discontinuous then developing an AI system that constitutes a sudden leap forward could give a company a large advantage over others, since the next best system would be years behind it in terms of prior progress in the field.²⁵ Consider the advantage that a company today would gain if they managed to develop something over a decade ahead of current systems used for cyber offense and defense, for example.

If progress in AI development is discontinuous then market forces and liability law may do little to encourage safe development.²⁶ The value of developing a system that gives a company a huge advantage—that could be used to undermine competition or seize resources, for example—would be largely divorced from the process of getting market feedback. And a company can only be held liable for accidents if these accidents are not catastrophic and the existing legal framework can both keep up with the rapidity of technological progress and enforce judgments against companies. Therefore if AI progress is discontinuous, ex ante safety measures like industry self-regulation or international oversight may be more effective than ex post safety measures like market response and liability.

1.5 Summary

Incentives to develop safe products generally come from the market, liability laws, and regulation (Rubin, 2011), as well as factors that motivate AI companies beside profits, such as a general desire to avoid doing harm. For AI companies, the profit motive to develop AI responsibly is likely to come from the additional revenue generated by AI systems that are more valuable to consumers, the avoidance

²¹Hadfield (2017) and Hadfield and Clark (2019) outline a regulatory framework for AI that could overcome barriers like information asymmetries and slow response times: key problems for the regulation of new technology.

²²Esty and Geradin (2001) offer an overview of different perspectives on regulatory competition, while Genschel and Plumper (1997) note that regulatory competition and international co-operation can actually increase levels of regulation. Erdélyi and Goldsmith (2018) argue that an international AI regulatory agency should be established, but on the grounds that AI has externalities that cross national boundaries.

²³For more on this problem, see Muehlhauser and Hibbard (2014). Anticipating, preventing, and responding to catastrophic AI-caused UEEs may present a key challenge in AI safety and policy.

²⁴See Ehrnberg (1995) on definitions of technological discontinuities. The AI discontinuity hypothesis should not be confused with the claim that there will be rapid AI development in the future—progress in AI development could be continuous but extremely rapid, e.g. hyperbolic (Christiano, 2017)—but that there will be a system that represents a sudden leap forward in AI capabilities. It may be possible to achieve a decisive advantage over competitors if progress is rapid but not discontinuous.

²⁵Bostrom (2017b) claims that such an AI could give a company a ‘decisive strategic advantage’: ‘a level of technological and other advantages sufficient to enable it to achieve complete world domination’ (p.96, *ibid.*). But the concerns we raise here apply even if the advantage is extreme but not decisive in this sense.

²⁶This may be true even if progress in AI development is continuous but rapid. Even if no single company has a profound advantage over others, mechanisms like regulation and liability could be too slow to catch up with the rate of technological progress. It is worth noting that if AI progress takes this shape then responsible AI development may be more like a one-shot game than an iterated game, which reduces developers’ incentives to cooperate on responsible development for reasons that we discuss in the next section.

of reputational harm from safety failures, the avoidance of widespread harms caused by AI systems (see note 20), and the avoidance of tort litigation or regulatory penalties.

A key factor that can influence the cost-benefit ratio of responsible AI development that we have not discussed, however, is the competitive environment in which the AI systems in question are being developed. In the next section we will explore the impact that competition between AI companies can have on the incentives that each company has to invest or fail to invest in responsible development.

2 The need for collective action on responsible AI development

We have argued that safer, more secure, and more socially valuable AI systems will tend to have a higher market value, be less likely to cause costly accidents that the company is held liable for, and so on. This means that if a company is guaranteed to be the first to develop a system of this type, we can expect that they will invest resources to ensure that their system is safe, secure, and socially beneficial to the extent that this is incentivized by regulators, liability law, and market forces. This means the more that positive and negative externalities of AI systems have been internalized via these mechanisms, the more that companies can expect to invest in responsible development.²⁷

In this section we will argue that, even with these incentives in place, competitive pressures can cause AI companies to invest less in responsible development than they otherwise would. Responsible AI development can therefore take the form of a collective action problem. We then identify and discuss five key factors that improve the prospects for cooperation between AI companies that could find themselves in a collective action problem over responsible development.

2.1 How competitive pressures can lead to collective action problems

To see how the competitive environment could affect investment in responsible development, suppose that several AI companies are working on a similar type of system. If there is a large degree of substitutability between the inputs of different aspects of development, we should not expect AI companies to invest in responsible development beyond the point at which the expected marginal return is lower than the expected marginal return from investing in other areas of development. Suppose each company places less value on coming second than on coming first, less value in coming third than in coming second, and so on. These companies will likely engage in a technological race: a competition to develop a technology in which the largest reward goes to the first company (Grossman and Shapiro, 1985).²⁸ The resulting dynamics may be similar to those we would expect to see in patent races between firms.²⁹

There are various strategies companies could use in a “winner takes more” race: they could try to develop and maintain a strong technical lead or they could try to maintain a close position behind the technical leader, for example.³⁰ For now, we will assume that the best strategy involves trying to develop and maintain a strong technical lead throughout the race.

Since speed is more valuable when racing against others, we should expect investment into responsible development to be lower when companies are racing against each other.³¹ Armstrong et al. (2016) point out that in an AI development race, responsible development could be prey to a “race to the bottom” dynamic. Consider what happens if one company decides to increase their development speed by decreasing their investment in safety, security, and impact evaluation. This increases their expected ranking in the race and decreases the expected ranking of others in the race. A decrease in expected

²⁷If the first company could prevent future competitors from entering the market (i.e. the first company could expect to be the only company), it is likely this would reduce but not eliminate market incentives to invest in responsible development (Sheshinski, 1976).

²⁸As we noted in the previous section, this is a non-trivial assumption that will not hold in all cases.

²⁹Patent races have positive effects on innovation, though at the cost of duplicating efforts (Judd et al., 2012).

³⁰The best strategy may depend on the competitive environment. Dasgupta and Stiglitz (1980) argue that monopolist companies will attempt to outspend their rivals on R&D to prevent a duopoly, while Doraszelski (2003) shows that there are conditions in which companies that are behind will invest to catch up.

³¹How much lower will depend on various features of the race, such as how close it is and the value placed on each position. Note that this argument assumes that investments with even worse expected marginal returns have already been cut. It also assumes that investments in responsible development contribute less to development speed than other available investments: not that they contribute nothing to development speed.

ranking gives competing AI companies an incentive to decrease their own investment in these areas in order to maintain or increase their expected ranking in the race.³²

We might ask why racing to the bottom on product safety is not ubiquitous in other industries in which decreasing time-to-market is valuable, such as in the pharmaceutical industry.³³ The most plausible explanation of this difference is that the cost of safety failures has been internalized to a greater extent in more established industries via external regulation, self-regulation, liability, and market forces. These mechanisms can jointly raise the “bottom” on product safety to a level that is generally considered acceptable by regulators and consumers.³⁴

In a race to the bottom on safety, competing AI companies could reduce their investment in responsible development to the point that winning the technology race—successfully developing the system they are racing to develop before others—is barely of net positive value for the winner even after all the first-mover advantages, including positive reputational effects, the ability to capture resources like data, hardware and talent, and creating switching costs for consumers, have been taken into account.³⁵

2.2 When competition has negative rather than positive effects

The race to the bottom on safety described above is a collective action problem: a situation in which all agents would be better off if they could all cooperate with one another, but each agent believes it is in their interest to defect rather than cooperate.³⁶ As Heckathorn (1989, p. 78) states, “the inclinations of individuals (that is, each actor’s preferences regarding his or her own behavior) are in conflict with regulatory interests (that is, each actor’s preferences regarding the behavior of others). The collective action problem arises when a group possesses a common interest, or faces a common fate.”

In a race to the bottom on safety, it is in each company’s interest to reduce their investment in responsible development in order to increase development speed. If all companies do this, however, there is a single equilibrium: one in which much or all of the value that could have been gained with coordination is destroyed. If each company defects, they will have a similar position in the race to the one that they would have had if they had all successfully coordinated, but they will be developing systems that are more risky than the ones they would have developed if they had all managed to successfully coordinate. In other words, the situation in which they find themselves is strictly worse than the situation in which coordination was successful.

Collective action problems between companies can have positive effects on consumers and the public. A price war is a collective action problem between companies with mostly positive effect on consumers, for example, as it results in lower prices.³⁷ Antitrust law exists to maintain competition between companies that has a positive effect on consumers and to prevent collusion between companies that has a negative effect on consumers (e.g. price fixing).

³²In this scenario, companies have full information about the investments made by other companies and their likelihood of winning. But this assumption is not necessary, since companies can invest in accordance with their expectation about the investments and win probabilities of other companies. Armstrong et al. (2016) explore scenarios in which AI companies have different levels of information about their own and others’ capabilities.

³³It is worth noting that similar concerns about the desire to develop quickly conflicting with risk management have been expressed in other industries that involve novel technology, such as the use of nanoparticles and nanotechnology in the food industry (Morgan, 2005; Cushen et al., 2012).

³⁴It could also be the that the best strategies in technological races do not involve trying to develop a strong technological lead, or that there are unidentified factors that make racing to the bottom on product safety undesirable: factors that may apply equally to the development of AI systems.

³⁵The company with the winning system could even consider their own system to be worse than developing nothing at all absent competition, though this would only happen if they considered the release of the alternative winning system to be even worse for them than the release of their own worse-than-nothing system.

³⁶This is a weakening of the definition that Jon Elster derives from Schelling (2006 [1978]), which states ‘First, each individual derives greater benefits under conditions of universal cooperation than he does under conditions of universal noncooperation. Second, each derives more benefits if he abstains from cooperation, regardless of what others do.’ (Elster, 1985, p.139). We simply replace ‘regardless of what others do’ with ‘given what we expect others will do.’ See Holzinger (2003) for a broader definition and taxonomy.

³⁷When companies engage in price wars, prices often end up close to their marginal cost of production (Bresnahan, 1987). Prices can even be temporarily set below the marginal cost of production in order to push competitors out of the market (Guiltinan and Gundlach, 1996), sometimes in violation of antitrust.

When there are negative effects from production that are not captured by the incentives facing producers (i.e. negative externalities), however, competition does not lead to the socially optimal outcome. If this outcome is also bad for the producers, it is a collective action problem for producers.

A race to the bottom on safety falls into this category if it results in AI systems with safety levels below what is socially optimal and below what AI companies would prefer. Pollution by companies is another example of a collective action problem between companies that has a negative effect on the public (Lévéque, 1999).

Before discussing strategies for cooperation such as self-regulation in more depth, however, it will be useful to understand the incentives that AI companies have to abide by norms that involve mutual investment in responsible AI development. This will be the focus of the remainder of this section.

2.3 Incentives to cooperate in collective action problems

In an AI development race, companies “cooperate” if they maintain some acceptable level of investment in responsible development and they “defect” if they fail to maintain this level of investment, thereby acting in their own interest (hypothetically) and against the collective interest. Encouraging companies to cooperate should therefore not be confused with encouraging them to stop competing. Companies agreeing not to compete across the investment in safety dimension does not imply that they will cease to compete across the R&D dimension. Competitive dynamics that contain cooperative elements are sometimes referred to as a “coopetition”.³⁸

If companies have incentives to prevent or mitigate collective action problems that have negative effects on consumers or the public then we should expect the companies themselves (and not just third parties like government regulators) to take steps to solve them. And companies often do attempt to cooperate to prevent or solve collective action problems of this sort. One example of a mechanism used to this end is industry self-regulation. (Gunningham and Rees, 1997).³⁹ Examples of self-regulation include Responsible Care: a self-regulation program in the US chemicals industry (Gamper-Rabindran and Finger, 2013), and the Institute of Nuclear Power Operations (INPO): an industry organization that conducts inspections and facilitates the sharing of best practices in the nuclear power industry (Davis and Wolfram, 2012; Hausman, 2014).⁴⁰

In order to identify features that affect the degree to which it is in a company’s interest to cooperate on responsible development, it will be helpful to highlight features that increase incentives to cooperate in collective action problems generally. To do this, consider the payoff matrix of a cooperate-defect game in which two agents (AI companies) can cooperate (develop responsibly) or defect (fail to develop responsibly). Here the first letter in each pair represent the expected payoff for Agent 1, and the second letter in each pair represents the payoff for Agent 2.⁴¹

		Agent 2	
		Cooperate	Defect
Agent 1	Cooperate	a_1, a_2	b_1, b_2
	Defect	c_1, c_2	d_1, d_2

Table 1: A Normal Form Cooperate-Defect Game

³⁸See Bengtsson and Kock (2000) and Tsai (2002).

³⁹Industry self-regulation can also be incentivized by government regulators via meta-regulation (Parker, 2007; Coglianese and Mendelson, 2010).

⁴⁰Other examples of self-regulation can be found in a variety of industries, as self-regulation is sometimes used to preempt government regulation (Lenox, 2007). How successful such self-regulation is at reducing the negative effects of collective action problems varies a great deal by industry. The INPO is generally considered to be a more successful self-regulatory scheme than Responsible Care, for example (Cohen and Sundararajan, 2015, pp. 126-7). This may be because the INPO, unlike Responsible Care, has an agreement with a government regulator, the Nuclear Regulatory Commission, which can monitor the program and provide meaningful sanctions, which may be required for successful self-regulation (King and Lenox, 2000). O’Keefe (forthcoming 2019) explores one possible form of antitrust-compliant self-regulation in the AI industry.

⁴¹We assume that these expected utilities have already factored in agents’ attitudes towards risk and discuss some of the simplifications of this framework below.

Let p be the probability that Agent 1 assigns to Agent 2 cooperating and let q be the probability that Agent 2 assigns to Agent 1 cooperating. We assume it is rational for Agent 1 to cooperate if the expected value of cooperation (the likelihood Agent 2 will cooperate times a_1 plus the likelihood Agent 2 will defect times b_1) is greater than the expected value of defection (the likelihood Agent 2 will cooperate times c_1 plus the likelihood Agent 2 will defect times d_1). We assume the same is true of Agent 2.⁴² This lets us identify five highly interrelated factors that increase an agent's incentive to cooperate. These factors are as follows, where expected values are relative to the agent's beliefs.⁴³

- (1) *High Trust*: being more confident that others will cooperate (p, q)⁴⁴
- (2) *Shared Upside*: assigning a higher expected value to mutual cooperation (a_1, a_2)
- (3) *Low Exposure*: assigning a lower expected cost to unreciprocated cooperation (b_1, c_2)
- (4) *Low Advantage*: assigning a lower expected value to not reciprocating cooperation (c_1, b_2)
- (5) *Shared Downside*: assigning a lower expected value to mutual defection (d_1, d_2)

The last four factors each refer to the expected value of an action conditional on the behavior of the other agent, such as cooperating and having your cooperation reciprocated. Note that the expected value of an action depends on how good the agent perceives the outcome to be and how likely the agent perceives it to be. This means that an agent could be in a 'low exposure' scenario if she considers unreciprocated cooperation to be not very valuable or not very likely or both. We can provide agents with evidence about the likelihood and value of each outcome by changing the world in some perceptible way, e.g. by offering a reward for responsible development, or by giving them evidence about the way the world already is, e.g. by correcting false beliefs.

It is useful to separate the degree of trust (factor 1) from incentives (factors 2-5) in order to discuss its role in cooperation, but trust is not independent of incentives or vice versa. If one agent comes to trust an agent more, this increases the expected value of the outcomes that involve cooperation.⁴⁵ The same is true in reverse: if the expected value of the outcomes that involve cooperation increase, it is more likely that the other agent will cooperate. In other words, increasing trust can increase incentives to cooperate, and increasing incentives to cooperate can increase trust between agents.⁴⁶

This means that if a company can provide information about itself that increases the probability the other assigns to it cooperating, this will increase the degrees of trust between the companies and make it more likely each company's trust threshold will be met. Two important facts follow from this. First, information that companies provide about their intentions and actions—how transparent they are—can play an important role in whether other companies will cooperate with them. Second, trust is prey to virtuous and vicious cycles. If one company demonstrably increases its trust in another, the other company should increase its trust in return. But if one company demonstrably decreases its trust in another, the other company should decrease its trust in return.⁴⁷

A real world race to the bottom on safety would unfold over many interactions. The factors identified here also increase the prospect of cooperation in sequential games, however.⁴⁸ And iterated collective

⁴²In other words, it is rational for Agent 1 to cooperate if $p \times a_1 + (1 - p) \times b_1 > p \times c_1 + (1 - p) \times d_1$ and it is rational for Agent 2 to cooperate if $q \times a_2 + (1 - q) \times c_2 > q \times b_2 + (1 - q) \times d_2$. These two agents are in a collective action problem if it is irrational for both agents to cooperate, but $a_1 > d_1$ and $a_2 > d_2$. If both sides of these equations are equal then defecting and cooperating are both rationally permissible for the agent. Note that if $a_1 + a_2 > d_1 + d_2$ but $a_1 \not> d_1$ or $a_2 \not> d_2$ (i.e. defecting is rational for at least one agent but mutual cooperation creates more total value for both agents than mutual defection does) then the likelihood of cooperation increases if redistribution is possible, i.e. if the agents can bargain towards a solution.

⁴³If Agent 1 and Agent 2 are not in an anti-coordination game then Agent 1's incentives to cooperate increase as (1) p increases, (2) the expected value of a_1 increases, (3) the expected value of c_1 increases, (4) the expected value of b_1 decreases, and (5) the expected value of d_1 decreases. Naturally, the inverse of each of these factors will decrease the agent's incentive to cooperate.

⁴⁴This is not the only definition of 'trust', but it is the one that is most relevant to the current analysis.

⁴⁵We say 'in the situations we consider here' because if the agents are in an anti-coordination game then increasing Agent 1's trust in Agent 2 will decrease Agent 1's incentives to cooperate.

⁴⁶Again, this will not be true in certain anti-coordination games.

⁴⁷This is one reason why a degree of 'forgiveness' can be strategically valuable: it can prevent errors, misinterpretations, or aberrant behavior from plunging both players into a vicious cycle of distrust prematurely, and can pull players out of such a cycle (Axelrod, 1980)

⁴⁸The main adjustment we need to make to the factors above in extensive form games will be to the first factor: high trust. If we let C_i mean that agent i cooperates and assume that agents can only either cooperate or defect, in

action problems are generally easier to solve than one-shot collective action problems because, in iterated collective action problems, players have an incentive (and opportunity) to cooperate early in the game in order to establish trust and avoid retaliation.⁴⁹ Using one-shot games to illustrate our points is therefore more likely to skew us towards undue pessimism about our ability to solve races to the bottom rather than undue optimism.

One shortcoming of our analysis, however, is that it appeals to an overly simplified conception of cooperation and defection. For example, we assume that the options available to agents can be divided into ‘cooperation’ and ‘defection’. In reality, cooperation will come in varying degrees—companies can invest different amounts in responsible development, for example—and it would be better to talk about the degree of cooperation that we can expect between agents.⁵⁰ We also assume that companies will make an intentional decision to coooperare or defect over time. In reality, companies could fail to foresee the consequences of investing very little into areas like safety, and may therefore defect without intending to. Third, we assume that both companies perfectly understand the actions and assertions of the other. In reality, it may not be clear whether a company is living up to an agreement to develop AI responsibly. If agreements are not clear then there may not be a bright line between defection and non-defection that companies can respond to (Chassang, 2010; Gibbons and Henderson, 2012). A more complete analysis of collective action problems in AI development should build a more realistic model of what cooperating and defecting during AI development would look like.

We have argued that in order to “solve” a collective action problem, we can try to transform it into a situation in which mutual cooperation is rational. If we can transform it into a situation in which agents have lower minimum trust thresholds (generally determined by the payoff matrix) and greater trust of each other—greater confidence that if they cooperate, others will reciprocate (Kydd, 2007, p. 9)—then we should expect a higher degree of mutual cooperation.⁵¹ Given this, we should expect “lower conflict” collective action problems—problems in which agents have stronger incentives to cooperate—to be easier to solve than “higher conflict” collective action problems—problems in which agents have weaker incentives to cooperate.⁵²

2.4 The cooperative factors in AI development

Whether an AI development race will result in a collective action problem and, if so, how bad it will be are both open questions.⁵³ But there are many features of an AI development race that affect both the likelihood and severity of collective action problems. For example, having close frontrunners would likely worsen a collective action problem—would reduce the tractability of resolving it—because this increases the expected value frontrunners will assign to not reciprocating the cooperation of others (low advantage) and therefore increases the probability they assign to not having their own cooperation

extensive form games our ‘high trust’ factor would say that the incentives for Agent 1 to cooperate with Agent 2 increase as $p(C_2|C_1)$ and $p(\neg C_2|\neg C_1)$ increase. The other four factors can remain largely unchanged.

⁴⁹In an iterated Prisoner’s Dilemma, for example, cooperation can be incentivized by things like the threat of retaliation and the promise of reciprocity (Axelrod, 1984; Nowak, 2006). The promise of reciprocity increases the expected value of mutual cooperation today (shared upside) and the threat of retaliation decreases the expected value of betraying the cooperation of others today (low advantage). The payoff structure of the iterated Prisoner’s Dilemma may therefore be more like that of a Stag Hunt (Seabright, 1993, p.123). See Mailath et al. (1991) on the extent to which important features of extensive form games can be preserved in normal form.

⁵⁰We are attempting to illustrate the general structure of reasons to cooperate in AI development rather than analyzing a particular case in detail.

⁵¹Sometimes collective action problems are the result of one or more agents having mistaken beliefs about the expected value of cooperating and defecting. When this is the source of the problem, it can be ‘solved’ by correcting these misconceptions.

⁵²Scenarios in which agents have stronger incentives to cooperate with one another involve less ‘conflict’ (Robinson and Goforth, 2005; Schelling, 1980). How easy it is to solve collective action problems depends both on the degree of conflict involved and the nature and magnitude of the resources we have at our disposal. For example, the Stag Hunt is easier to solve than the Prisoner’s Dilemma. All possible adjustments to the Prisoner’s Dilemma that result in a solution will, if applied to the Stag Hunt, result in a solution to this problem as well. But only some adjustments to payoffs and probabilities that solve the former would also solve the latter.

⁵³It is also worth bearing in mind that scenarios can be superficially similar to collective action problems even though it is in everyone’s interest to cooperate.

reciprocated (high trust).⁵⁴ Similarly, a misaligned perception of the risks associated with different AI systems could worsen a collective action problem if it causes less cautious companies to assign a lower cost to not reciprocating cooperation (low advantage), which could increase the probability that cautious companies assign to having their cooperation unreciprocated by less cautious companies (high trust) and increase the expected harm that cautious companies expect to arise from incautious companies getting ahead this way (low exposure).

Features that affect the likelihood and severity of a collective action problem for responsible development can be used to decrease its likelihood and severity if they are features that we can control. For example, fundamental distrust between companies is likely to worsen a collective action problem because companies are less likely to expect that their cooperation will be reciprocated (high trust). Building trust between AI companies can therefore decrease the severity of collective action problems. An AI race development in which the expected value of winning is much greater than the expected value of losing is also likely to have a worse collective action problem (low exposure and low advantage).⁵⁵ If close frontrunners worsen collective action problems, AI companies may agree to take steps to avoid engaging in a harmful race to the bottom on safety. For example, citing concerns about race dynamics, OpenAI (2018) have stated that “if a value-aligned, safety-conscious project comes close to building AGI before we do, we commit to stop competing with and start assisting this project.”

The mechanisms to incentivize investment in product safety outlined in the previous section—market forces, regulation, and liability—all operate to prevent collective action problems for product safety. Consumers often pay less for products that are unsafe (low advantage and shared downside) and more for safe products (shared upside and low exposure). Government regulation either removes the option to underinvest in safety or increases the cost of underinvesting in safety via sanctions and fines (low advantage and shared downside). And the possibility of being held liable for harms caused by unsafe products decreases the expected value of underinvesting in safety to get ahead (low advantage and shared downside).

Market forces, regulation, and liability are all mechanisms operating outside of the AI industry that affect the incentives that AI companies have to develop responsibly. But if responsible AI development is a collective action problem then each AI company expects to benefit from being in a better equilibrium and therefore has an incentive to ensure that the AI industry itself collectively coordinates to maintain some acceptable level of responsible development. Companies should be willing to invest in cooperative mechanisms to the degree that these mechanisms increase the likelihood that they will be able to capture the cooperation surplus: the additional expected value that cooperation would generate for them.⁵⁶

This means that industry-led mechanisms like greater self-regulation could also be developed to incentivize responsible AI development.⁵⁷

2.5 Summary

In this section we argued that responsible AI development may take the form of a collective action problem. We also identified five factors that generally increase the likelihood of mutual cooperation and can help solve such collective action problems. In the next section we will translate this into more concrete suggestions for increasing cooperation on safety between AI companies.

⁵⁴This is consistent with the conclusion of Armstrong et al. (2016) that frontrunners will take more risks if they have a close competitor. The claim that competition is more intense among close competitors has also been made in the literature on R&D races also (Grossman and Shapiro, 1985; Harris and Vickers, 1987).

⁵⁵We have focused on cases that involve cooperation, but we can use the more cooperation-neutral factors in note 43 to look at the expected cost to one company if another company wins regardless of the degree of cooperation between the two companies. To give another example of this, Armstrong et al. (2016) discuss the level of enmity between companies. Higher enmity would then be expected to worsen a collective action problem by increasing the cost of losing the race to the other company (low exposure).

⁵⁶This concept is similar to the Harsanyi dividend, which quantifies the value created by a coalition (Harsanyi, 1963). The value of trust as a commodity is explored by Dasgupta (2000). Companies should be willing to pay more for credible demonstrations of their intention to cooperate.

⁵⁷King and Lenox (2000) highlights the difficulties of self-regulation by looking at the chemical industry’s Responsible Care program. A self-regulatory program that is considered more successful, however, is the Institute of Nuclear Power Operations (INPO). See Coglianese and Mendelson (2010) for an analysis of both.

3 Strategies to improve AI industry cooperation on safety

In the previous section, we argued five factors make it more likely that AI companies will cooperate if they are faced with a collective action problem: (1) being more confident that others will cooperate, (2) assigning a higher expected value to mutual cooperation, (3) assigning a lower expected cost to unreciprocated cooperation, (4) assigning a lower expected value to not reciprocating cooperation, (5) assigning a lower expected value to mutual defection.

These five factors give high-level direction regarding how to ensure that the fruits of cooperation in AI are realized. However, it is not always obvious what these five factors mean in the real world, so there is a need for translating these factors into tangible policy strategies that various actors can implement in order to improve cooperation prospects.

It is impossible to prescribe such strategies fully in advance, because we lack information about the future which would be needed in order to make informed future decisions, and because a particular policy proposal could be effective if well-implemented but counterproductive if poorly executed. However, while detailed, long-term policy prescriptions would be premature today, there are several coarse-grained strategies that seem robustly desirable even if some of the low-level details require research, dialogue, and passage of time before they can be clarified.

We believe that the four strategies we identify in this section are robustly desirable in the sense that they all have substantial benefits with respect to at least one of the factors above, and are unlikely to be very harmful with respect to the others.

3.1 Promote accurate beliefs about the opportunities for cooperation

As noted in prior sections, there are multiple competing conceptions of AI development. In cases where people are demonstrably uninformed about key aspects of AI development, it is likely beneficial to correct them, and more generally for stakeholders to make nuanced public statements consistent with the spirit of AI development that involves cooperation on norms of responsible development.

Some misconceptions that should be corrected in order to improve prospects for such cooperation include incorrect beliefs that safety and security risks can be safely ignored (Brundage, Avin, et al., 2018; Amodei, Olah, et al., 2016; Ortega, Maini, et al., 2018), an unwarranted focus on relative gains and losses instead of absolute gains and losses (shared upside, low exposure, low advantage, shared downside), and mistaken belief in interests being more misaligned than they are (low exposure and low advantage). In addition to correcting specific misconceptions, there is also likely value in proactively informing people about the case for cooperating on responsible development generally.

For example, recent years have seen substantial effort by researchers and activists to highlight the biases being learned by deployed AI systems in critical societal domains such as criminal justice and in widely used technological platforms such as recommender systems. This work has highlighted the risks of incautious development to a large and growing swathe of the AI community. Similarly, concerns have been raised about both the bias, efficacy, and other properties of medical AI systems, as well as self-driving vehicles and other emerging technologies. Analyzing and communicating these sorts of risks is critical for generating interest in cooperation among a sufficiently wide range of actors, as well as in identifying appropriate norms around research, publication, and deployment given the safety risks and the ways of mitigating them that have been identified.

In many cases, common knowledge that multiple parties share a concern or interest can be critical for the initiation of cooperation, and a misconception that parties lack such a shared concern or interest could be damaging to cooperation on issues like safety. Avoiding such misunderstanding may be particularly important in the case of international cooperation on responsible AI development across distinct countries with different languages and cultural frames of reference.

Propagating accurate information about existing beliefs can also be valuable, as it allows multiple parties to stabilize their expectations. For example, the Asilomar AI Principles (Future of Life Institute, 2017) commit the many signatories to arms race avoidance, and various statements of principles before and after this have similarly committed many actors to various (admittedly still abstract) cooperative statements and actions. Expanding the breadth and depth of such dialogue, especially across cultural and language boundaries, will be critical in fostering understanding of the large gains from mutual

responsible development (shared upside) and the large losses from mutual irresponsible development (shared downside), and in establishing common knowledge that such understanding exists (high trust).

It is possible to create positive spirals of trust, in which an increase in one party's trust causes the trusted party to increase their trust in turn. We can also stumble into negative trust spirals, however, in which a loss of trust leads to further distrust between parties. It is therefore also important to avoid feeding into unnecessarily adversarial rhetoric about AI development, lest it become self-fulfilling (Kreps, 2019).

3.2 Collaborate on shared research and engineering challenges

On a range of possible research challenges—from basic AI research to applied AI projects to AI safety and security research—it can be beneficial for multiple parties to actively pool resources and ideas, provided this can be done in a way that is procompetitive and compliant with antitrust laws (FTC/DoJ, 2000), does not raise security concerns for the companies participating, and so on.

Joint research can provide value for cooperation via useful technical insights (such as solutions to safety problems; low exposure and low advantage), stabilizing expectations regarding who is working on what via public information about joint investments as well as interpersonal dialogue (versus work being more shrouded in secrecy; high trust and shared downside), concretizing the joint upsides of AI (e.g. AI for good collaborations; shared upside), and facilitating more societally beneficial publication and deployment decisions by various actors (e.g. via collaborative analysis of the risks of specific systems; shared upside).⁵⁸ Note that we refer specifically here to active and explicit research collaboration, of which some already occurs, alongside a much greater amount of implicit collaboration on AI research that already exists due to the high degree of openness in the AI research community.

Active and explicit research collaboration in AI, especially across institutional and national borders, is currently fairly limited in quantity, scale, and scope. This is for a range of reasons. In order to maintain legitimate academic and industrial competition, researchers or their managers may be averse to publishing certain research outputs early or at all. And research ideas, results, datasets, and code can be hard to disentangle from proprietary product plans and technical infrastructure. Furthermore, safety or security considerations can in some cases make the joint analysis of a particular system more challenging than it would otherwise be (Radford, Wu, et al., 2019). There are also linguistic and logistical barriers to collaborating across long distances and across different cultures and languages.

While we acknowledge that such challenges exist, we advocate a more thorough mapping of possible collaborations across organizational and national borders, with particular attention to research and engineering challenges whose solutions might be of wide utility. Areas to consider might include joint research into the formal verification of AI systems' capabilities and other aspects of AI safety and security with wide application; various applied "AI for good" projects whose results might have wide-ranging and largely positive applications (e.g. in domains like sustainability and health); coordinating on the use of particular benchmarks; joint creation and sharing of datasets that aid in safety research; and joint development of countermeasures against global AI-related threats such as the misuse of synthetic media generation online.

3.3 Open up more aspects of AI development to appropriate oversight and feedback

Openness about one's beliefs, actions, and plans is critical to establishing trust generally. In the case of AI development, those building and deploying AI systems need to provide information about their development process so that users can make informed decisions. Likewise, governments need to be able to appropriately oversee safety-critical AI systems, and (in the absence of relevant regulation) companies need to be able to provide information to one another that shows they are following appropriate norms.

The general appeal of openness for cooperation-related reasons does not imply that all aspects of AI development should always be open, and as AI systems become more capable, it will be increasingly

⁵⁸For example, OpenAI's approach to the release of the GPT-2 language model family (Radford, Wu, et al., 2019) involves staged release, in which a model is released incrementally due to safety and security concerns, and partnership-based sharing, in which a model is shared with a small number of research partners to enable research on that system without necessarily requiring broad-based access. This experiment in responsible publication, and others like it such as the Allen Institute for Artificial Intelligence and the University of Washington's approach on their Grover family of language models, may help to "derisk" this particular form of research-level collaboration discussed in the next subsection.

important to decide responsibly what should and shouldn't be made open (Brundage, Avin, et al., 2018; Bostrom, 2017a; Krakovna, 2016). Full transparency is problematic as an ideal to strive for, in that it is neither necessary nor sufficient for achieving accountability in all cases (Desai and Kroll, 2017; Ananny and Crawford, 2018). Further, some information about AI development cannot or should not be shared for reasons of safety, security, ethics, or law. For example, AI developers might legitimately be wary of releasing code that is intimately tied to proprietary infrastructure, and should certainly be wary of releasing private data as well as AI systems that are easily amenable to misuse.

Given that full openness is rarely called for, but that some openness is required for building trust, there is a need for continuing effort to implement existing modes of trust-building in AI, as well as to discover new ones. Different mechanisms for achieving openness regarding how AI systems are developed and operated include, e.g., publicizing decision-making principles and processes, explaining publication/release decisions, sharing accessible information about how particular AI systems and broad classes of AI systems work, allowing external visitors to the lab, and opening up individual AI systems to detailed scrutiny (e.g. via bug bounties or open sourcing).

Such openness is critical in allowing reputation to play its stabilizing role in cooperation. Indeed, some actors have explicitly pointed to the challenges of monitoring the development and use of lethal autonomous weapons as a reason not to agree to strict rules, suggesting that the inability to track others' behavior reliably could be a bottleneck on some forms of mutually beneficial cooperation (e.g. joint restraints on weapons development). In cases such as this, a richer set of tools for opening up actors to critical scrutiny and feedback (while managing the associated risks) would be useful, and we encourage continued exploration of approaches such as those mentioned above as well as others in order to widen the range of cooperative actions available to AI developers.

In combination, the appropriate application of transparency mechanisms such as these should reduce the severity of concerns about others behaving irresponsibly (low exposure), reduce the temptation to defect in partially competitive situations (low advantage), and increase confidence that others' statements about their behavior are accurate (high trust). Openness is a particularly powerful strategy, and applicable to a wider range of cooperation problems, if it can be gradually ratcheted up in an iterative fashion between parties, as opposed to happening all at once. This gradual approach can reduce the temptation to defect at any particular stage (low advantage) and increase confidence in others cooperating (shared downside).

3.4 Incentivize adherence to high standards of safety

Cooperative actors might want to introduce additional incentives (reward and/or punishment) related to responsible AI development beyond those that exist today, or would exist by default in the future. E.g. such actors might strongly value compliance with certain norms intrinsically, and prefer that those who comply with appropriate norms be rewarded; or one might want to deliberately bring about an incentive for oneself to act in a certain way, as a commitment mechanism; one might also want to use incentives as a complement to other governance tools such as monitoring of behavior and direct regulation; and one might want to generally influence the incentives of many actors in a particular direction, and support policies that bring this about.

There are several categories of incentives that one might want to consider in this context. Creating incentives for key actors to act cooperatively, if done effectively, would help with all five factors simultaneously. Potential incentives include:

- Social incentives (e.g. valorizing or criticizing certain behaviors related to AI development) can influence different companies' perceptions of risks and opportunities
- Economic incentives (induced by governments, philanthropists, industry, or consumer behavior) can increase the share of high-value AI systems in particular markets or more generally, and increase attention to particular norms⁵⁹
- Legal incentives (i.e. proscribing certain forms of AI development with financial or greater penalties) could sharply reduce temptation by some actors to defect in certain ways.
- Domain-specific incentives of particular relevance to AI (e.g. early access to the latest generation of computing power) could be used to encourage certain forms of behavior.

⁵⁹Mutual agreements to distribute the economic gains from winning the AI development (O'Keefe et al., forthcoming 2019) could also decrease the severity of collective action problems.

As argued in each case above, these strategies are robustly desirable from the perspective of enabling cooperation, but our articulation of them leaves many questions unanswered. In particular, sharpening these recommendations and adapting them over time will require technical and social scientific research, creative institutional design, and bold policy experimentation, e.g. via regulatory markets as discussed in Hadfield and Clark (2019).

4 Conclusion and Future Directions

In this paper we have argued that competition between AI companies could create a collective action problem for responsible AI development. We have identified five key factors that make it more likely that companies will cooperate on responsible development: high trust, shared upside, low exposure, low advantage, and shared downside. We have shown that these five factors can help us to identify strategies to help AI companies develop responsibly and thereby realize the gains from cooperation. This also has important positive externalities for consumers and the general public.

If our analysis is on the right track then it is best thought of as the beginning of a program of research, rather than the last word on the subject. Much work needs to be done to identify whether collective action problems for responsible AI development will occur if we vary who is developing AI, how many entities are developing AI, what systems they are developing, and so on. More work must also be done to identify and evaluate strategies that can prevent or mitigate these kinds of collective action problems across a wide range of possible scenarios.

The possible future research directions on this issue are broad and we do not aim to provide a comprehensive list of them here, but examples of potentially fruitful research questions include:

1. How might the competitive dynamics of industry development of AI differ from government-led or government-supported AI development?
2. What is the proper role of legal institutions, governments, and standardization bodies in resolving collective action problems between companies, particularly if those collective action problems can arise between companies internationally?
3. What further strategies can be discovered or constructed to help prevent collective action problems for responsible AI development from forming, and to help solve such problems if they do arise? What lessons can we draw from history or from contemporary industries?
4. How might competitive dynamics be affected by particular technical developments, or expectations of such developments?

As we noted at the outset, there is substantial uncertainty about the nature and pace of developments in AI. If the impact of AI systems on society is likely to increase, however, then greater attention must be paid to ensuring that the systems being developed and released are safe, secure, and socially beneficial. In this paper we argued that existing incentives to develop AI responsibly may be weaker than is ideal, and that this may be compounded by competitive pressure between companies, leading to a collective action problem on the responsible development of AI.

That such collective action problems will arise or that they will be maintained if they do arise is far from a foregone conclusion, however. Finding ways of preventing and solving these problems may require new ways of building trust in novel technological contexts, and in some cases to assume some risk in the expectation that others will reciprocate in turn. While intellectually and politically challenging, we think such efforts are integral to realizing the positive-sum potential of AI.

Acknowledgments

We are grateful to Michael Page, Jack Clark, Larissa Schiavo, Carl Shulman, Luke Muehlhauser, Geoffrey Irving, Sarah Kreps, Paul Scharre, Michael Horowitz, Robert Trager, Tamay Besiroglu, Helen Toner, Cullen O’Keefe, Rebecca Crootof, Ben Garfinkel, Adam Gleave, Jasmine Wang, and Toby Shevlane for valuable feedback on earlier versions of this paper.

References

- Daron Acemoglu and Pascual Restrepo. Artificial intelligence, automation and work. Technical report, National Bureau of Economic Research, 2018.
- AI Impacts. Likelihood of discontinuous progress around the development of agi. <https://aiimpacts.org/likelihood-of-discontinuous-progress-around-the-development-of-agi/>, Feb 2018. (Accessed on 03/12/2019).
- Dario Amodei, Chris Olah, et al. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.
- Mike Ananny and Kate Crawford. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3):973–989, 2018.
- Wilma Rose Q Anton, George Deltas, and Madhu Khanna. Incentives for environmental self-regulation and implications for environmental performance. *Journal of environmental economics and management*, 48(1):632–654, 2004.
- Reiko Aoki. R&d competition for product innovation: An endless race. *The American Economic Review*, 81(2):252–256, 1991.
- Stuart Armstrong, Nick Bostrom, and Carl Shulman. Racing to the precipice: a model of artificial intelligence development. *AI & Society*, 31(2):201–206, 2016.
- R. M. Axelrod. *The Evolution of Cooperation*. Basic books. Basic Books, 1984. ISBN 9780465021215. URL <https://books.google.com/books?id=NJZBCGbn98C>.
- Robert Axelrod. More effective choice in the prisoner’s dilemma. *Journal of Conflict Resolution*, 24 (3):379–403, 1980.
- Omri Ben-Shahar. Should products liability be based on hindsight? *Journal of Law, Economics, & Organization*, pages 325–357, 1998.
- Maria Bengtsson and Sören Kock. ”coopetition” in business networks—to cooperate and compete simultaneously. *Industrial marketing management*, 29(5):411–426, 2000.
- Nick Bostrom. Strategic implications of openness in ai development. *Global Policy*, 8(2):135–148, 2017a.
- Nick Bostrom. *Superintelligence*. Oxford University Press, 2017b.
- Yves Breitmoser, Jonathan HW Tan, and Daniel John Zizzo. Understanding perpetual r&d races. *Economic Theory*, 44(3):445–467, 2010.
- Timothy F Bresnahan. Competition and collusion in the american automobile industry: The 1955 price war. *The Journal of Industrial Economics*, pages 457–482, 1987.
- Miles Brundage, Shahar Avin, et al. The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. *arXiv e-prints*, art. arXiv:1802.07228, Feb 2018.
- Erik Brynjolfsson, Daniel Rock, and Chad Syverson. Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics. In *The economics of artificial intelligence: An agenda*. University of Chicago Press, 2018.
- Jian Cai, Kent Cherny, Todd Milbourn, et al. Compensation and risk incentives in banking and finance. *Economic Commentary*, (2010-13), 2010.
- G. Calabresi. *The Costs of Accidents: A Legal and Economic Analysis*. Legal and Economic Analysis. Yale University Press, 1970. ISBN 9780300011142. URL <https://books.google.com/books?id=PEMLlb7XxyMC>.
- John L. Campbell. Why would corporations behave in socially responsible ways? an institutional theory of corporate social responsibility. *Academy of Management Review*, 32(3):946–967, jul 2007. doi: 10.5465/amr.2007.25275684. URL <https://doi.org/10.5465/amr.2007.25275684>.

- D Chalmers. The singularity: A philosophical analysis. *Science fiction and philosophy: From time travel to superintelligence*, pages 171–224, 2009.
- Sylvain Chassang. Building routines: Learning, cooperation, and the dynamics of incomplete relational contracts. *American Economic Review*, 100(1):448–65, 2010.
- Yongmin Chen and Xinyu Hua. Competition, product safety, and product liability. *The Journal of Law, Economics, and Organization*, 33(2):237–267, 2017.
- Paul Christiano. AI “safety” vs “control” vs “alignment”. AI Alignment, <https://ai-alignment.com/ai-safety-vs-control-vs-alignment-2a4b42a863cc>, Nov 2016. (Accessed on 03/13/2019).
- Paul Christiano. Hyperbolic growth. *The sideways view*, <https://sideways-view.com/2017/10/04/hyperbolic-growth/>, Oct 2017. (Accessed on 03/21/2019).
- Paul Christiano. Takeoff speeds. *The sideways view*, <https://sideways-view.com/2018/02/24/takeoff-speeds/>, Feb 2018. (Accessed on 03/21/2019).
- Jack Clark and Dario Amodei. Faulty reward functions in the wild. Open AI Blog, <https://openai.com/blog/faulty-reward-functions/>, Dec 2016. (Accessed on 03/13/2019).
- Iain M Cockburn, Rebecca Henderson, and Scott Stern. The impact of artificial intelligence on innovation. Technical report, National Bureau of Economic Research, 2018.
- Cary Coglianese and Evan Mendelson. Meta-regulation and self-regulation. *Regulation*, pages 12–11, 2010.
- Molly Cohen and Arun Sundararajan. Self-regulation and innovation in the peer-to-peer sharing economy. *U. Chi. L. Rev. Dialogue*, 82:116, 2015.
- Consumer Reports. Volkswagen and Audi Recall for Brake Issues. <https://www.consumerreports.org/car-recalls-defects/volkswagen-audi-recall-sedans-suvs-for-brake-issues/>, Aug 2018. (Accessed on 03/12/2019).
- Nina W Cornell, Roger G Noll, and Barry R Weingast. Safety regulation. 1976.
- Missy Cummings. *Artificial intelligence and the future of warfare*. Chatham House for the Royal Institute of International Affairs, 2017.
- Maeve Cushen, J Kerry, M Morris, Malco Cruz-Romero, and Enda Cummins. Nanotechnologies in the food industry—recent developments, risks and regulation. *Trends in food science & technology*, 24(1):30–46, 2012.
- Allan Dafoe. AI Governance: A Research Agenda. *Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK*, 2018.
- Carl J Dahlman. The problem of externality. *The journal of law and economics*, 22(1):141–162, 1979.
- Partha Dasgupta. Trust as a commodity. *Trust: Making and breaking cooperative relations*, 4:49–72, 2000.
- Partha Dasgupta and Joseph Stiglitz. Uncertainty, industrial structure, and the speed of r&d. *The Bell Journal of Economics*, pages 1–28, 1980.
- Andrew F Daughety and Jennifer F Reinganum. Product safety: liability, R&D, and signaling. *The American Economic Review*, pages 1187–1206, 1995.
- Andrew F Daughety, Jennifer F Reinganum, et al. Economic analysis of products liability: theory. *Research Handbook on the Economics of Torts*, pages 69–96, 2013.
- Andrew F Daughety, Jennifer F Reinganum, et al. Market structure, liability, and product safety. *Handbook of Game Theory and Industrial Organization, Volume II: Applications*, 2:225, 2018.
- Lucas W Davis and Catherine Wolfram. Deregulation, consolidation, and efficiency: Evidence from us nuclear power. *American Economic Journal: Applied Economics*, 4(4):194–225, 2012.

- Deven R. Desai and Joshua A. Kroll. Trust but verify: A guide to algorithms and the law. *Harvard Journal of Law & Technology*, 31(1):1, 2017.
- Timothy M Devinney. Is the socially responsible corporation a myth? the good, the bad, and the ugly of corporate social responsibility, 2009.
- Ulrich Doraszelski. An r&d race with knowledge accumulation. *Rand Journal of economics*, pages 20–42, 2003.
- Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.
- William R Dunn. Designing safety-critical computer systems. *Computer*, 36(11):40–46, 2003.
- Ellinor Ehrnberg. On the definition and measurement of technological discontinuities. *Technovation*, 15(7):437–452, sep 1995. doi: 10.1016/0166-4972(95)96593-i. URL [https://doi.org/10.1016/0166-4972\(95\)96593-i](https://doi.org/10.1016/0166-4972(95)96593-i).
- Jon Elster. Rationality, morality, and collective action. *Ethics*, 96(1):136–155, 1985.
- Olivia J Erdélyi and Judy Goldsmith. Regulating artificial intelligence: Proposal for a global solution. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 95–101. ACM, 2018.
- Daniel C Esty and Damien Geradin. *Regulatory competition and economic integration: comparative perspectives*. Oxford University Press, 2001.
- Federal Trade Commission and U.S. Department of Justice. Antitrust guidelines for collaborations among competitors. <https://www.ftc.gov/sites/default/files/attachments/dealings-competitors/ftcdojguidelines.pdf>, April 2000.
- Future of Life Institute. AI Principles. <https://futureoflife.org/ai-principles/?cn-reloaded=1>, 2017. (Accessed on 04/01/2019).
- Shanti Gamper-Rabindran and Stephen R Finger. Does industry self-regulation reduce pollution? responsible care in the chemical industry. *Journal of Regulatory Economics*, 43(1):1–30, 2013.
- Philipp Genschel and Thomas Plumper. Regulatory competition and international co-operation. *Journal of European Public Policy*, 4(4):626–642, 1997.
- Robert Gibbons and Rebecca Henderson. *What do managers do?: Exploring persistent performance differences among seemingly similar enterprises*. Harvard Business School, 2012.
- Joseph T Gilbert and Philip H Birnbaum-More. Innovation timing advantages: From economic theory to strategic application. *Journal of Engineering and Technology Management*, 12(4):245–266, 1996.
- Irving John Good. Speculations concerning the first ultraintelligent machine. In *Advances in computers*, volume 6, pages 31–88. Elsevier, 1966.
- Gene M Grossman and Carl Shapiro. Dynamic R&D competition, 1985.
- Joseph P Guiltinan and Gregory T Gundlach. Aggressive and predatory pricing: A framework for analysis. *Journal of marketing*, 60(3):87–102, 1996.
- Neil Gunningham and Joseph Rees. Industry self-regulation: an institutional perspective. *Law & Policy*, 19(4):363–414, 1997.
- Gillian Hadfield and Jack Clark. Regulatory markets for AI safety. *Safe Machine Learning workshop at ICLR*, 2019.
- G.K. Hadfield. *Rules for a Flat World: Why Humans Invented Law and how to Reinvent it for a Complex Global Economy*. Titolo collana. Oxford University Press, 2017. ISBN 9780199916528. URL <https://books.google.com/books?id=TBYBDQAAQBAJ>.
- Christopher Harris and John Vickers. Racing with uncertainty. *The Review of Economic Studies*, 54(1):1, jan 1987. doi: 10.2307/2297442. URL <https://doi.org/10.2307/2297442>.

- John C Harsanyi. A simplified bargaining model for the n-person cooperative game. *International Economic Review*, 4(2):194–220, 1963.
- Catherine Hausman. Corporate incentives and nuclear safety. *American Economic Journal: Economic Policy*, 6(3):178–206, 2014.
- Douglas D. Heckathorn. Collective action and the second-order free-rider problem. *Rationality and Society*, 1(1):78–100, 1989. doi: 10.1177/1043463189001001006. URL <https://doi.org/10.1177/1043463189001001006>.
- Spencer Henson and Julie Caswell. Food safety regulation: an overview of contemporary issues. *Food policy*, 24(6):589–603, 1999.
- Anthony Heyes. Is environmental regulation bad for competition? a survey. *Journal of Regulatory Economics*, 36(1):1–28, 2009.
- Fred Matthew Hohman, Minsuk Kahng, Robert Pienta, and Duen Horng Chau. Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE transactions on visualization and computer graphics*, 2018.
- Katharina Holzinger. The problems of collective action: A new approach. *MPI Collective Goods Preprint*, (2003/2), 2003.
- Keith N Hylton. The law and economics of products liability. *Notre Dame L. Rev.*, 88:2457, 2012.
- Geoffrey Irving, Paul Christiano, and Dario Amodei. Ai safety via debate. *arXiv preprint arXiv:1805.00899*, 2018.
- Neil Johnson, Guannan Zhao, Eric Hunsader, Hong Qi, Nicholas Johnson, Jing Meng, and Brian Tivnan. Abrupt rise of new machine ecology beyond human response time. *Scientific reports*, 3: 2627, 2013.
- By Kenneth L Judd, Karl Schmedders, and Şevin Yeltekin. Optimal rules for patent races. *International Economic Review*, 53(1):23–52, 2012.
- Andrew A King and Michael J Lenox. Industry self-regulation without sanctions: The chemical industry’s responsible care program. *Academy of management journal*, 43(4):698–716, 2000.
- Steven Klepper. Entry, exit, growth, and innovation over the product life cycle. *The American Economic Review*, 86(3):562–583, 1996. ISSN 00028282. URL <http://www.jstor.org/stable/2118212>.
- Charles D. Kolstad, Thomas S. Ulen, and Gary V. Johnson. Ex post liability for harm vs. ex ante safety regulation: Substitutes or complements? *The American Economic Review*, 80(4):888–901, 1990. ISSN 00028282. URL <http://www.jstor.org/stable/2006714>.
- Victoria Krakovna. Clopen ai: Openness in different aspects of ai development. <https://vkrakovna.wordpress.com/2016/08/01/clopen-ai-openness-in-different-aspects-of-ai-development/>, Aug 2016. (Accessed on 06/27/2019).
- Victoria Krakovna. Specification gaming examples in ai. <https://vkrakovna.wordpress.com/2018/04/02/specification-gaming-examples-in-ai/>, April 2018. (Accessed on 05/14/2019).
- Sarah Kreps. Is this a sputnik moment for artificial intelligence? or why the cold war past should not be prologue (unpublished manuscript). 2019.
- Andrew H Kydd. *Trust and mistrust in international relations*. Princeton University Press, 2007.
- William M. Landes and Richard A. Posner. A positive economic analysis of products liability. *The Journal of Legal Studies*, 14(3):535–567, 1985. ISSN 00472530, 15375366. URL <http://www.jstor.org/stable/724257>.
- Howard Latin. Good science, bad regulation, and toxic risk assessment. *Yale J. on Reg.*, 5:89, 1988.
- Edward Lazear. Intergenerational externalities. *Canadian Journal of Economics*, pages 212–228, 1983.

Jan Leike, Miljan Martic, Victoria Krakovna, Pedro A Ortega, Tom Everitt, Andrew Lefrancq, Laurent Orseau, and Shane Legg. Ai safety gridworlds. *arXiv preprint arXiv:1711.09883*, 2017.

M Lenox. The prospects for industry self-regulation of environmental externalities. *Making global regulation effective: What role for self-regulation*, 2007.

François Lévéque. Externalities, collective goods and the requirement of a state's intervention in pollution abatement. In *Voluntary Approaches in Environmental Policy*, pages 17–26. Springer, 1999.

Marvin B Lieberman and David B Montgomery. First-mover advantages. *Strategic management journal*, 9(S1):41–58, 1988.

George J Mailath, Larry Samuelson, and J Swinkels. Extensive form reasoning in normal form games. *Foundations in Microeconomic Theory*, page 451, 1991.

Constantinos C Markides and Paul A Geroski. *Fast second: How smart companies bypass radical innovation to enter and dominate new markets*, volume 325. John Wiley & Sons, 2004.

Talya Minsberg. The newfound power of tech workers. *The New York Times*, <https://www.nytimes.com/2019/03/02/business/tech-employees-protests.html>, Mar 2019. (Accessed on 03/14/2019).

Kara Morgan. Development of a preliminary framework for informing the risk analysis and risk management of nanoparticles. *Risk Analysis: An International Journal*, 25(6):1621–1635, 2005.

Luke Muehlhauser and Bill Hibbard. Viewpoint exploratory engineering in artificial intelligence. *Communications of the ACM*, 57:32–34, 09 2014. doi: 10.1145/2644257.

Martin A Nowak. Five rules for the evolution of cooperation. *science*, 314(5805):1560–1563, 2006.

Walter Y Oi et al. The economics of product safety. *Bell Journal of Economics*, 4(1):3–28, 1973.

Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The building blocks of interpretability. *Distill*, 3(3):e10, 2018.

OpenAI. Openai charter. <https://openai.com/charter/>, April 2018. (Accessed on 04/01/2019).

Pedro A. Ortega, Vishal Maini, et al. Building safe artificial intelligence: specification, robustness, and assurance. <https://medium.com/@deeppindsafetyresearch/building-safe-artificial-intelligence-52f5f75058f1>, Sep 2018. (Accessed on 03/13/2019).

Cullen O’Keefe. Antitrust-compliant ai industry self-regulation. forthcoming 2019.

Cullen O’Keefe et al. The windfall clause. forthcoming 2019.

Christine Parker. Meta-regulation: legal accountability for corporate social responsibility. 2007.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems*, pages 5680–5689, 2017.

A. Mitchell Polinsky and Steven Shavell. The uneasy case for product liability. *Harv. L. Rev.*, 123: 1437, 2009.

Adrien Querbes and Koen Frenken. Evolving user needs and late-mover advantage. *Strategic organization*, 15(1):67–90, 2017.

Alec Radford, Jeffrey Wu, et al. Language models are unsupervised multitask learners. *OpenAI Blog*, 2019.

Mooweon Rhee and Pamela R Haunschild. The liability of good reputation: A study of product recalls in the us automobile industry. *Organization Science*, 17(1):101–117, 2006.

DR Robinson and DJ Goforth. Conflict, no conflict, common interests, and mixed interests in 2×2 games. *Department of Mathematics and Computer Science*, 2005.

Paul H Rubin. Markets, tort law, and regulation to achieve safety. *Cato J.*, 31:217, 2011.

- Maurice Schellekens. Self-driving cars and the chilling effect of liability law. *Computer Law & Security Review*, 31(4):506–517, 2015.
- T.C. Schelling. *Micromotives and Macrobbehavior*. W. W. Norton, 2006 [1978]. ISBN 9780393069778. URL <https://books.google.com/books?id=DenWKRgqzWMC>.
- Thomas C Schelling. *The strategy of conflict*. Harvard university press, 1980.
- Paul Seabright. Managing local commons: theoretical issues in incentive design. *Journal of economic perspectives*, 7(4):113–134, 1993.
- Michael Segal. How automation is changing work. *Nature*, 563(7733):S132–S135, Nov 2018. doi: 10.1038/d41586-018-07501-y. URL <https://doi.org/10.1038/d41586-018-07501-y>.
- Murray Shanahan. *The technological singularity*. MIT Press, 2015.
- Steven Shavell. Liability for harm versus regulation of safety. *The Journal of Legal Studies*, 13(2): 357–374, 1984.
- Eytan Sheshinski. Price, quality and quantity regulation in monopoly situations. *Economica*, 43(170): 127–137, 1976.
- Brad Smith. Facial recognition technology: The need for public regulation and corporate responsibility. <https://blogs.microsoft.com/on-the-issues/2018/07/13/facial-recognition-technology-the-need-for-public-regulation-and-corporate-responsibility/>, Jul 2018a. (Accessed on 03/20/2019).
- Brad Smith. Facial recognition: It's time for action. *Microsoft on the Issues*, <https://blogs.microsoft.com/on-the-issues/2018/12/06/facial-recognition-its-time-for-action/>, Dec 2018b. (Accessed on 03/20/2019).
- Joseph Stiglitz. Regulation and failure. *New perspectives on regulation*, 576, 2009.
- Cass R Sunstein. Irreversible and catastrophic. *Cornell L. Rev.*, 91:841, 2005.
- The Atlantic. Who is liable for a death caused by a self-driving car? <https://www.theatlantic.com/technology/archive/2018/03/can-you-sue-a-robocar/556007/>, Mar 2018. (Accessed on 03/12/2019).
- Wenpin Tsai. Social structure of “coopetition” within a multiunit organization: Coordination, competition, and intraorganizational knowledge sharing. *Organization science*, 13(2):179–190, 2002.
- Hayley Tsukayama and Jamie Williams. If a pre-trial risk assessment tool does not satisfy these criteria, it needs to stay out of the courtroom | electronic frontier foundation. EFF, <https://www.eff.org/deeplinks/2018/11/if-pre-trial-risk-assessment-tool-does-not-satisfy-these-criteria-it-needs-stay>, Nov 2018. (Accessed on 03/13/2019).
- Roman V. Yampolskiy. Taxonomy of pathways to dangerous AI. *arXiv preprint arXiv:1511.03246*, 2015.
- Eliezer Yudkowsky. Intelligence explosion microeconomics. <https://intelligence.org/files/IEM.pdf>, 2013. (Accessed on 03/21/2019).
- Remco Zwetsloot and Allan Dafoe. Thinking About Risks From AI: Accidents, Misuse and Structure, Feb 2019. *Lawfare*, <https://www.lawfareblog.com/thinking-about-risks-ai-accidents-misuse-and-structure> (Accessed 03/07/2019).