PUBLICATION POLICIES AND MODEL-SHARING DECISIONS: A LITERATURE REVIEW AND RECOMMENDATIONS FOR AI LABS

Akash R. Wasil
Independent
wasil@sas.upenn.edu

Charlotte Siegmann
Global Priorities Institute
Oxford University
csiegmann@outlook.com

Carson Ezell
Harvard University
Harvard AI Safety Team
cezell@college.harvard.edu

Aris Richardson
Independent
alarichardson@berkeley.edu

Abstract

Research organizations that develop advanced artificial intelligence (AI) systems face difficult decisions about how widely to share their research findings, methods and models. Designing responsible publication policies and model-sharing practices is an important priority for responsible AI labs. However, designing such policies is difficult: findings in AI can contribute to catastrophic risks, and these effects often occur though complicated mechanisms (e.g., by increasing hype, worsening race dynamics, and influencing other structural risk factors). To inform responsible publication policies and model-sharing practices, we reviewed fields with a history of dual-use information-sharing concerns to generate recommendations for AI labs. In section one (Background), we review risks associated with advanced AI systems (accident risks, misuse risks, structural risk factors) and discuss how publication policies and model-sharing policies can influence these risks. In section two (Literature Review), we review information-sharing policies in fields that regularly encounter dual-use concerns, such as the biological sciences, cybersecurity and nuclear technologies; we also review existing norms in AI/machine learning research. In section three (Proposals), we draw from our literature review to make recommendations to AI labs. We recommend establishing Catastrophic Risk Review Boards that consist of members from safety and security teams, applying catastrophic risk questionnaires prior to publication and model-sharing decisions, and implementing the responsive access paradigm that involves storing findings in private time-stamped repositories. We conclude by identifying proposals and questions that could be explored in future research.

1 Introduction

Several research groups are aiming to advanced artificial intelligence (AI) systems that may rival human performance in the 21st century. The field currently consists of a few major research groups (e.g., OpenAI, Deepmind, Anthropic, Google Brain, Meta), several smaller research groups, and new research groups joining every few months.

Research groups aiming to develop powerful AI systems (which we'll refer to as "AGI labs" for convenience) have attracted considerable attention in recent years. For instance, in the last year, Microsoft invested $10B into OpenAI. Google, which already owned DeepMind and Google Brain, invested $300M into Anthropic. There are also several startups that are aiming to build AGI; notable examples include AdeptAI and Generally Intelligent, both of which were founded in 2022. Given the high amount of investment in the space, it seems likely that new AGI labs will be created each year.

While the benefits of advanced AGI are enormous, there is also widespread recognition that AGI could have catastrophic harms. Some leaders of AGI labs have publicly made statements acknowledging extreme risks from AI. For example:

> "The bad case, and I think this is important to say, is lights out for all of us..." — Sam Altman, CEO of OpenAI (Jackson, 2023)

> "When it comes to very powerful technologies—and obviously AI is going to be one of the most powerful ever—we need to be careful. Not everybody is thinking about those things. It's like experimentalists, many of whom don't realize they're holding dangerous material." — Demis Hassabis, CEO of DeepMind (Perrigo, 2023b)

> "So far, no one knows how to train very powerful AI systems to be robustly helpful, honest, and harmless. Furthermore, rapid AI progress will be disruptive to society and may trigger competitive races that could lead corporations or nations to deploy untrustworthy AI systems. The results of this could be catastrophic, either because AI systems strategically pursue dangerous goals, or because these systems make more innocent mistakes in high-stakes situations." (Anthropic, 2023)

To reduce risks, there has been some interest in stricter publication policies, model-sharing policies, and information-sharing norms. In an interview with TIME, DeepMind CEO Demis Hassabis suggested that the AI industry's culture of publishing findings openly may need to end (Perrigo, 2023b). Furthermore, in a recent blog post, Anthropic mentioned that it rarely publishes AI capabilities in order to avoid advancing the rate of AI progress (Anthropic, 2023). These statements contrast with open publication practices that have been common in the field of AI (and adjacent disciplines like machine learning and software engineering). Open publication practices have several benefits, but they also directly and indirectly increase potential risks from advanced AI systems (Bostrom, 2017; Shevlane and Dafoe, 2020).

AI progress is often dual-use (the same information can be used for beneficial purposes or harmful purposes) and AI advances can contribute to harmful race dynamics. Research that aims to identify best practices from other fields with dual-use concerns could be valuable. In an influential report about the malicious use of AI, the authors identified four high-level recommendations, and two of these areas were focused on navigating dual-use concerns: (1) researchers and engineers in AI labs should take dual-use concerns seriously and adopt research norms to address these concerns & (2) Future work should identify best practices from research areas with mature methods for addressing dual-use concerns (Brundage et al., 2018).

Publication and model-sharing decisions by AGI labs often have important consequences for the entire AI field. For example, the release of GPT-3 likely sped up the diffusion process of GPT3-like models by creating more publicity, revealing algorithmic insights, and contributing to the proliferation of open-source tools (Cottier, 2022). Other examples include the release of the Chinchilla scaling laws paper (which may have increased the number of actors capable of developing powerful language models), the release of ChatGPT (which appears to have increased acceleration and hype relating to AI progress), and the release of Bing Chat (which increased awareness about potential risks and safety concerns from AI systems; Hoffmann et al., 2022; Metz and Weise, 2023; Perrigo, 2023a).

How might we preserve some of the benefits of an open scientific research culture while mitigating risks from sharing dual-use information? To explore this question, we reviewed information-sharing practices from disciplines with more mature methods for addressing dual-use concerns. Then, inspired by best practices from other fields, we propose recommendations to AI labs. Our paper is organized as follows:

1. In section one (Background), we review some risks associated with advanced AI systems and describe how information-sharing norms affect these risks.
2. In the section two (Literature Review), we review information-sharing practices in the areas of biosecurity, cybersecurity, nuclear technologies, and AI.
3. In section three (Proposals), we recommend three information-sharing policies that AGI labs can adapt and implement.
4. In section four (Future Research), we list policies that could be further developed by future research, as well as future research that can inform information-sharing policies more generally.

2 Background

2.1 Potential risks from advanced AI systems

While the benefits of advanced AGI are enormous, AGI development and deployment could also cause a catastrophes. Broadly, there are three kinds of risks that are regularly discussed in AGI safety literature:

Table 1: AI risk types and explanations

Risk type	**Explanation**
Accident risks	An AI lab could develop a powerful AI system without knowing how to control it. The AI system could be capable enough to overpower humanity and produce catastrophic harms (see Bostrom 2012; Carlsmith 2022; Ngo et al. 2023).
Misuse risks	A nefarious group could explicitly instruct an AI to cause harm or undermine security (see Brundage et al. 2018).
Structural risk factors	Various cultural, economic, diplomatic, and competitive pressures can adversely affect safe AI development and deployment. As an example, pressure to outcompete competitors may produce incentives to "race" to develop and deploy AI models prematurely (see Zwetsloot and Dafoe 2019).

Accident risks. AI systems could pose catastrophic risks even if no one explicitly "commands" them to harm anyone. They are risks caused by accident or negligence (e.g., failing to develop a system that could be controlled) rather than malice (e.g., explicitly instructing a system to cause harm).

More specifically, it can be extremely difficult for AI researchers to determine the objective of a system. Current techniques for developing AI systems often involve "trial-and-error-learning", in which AIs receive reinforcement for performing behaviors that appear to be desirable. However, such techniques do not provide researchers with a detailed understanding of how the system learns, what its final objective is, or what its internal cognition looks like. Furthermore, it is often difficult for humans to provide accurate feedback on tasks, and this problem is expected to become more difficult as systems get closer to AGI (especially when systems are expected to solve problems that humans cannot solve). While these dangers tend to be limited with current systems, many AGI safety researchers fear that these issues would lead to catastrophic consequences for more powerful AI systems. Furthermore, some dangers may only present themselves once systems are sufficiently intelligent (e.g., intelligent enough to become situationally aware, develop long-term goals and meaningfully deceive humans). Note that a thorough review of these risks is outside the scope of this paper [1].

Misuse risks. Misuse risks are risks that involve someone explicitly instructing an AI to cause harm or undermine security. Misuse risks can be digital (e.g., AI-assisted cyberattacks), physical (e.g., AI-operated drone strikes), or political (e.g., AI-assisted surveillance; Brundage et al., 2018). For example, imagine that a nefarious group used a powerful AI to hack into important military organizations, launch attacks on enemies, and create persuasive online propaganda techniques. With sufficiently powerful AGI, such a group might even be able to overpower humanity and permanently affect the trajectory of human civilization. As another example, an authoritarian government may be able to weaponize AGI to forcefully maintain its regime, defeat foreign enemies, and potentially even establish itself as the only major global power.

Structural risk factors. Structural risk factors involve the cultural, economic, diplomatic, and competitive environments that affect the development and deployment of AI systems. A structural perspective on AI risk is meant to call attention to incentive structures and contextual factors that can affect AI risk (Zwetsloot and Dafoe, 2019). Structural risk factors generally increase the likelihood of both accident risks and misuse risks. Competitive market forces are an example of a structural risk factor in AGI development. Pressure to outcompete competitors may produce incentives to build and deploy AI models prematurely (Bostrom, 2017; Dafoe, 2018). Cautious AGI labs that are highly concerned about safety may feel pressure to cut back on safety measures in order to stay competitive with AGI labs that are less cautious (or perceived as less cautious).

2.2 Publication policies, model-sharing decisions, and AI risk

AGI labs are frequently faced with complicated choices about what information they should share with whom. Historically, the AI community is thought to value a culture of openness and diffusion (Fischer et al., 2021). Papers are

[1] To better understand these risks, we recommend the following resources: Bostrom (2012) for a conceptual understanding of why systems might learn undesirable goals; Hubinger et al. (2021) for a more mechanistic model of how systems could learn to optimize for undesirable goals; Shah et al. (2022) for an empirical understanding of goal misgeneralization with current models; Ngo et al. (2023) for a more technical understanding of various problems with AI alignment; Carlsmith (2022) for why systems may be power-seeking and how this could lead to catastrophic harms; Branwen (2022) for a hypothetical takeover scenario.

often released on publicly available repositories (e.g., arxiv.org), code is often freely available, and many datasets are publicly available. In general, scientific diffusion has several advantages, such as (a) expanding access to information, (b) allowing other researchers to detect bugs or mistakes, (c) improving the reproducibility of a field's findings, and (d) increasing the rate of scientific progress.

However, diffusion can be dangerous when sharing information has the potential to cause harm through (a) accidents, (b) misuse, or (c) structural risks. We summarized these downsides in the table above. Some fields have more mature norms around what kind of information should be shared, how it should be shared, and with whom it should be shared. Such fields often have strong explicit norms (e.g., information-sharing policies and procedures for scientists) and implicit norms (e.g., a general culture in which individual scientists are expected to be cautious) that reduce the likelihood that potentially-dangerous findings are widely or irresponsibly shared. Examples of fields with more mature (but still imperfect and limited) information-sharing norms include biological sciences, cybersecurity, and nuclear technologies (Miller and Selgelid, 2007; Miller, 2018; Riebe and Reuter, 2019). We reasoned that important lessons can be learnt from these fields nonetheless, because experts have spent years thinking and debating about how to balance the benefits and harms of information-sharing (Miller and Selgelid, 2007; Miller, 2018; Riebe and Reuter, 2019).

Importantly, publication decisions and model-sharing decisions are not "all or nothing." Labs have flexibility over what they release, and there is a variety of relevant information (e.g., model weights, prompting access or the ability to fine-tune models, knowledge about the capabilities of the model, and information about the training process). They can decide whether to share the information with the public, screened applications, AI safety researchers or other trusted third parties, or specific individuals. They can also design systems that make it less likely for information to spread beyond the selected parties. Moreover, they can decide when and how to release the information; options include staged release (in which increasingly powerful versions of a model are released over time), structured access (in which developers restrict the kinds of capabilities that individuals can access to reduce risks), and discretion regarding which parties to include in sharing decisions (see Shevlane, 2022). When we refer to publication policies and model-sharing policies, we refer to all of these dimensions and design principles.

As AI systems become increasingly powerful, AGI labs may wish to adopt new publication-sharing and model-sharing policies to reduce catastrophic risks.

Reducing accident risks. Recall that catastrophic accident risks occur when a system is developed or deployed that humanity cannot control. Accident risks are minimized if the first group to develop AGI is a group that has a strong safety culture and applies a strong set of precautions in AI development and deployment. For example, a safety-conscious lab might have plans in place to detect potential concerns around misalignment (e.g., by developing and applying tools to detect and eliminate deception), eliminate potential sources of risk (e.g., by developing and applying state-of-the-art safety techniques), carefully evaluating the system in a wide range of environments before widespread deployment (e.g., by constructing test environments in which the systems are most likely to misbehave), and adopting other safety standards (see Barrett et al. 2023; Dafoe 2018; Schuett and Anderljung 2022).

Accounting for the unilateralist's curse (Bostrom et al., 2016) increases the risk of accidents: even if 90% of the groups trying to build AGI are highly safety-conscious, the 10% that are least safety-conscious may be most likely to scale and deploy powerful models. This reasoning suggests that indiscriminate information-sharing policies (e.g., publicly releasing materials) disproportionately favors incautious actors. Incautious actors developing state-of-the-art models

implement fewer technical safety techniques, thereby increasing the risk of catastrophic accidents. Conversely, strong information-sharing policies could reduce the risk that less safety-conscious actors build AGI, widen the gap between top labs and competitors, allow leading labs to invest more time into safety research, and reduce the chance of accident risks.

Reducing misuse risks. As the AI technology becomes more powerful, there will be more malicious actors trying to acquire and misuse these tools. In the short-term, unrestricted access to information (e.g. details about how an AI system was trained or the model weights) can help nefarious actors deploy the models for bad purposes, e.g. persuasion or surveillance, see Bai et al. (2023). Sharing can also accelerate an actor's progress toward AGI. This is especially salient given that much of the top talent in AI are currently located in the US and UK. In practice, this makes it difficult for scientists in authoritarian regimes to discover such insights (e.g., state-of-the-art algorithms, architectures, and scaling laws). However, a culture of scientific openness and diffusion makes it relatively easy for groups to *copy* and *reproduce* insights (Brundage et al., 2018).

In the long-term, if a malicious actor were to get access to AGI, AGI could be used to radically shift the balance of power in the world. In extreme cases, this could present a catastrophic risk, either through a great power conflict (e.g., war involving AGIs) or a future shaped by the interests of an unstoppable authoritarian state (e.g., value lock-in; see Finnveden et al. 2022).

Reducing structural risks from race dynamics. We use the term *race dynamics* to refer to pressure to be the "first" to hit various AI milestones and to "keep up" with competitors (e.g., to get press attention and investors; to capture the benefits of AGI). Race dynamics are a structural risk factor that can push actors to be less cautious (e.g., by investing fewer resources into safety research or prematurely scaling/deploying models). In scenarios where there are relatively few groups developing AGI, race dynamics are likely to be less strong– especially if the few groups are safety-conscious and recognize each other as safety-conscious. In contrast, imagine a scenario in which there are 20 major AGI labs, and they are all fairly "close in the race" (e.g., all labs are within 3 months of catching up to the leading lab). In this scenario, structural risks stemming from race dynamics are likely to be higher. On one hand, the race is more intense, meaning that an incautious actor is more likely to develop AGI. Furthermore, the race is perceived as more intense, meaning that all actors will feel more pressure to cut back on safety and caution. This can lead to an unfortunate cycle, in which safety-conscious actors feel pressure to cut back on safety in order to stay competitive, which leads the average safety-consciousness to decrease, which produces even *more* pressure for labs to cut back on safety (for more on race dynamics, see Bostrom 2017; Dafoe 2018).

Responsible information-sharing policies can reduce race dynamics between leading labs by reducing AI hype, limiting the number of actors who have the knowledge needed to build AGI, and reducing pressure on leading labs to prematurely accelerate. Notably, this relies on the leading actors being relatively responsible. If this is not the case, concentrating power in a small number of firms could be undesirable (see Solaiman 2023). While distributing power among various actors is generally a valuable aim, this aim can be counterproductive in situations with dangerous technologies that involve race dynamics. For example, as mentioned earlier, competitive pressures can produce a unilateralist's curse that favors the least cautious actors (Bostrom et al., 2016). To the extent that policies can reduce the likelihood that incautious actors are able to develop state-of-the-art systems, such policies could play an important role in reducing the likelihood of accident risks, misuse risks, and dangerous race dynamics that contribute to catastrophic risks.

Disclosure can be irreversible. Once information is released to an audience, it can't be taken (instead trying to hide the information might even make it worse). In contrast, not deciding to disclose information is a reversible decision. One can still decide to release the information at any future point. This may cause an asymmetry between disclosing and not disclosing information and should be taken into account when making decisions under deep uncertainty.

With these potential benefits in mind, information-sharing policies also come with costs.

Disclosure can help safety researchers. In many fields, greater scientific openness can improve safety by allowing external researchers to detect bugs and flaws. Naive information-sharing policies may run the risk of denying access to information to individuals who could help with AGI safety research by detecting flaws, performing red-teaming efforts, and understanding in what contexts the systems are dangerous.

Information-sharing policies can lead to slower progress. Another potential disadvantage is that information-sharing policies could slow down the rate of AGI development, thus delaying the potential benefits of AGI. As has been discussed in previous research, however, we do not expect this effect to be large: both because the actual effect on AGI timelines is likely to be small (especially relative to the points about race dynamics) and because delaying AGI timelines has the beneficial effect of providing more time for AGI safety research.

Information-sharing policies can clash with corporate interests. AGI labs often have commercial incentives to publish papers or release models. Sharing research findings can generate investment opportunities and help an AGI lab attract talented researchers. All else equal, we believe the information-sharing policies can be implemented in ways that are attentive to the corporate interests of AGI labs. Nonetheless, we acknowledge that there will occasionally be unavoidable trade-offs between information-sharing policies and corporate interests.

2.3 The present paper

Thus far, we have presented reasons why information-sharing policies may help AGI labs reduce accident risks, misuse risks, and structural risk factors from AGI. We also briefly covered some potential negative effects of information-sharing policies and some benefits of scientific openness that would ideally be preserved.

For the rest of the paper, we focus on specific information-sharing policies for AGI labs. First, we present a literature review of information-sharing policies in other fields that involve sensitive or dual-use information. When reviewing government policies, we focus on the United States and the United Kingdom, because the leading AGI labs are currently based in these countries. Then, we provide concrete suggestions for AI labs. Finally, we offer directions for future research.

3 Literature review: Information security practices across fields

3.1 Biological sciences

> "There are times when research intended to help find cures for infectious diseases could also help terrorists make a bioweapon. The new panel must consider whether the benefits of publishing such a paper are outweighed by the risks it might pose to national security." – Joe Palca, NPR (Palca, 2005)

In the biological sciences, some types of infectious disease research are considered dual-use: the same scientific finding could be used for beneficial purposes (e.g., helping other researchers make medical discoveries or build future vaccines) or harmful purposes (e.g., helping malevolent actors develop dangerous pathogens or biological weapons).

In some cases, an informed risk assessment would require access to security expertise or classified information, such as knowledge about bioterrorist groups and the likelihood that they would have access to certain kinds of pathogens (Miller and Selgelid, 2007). In addition, synthetic biology researchers are not trained to adopt a "security mindset" and proactively consider the security implications of their work by default (Diggans and Leproust, 2019). For these reasons, it is especially important to establish processes, guidelines, and consultations or collaborations that can inform high-stakes decisions. As one example, the US National Research Council (NRC) taxonomy provides guidance to biological scientists. For example, it defines certain "experiments of concern", such as experiments which could increase the likelihood of developing a biological weapon (National Research Council, 2004). As another example, The American Society for Microbiology requires peer reviewers to bring dual-use concerns to the attention of editors (McLeish and Nightingale, 2007).

In industry, some companies produce synthetic DNA for consumers. However, malevolent actors can use synthetic DNA for dangerous purposes. As a result, some companies have implemented screening procedures to ensure that (a) there is no evidence that the customer is a bad actor and (b) the DNA does not encode for a dangerous pathogen or toxin that the customer is not authorized to access (Hoffmann et al., 2023; Pálya and Delaney, 2023). The US Department of Health and Human Services created guidance for DNA screening in 2010 (Department of Health and Human Services, 2010) that is currently being revised, although the guidelines are voluntary. One could imagine similar practices being used by AI companies to screen customers and their proposed use cases before sharing papers or API access.

One concrete example of an intervention in the biological sciences was the establishment of the National Science Advisory Board for Biosecurity (NSABB). After the 2001 Anthrax attacks, there was increased awareness about the potential of biological research to be used for harmful purposes (Casadevall et al., 2014). To address such concerns, the US government established the NSABB to address issues relating to biosecurity and dual-use research. The NSABB reports to the US Department of Health and Human services, and consists of up to 25 members with expertise in a variety of relevant fields (e.g., molecular biology, national defense, technology, immunology, and public health). While they are advisory and non-binding, the group has published codes of conduct for dual-use research, strategies to address biosecurity concerns relating to synthetic biology, educational resources for scientists about dual-use research, frameworks for risk-benefit assessments, and proposed regulations for gain-of-function research (National Science Advisory Board for Biosecurity, 2007, 2010, 2015).

Additionally, the NSABB can review publications and request that scientists remove sensitive information. For example, in 2011, the NSABB reviewed a paper that was submitted to *Science*. The authors of the paper discovered a way to modify highly pathogenic avian influenza H5N1 such that it could be transmitted by aerosol or respiratory droplets; in other words, they found a way to turn the H5N1 into an airborne virus in mammals. This prompted the NSABB to recommend that the scientists remove critical methodological details in the article, due to concerns that such information could be used by nefarious actors to create a deadly and transmissible human pandemic.

The NSABB recommendation was followed by a voluntary year-long moratorium, in which leading H5N1 researchers agreed to pause gain-of-function research while scientists and government officials evaluated the risks from this research, developed new safety standards, and updated policies relating to gain-of-function research. Eventually, the moratorium was lifted, the H5N1 original article was published in Science, and additional gain-of-function research resumed (Malakoff, 2013). Notably, the intervention by the NSABB, as well as the voluntary moratorium, bought time for

researchers and governments to develop and implement new safety standards designed to increase oversight and reduce risks from gain-of-function research.

Notably, many scholars have pointed out that the current regulations for dual-use biological research may be insufficient, and many of these regulations lack sufficient power. For example, NSABB recommendations do not need to be followed. While the US has policies for required reviews of potentially dual-use research, these only apply if the projects are funded by particular government agencies (Pannu et al., 2022). In some cases, scientists have declared that they would have published a paper even if the NSABB had recommended against publication (Kennedy, 2005). Furthermore, in some cases, a government agency might not have enough expertise to conduct an appropriate review process. Review processes conducted by independent organizations may demonstrate greater expertise, and such processes can also be developed within academic institutions, private organizations, or independent coalitions rather than a national government.

There are also several cases in which individual scientists have voluntarily taken actions to responsibly assess, communicate, and mitigate risks. For example, in 2012, scientists discovered a new strain of a highly lethal nerve toxin. To prevent the new toxin from being misused, they excluded its genetic sequence from their initial publication (Barash and Arnon, 2014). Meanwhile, they shared the sequence with colleagues in order to quickly develop an antitoxin. There are also programs to train scientists to identify and avoid working on risky research.

3.2 Cybersecurity

Information-sharing in cybersecurity can benefit defenders or attackers. Organizations might share information to help others detect threats, alert groups that vulnerabilities have been exploited, or work together to patch vulnerabilities (Skopik et al., 2016; Tosh et al., 2015). However, sharing information widely can allow attackers to exploit vulnerabilities. Furthermore, organizations can face legal penalties for sharing confidential data, and they may face public scrutiny for revealing vulnerabilities in their systems (Tosh et al., 2015). There are also technical bottlenecks to sharing cybersecurity data. Cybersecurity data standards are often not interoperable, there are not many ways to control the sharing of sensitive information, and there are not many systems for automated sharing (which is increasingly necessary to stay ahead of rapidly evolving threats, Dandurand and Serrano, 2013).

Various organizations have created standards to improve the interoperability of cybersecurity data: the MITRE Corporation, Internet Engineering Task Force (IETF), National Institute of Standards and Technology (NIST), and International Telecommunications Union (ITU; Dandurand and Serrano, 2013; Kampanakis, 2014). While interoperable standards facilitate data sharing among organizations that wish to share data, they do not resolve other bottlenecks to data sharing, such as a lack of trust or legal uncertainties.

The US Government has taken measures to make it easier for individuals to share safety-relevant information with trusted officials. For example, the 2015 Cybersecurity Information Sharing Act (CISA) offered legal immunity to private actors that share cyber threats and vulnerabilities with organizations or the government. However, the measure did not fully eliminate liability concerns (Pala and Zhuang, 2019) or improve trust (Sedenberg and Dempsey, 2018) for non-private organizations. For example, guidelines for information sharing under CISA suggest removing private information before sharing data, and organizations may still face liability if they do not follow best practices(Pala and Zhuang, 2019).

Coordination and trust among relevant actors can be increased through computer emergency response teams (CERTs), sometimes called computer security incident response teams (CSIRTs; Choucri et al., 2016; Ruefle et al., 2014). CERTs focus on the prevention and mitigation of cyber threats through frequent data sharing between organizations and rapid response capabilities (e.g. maintaining hotlines or issuing alerts). CERTs can be created between organizations or initiated at a national level (Ruefle et al., 2014). CERTs allow for the standardization of cyber information data structures, automated processes for data sharing, and shared norms for sensitive information controls. For example, the Traffic Light Protocol is a common framework for describing information recipient behavior (Ruefle et al., 2014). It includes four levels of security: white (open to sharing with anyone), green (share with closely trusted individuals), amber (share only on a need-to-know basis), and red (do not share with anyone else; Ruefle et al., 2014). There are also organizations to facilitate the sharing of best practices among CERTs, including the CERT Coordination Center (CERT/CC) and the Forum of Incident Response and Security Teams (FIRST; (Choucri et al., 2016)).

The CERT/CC has made recommendations related to coordinated vulnerability disclosure (CVD). Under CVD, a group of actors that coordinate to identify vulnerabilities, share the vulnerabilities with each other and other safety-conscious actors, and work together to identify ways to address the vulnerability. Vulnerabilities and mitigation measures are shared with the public only *after* the vulnerabilities have been addressed to prevent increased risks (Householder et al., 2017).

The United States Government also promotes sector-specific cybersecurity information sharing between public and private actors through Information Sharing and Analysis Centers (ISACs) and Information Sharing and Analysis Organizations (ISAOs). ISACs were first established by Presidential Decision Directive-63 in 1998. They consist of groups of organizations in the same sector related to critical infrastructure. ISACs facilitate the sharing of threat-related information, including cyber or physical threats (Ezhei and Tork Ladani, 2017), to be analyzed by a group of industry experts and shared with other members of the ISAC as appropriate (Choucri et al., 2016). ISAOs were created by Executive Order 13691 in 2015 to promote information sharing among organizations that would benefit from it but do not fit into an established sector. This includes facilitating the process of getting security clearance for private sector individuals in ISAOs (Choucri et al., 2016). One example of an ISAO is the Maritime and Port Security ISAO (MPS-ISAO), which brings together a variety of actors relevant to the maritime industry (Vijayan, 2022). Activities of the MPS-ISAO include engaging in cybersecurity exercises, identifying risks to critical infrastructure created by third-party contractors, and aligning response protocols across the industry to physical and cyber threats (Kobza, 2017).

3.3 Nuclear technologies

The United States Government took significant measures to significantly restrain non-governmental nuclear research and access to information about nuclear technologies after World War II. In the Atomic Energy Act of 1946, information about nuclear production and the usage of nuclear materials was “born secret”: it was classified by default, in contrast to standard procedures, in which the data had to be specifically classified (Morland, 2004). The reformed Atomic Energy Act of 1954 allowed for private actors to engage in nuclear research and obtain nuclear-related patents, but they needed to have a license which required a security clearance (Cheh, 1979).

The United States also implemented export controls to prevent the proliferation of nuclear materials. Some nuclear materials are included on the Commerce Control List, which is a list of dual-use items with export controls administered

by the Department of Commerce (Fergusson and Kerr, 2013). Some nuclear materials are also on the US Munitions List, which is a list of controlled defense items administered by the Department of State (Fergusson and Kerr, 2013).

States also coordinated internationally to agree to export controls and standards for managing nuclear materials to mitigate risks from proliferation. For example, the Nuclear Suppliers Group (NSG) was formed in 1974 in response to an Indian nuclear test. The NSG standardized export policies on non-weapons nuclear materials and equipment (Burr, 2014). Other multilateral export control regimes for nuclear materials and other weapons of mass destruction include the Missile Technology Control Regime and Wassenaar Arrangement (Fergusson and Kerr, 2013). International agreements are also used to legally bind non-nuclear states to standards; examples include the Non-Proliferation Treaty (NPT) and the safeguards agreements with the International Atomic Energy Agency (IAEA). Notably, the NPT involved a bargain: non-nuclear weapon states agreed not to seek nuclear weapons in return for access to nuclear energy technology for peaceful use.

Cooperation between leading actors has been important in nuclear non-proliferation. Despite the competitive pressures, leading powers (historically the US and Soviet Union) jointly recognized the dangers of nuclear weapons and the unique dangers of allowing new actors to develop nuclear weapons. As a result, leading powers enacted measures that specifically targeted non-leading actors (Colgan and Miller, 2019). An example is the NPT, which instituted legally-binding measures to prevent actors not already possessing nuclear weapons from obtaining them.

The NPT notably promoted the peaceful use of nuclear technology by non-nuclear powers, and these provisions were likely necessary for it to gain widespread agreement. The technology needed for harmful usage had to be clearly distinguishable from the technology needed for peaceful usage. Insofar as this was possible, the leading actors agreed to assist others in using nuclear technology for peaceful purposes. For the AI context, we may learn that restricting access to dangerous information (e.g. model weights) may be more acceptable if leading labs commit to sharing information that is unambiguously safe (e.g., sharing prompting or fine-tuning access with safety researchers).

Export controls can control the spread of physical materials. However, information that can spread digitally is more difficult to control, especially when many actors already have access to information. For example, some equipment to enrich uranium in a centrifuge can be 3D-printed, and it is more difficult to prevent the spread of computer-aided design (CAD) files than the materials themselves (Christopher, 2015). While laws may prohibit sharing such files online, after an illegal posting, the file can be shared and downloaded many times before it is being removed. As one example, information about a handgun called “The Liberator” was illicitly shared in 2013 (Christopher, 2015).

Digital or otherwise “intangible” materials are often included under export controls if they are used to produce controlled technologies (Stewart, 2015). The Wassenaar Arrangement defines a list of various intangibles that are controlled (e.g., diagrams and specifications, technical data, and instruction or consulting; Stewart, 2015). In addition, intangible technologies are generally exempt from export control regimes if they relate to “basic scientific research” or are in the “public domain” (European Union, 2021). Malicious actors may not have the skills necessary to develop dangerous technologies based on public information, so such exemptions from controls on technical assistance may increase risks. Furthermore, industry actors may be incentivized to place information in the public domain or share it scientifically in order to become exempt from export controls for dual-use technologies they develop.

Export controls can also be difficult to enforce. In the US, the Bureau of Industry and Security (BIS) is responsible for reviewing and approving export licenses and enforcing export controls on-site. However, BIS administration of export controls is presently underfunded and lacks integration with modern technologies that would allow for more effective enforcement (Allen et al., 2022). Actors can engage in theft or smuggling of controlled items, or they can establish shell companies to reduce suspicion of ties and have licenses for exports approved. Rules such as the Foreign Direct Product Rule, which establishes that products produced outside the US with American equipment or inputs are subject to export controls, make export controls stronger on paper but are also more difficult to enforce (Allen et al., 2022).

3.4 AI and Machine Learning

In machine learning research, publication norms and information-sharing practices differ between academia and industry. ML academia has relatively strong norms toward diffusion. Some universities have initiatives designed to promote open-source efforts (e.g., UC Berkeley's Berkeley AI Research Open Commons program). Corporate groups often have stricter internal information-sharing policies. For example, Apple has been known to go to great lengths to ensure that its products are kept secure. They've been known to develop code names for projects, strict NDAs, and large fines (up to $50M) for breaking confidentiality agreements (Gordon, 2020). Google holds background checks and yearly security trainings for all employees. Furthermore, all employees have NDAs, and some have been fired for unintentionally or intentionally leaking information. Leaked information often provides information that can reduce the time it takes for competitors to build model replicas.

Recently, ethics boards and risk assessments have become more common in AI and ML and research. For example, In 2021, the Stanford Institute for Human-Centered Artificial Intelligence required that all research projects being funded must gain ethics board approval before starting. NeurIPS recently implemented the NeurIPS paper checklist, which "prompts authors to reflect on the potential negative societal impact of their work" (Beygelzimer et al., 2021). NeurIPS also established an Ethics Review Board to evaluate the risks of submitted research. Furthermore, ML scholars developed a checklist designed to help academics analyze the impact of their work on long-term catastrophic risks from advanced AI systems (Hendrycks and Mazeika, 2022).

AI labs vary in their information-sharing policies. For example, one AI lab has an internal information-sharing policy that follows a "need to know" principle: staff are not shared on information unless it is relevant to their work. Information is classified as secret (only shareable with specific individuals), private (only shareable with a broadly defined group), and public (shareable with everyone; Leahy et al., 2022). To our knowledge, other major AGI labs have not disclosed similar policies. Furthermore, when making model-sharing decisions, labs often employ a structured access approach: they provide limited access to models and control the ways in which users can engage with the models. The purpose is to allow access to AI capabilities that could be useful while minimizing access to capabilities that could be dangerous (Shevlane, 2022).

For the release of GPT-2, OpenAI employed a staged release policy: they started by releasing a small version, intentionally withholding larger models due to concerns about misuse. Over the course of 9 months, they gradually released larger models. The purpose was to have time to see the effects of smaller models and update risk assessments accordingly (Solaiman et al., 2019). More recently, for the release of GPT-4, OpenAI intentionally decided not to disclose certain details about the model: "Given both the competitive landscape and the safety implications... this

report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar" (OpenAI, 2023).

4 Proposals: Adapting information security ideas for AGI labs

Informed by our literature review, we describe three proposals that could be implemented to improve publication and model-sharing decisions.

Our proposals are meant to complement previous efforts to improve information-sharing policies at AGI labs. Examples include the structured access paradigm (Shevlane, 2022), staged release strategies (Solaiman et al., 2019), and calls for strong information security at AGI labs (Ladish and Heim, 2022). While we do not focus on these proposals (as they have been described in previous papers), we commend the authors who designed them and the AGI labs who have implemented them.

Each of our proposals can either be implemented in various ways by various actors. They could be implemented by an individual lab, a group of AI labs, external third parties that oversee industry-wide regulation, or binding government regulation. For simplicity, we write about each proposal as if it were being implemented by an individual lab. However, we would be excited to see ambitious efforts to implement promising policies on a wider scale.

There are a few principles to keep in mind when reviewing these proposals. First, we do not believe these proposals are panaceas, and they are meant to *supplement and strengthen* strong cultures that prioritize safety. Second, and relatedly, there are ways to implement each of these proposals poorly, in ways that do not actually reduce catastrophic risks. Each of these proposals is "gameable", and the proposals rely (in part) on labs making genuine commitments to prioritize the reduction of catastrophic risks. Third, each of the proposals could be implemented in various ways. In the following section, we intend to describe plausible versions of each proposal, but there are many possible alterations that can be explored and addressed in future work.

Fourth, and relatedly, we believe that each proposal would ideally be adapted for specific AGI labs. There are several differences between labs (e.g., organizational structure and lab culture, which teams handle concerns around safety and how they interact, how catastrophic risks are currently managed, the amount of decision-making power that members of safety teams possess differ by labs). Ideally, these differences would be factored in when proposals are implemented.

Finally, it is important to consider catastrophic risks as early as possible in the research cycle. For example, researchers could be encouraged to submit a brief proposal to the Catastrophic Risk Review Board *before* starting a research project or writing a paper. Similarly, researchers could be encouraged to consult risk assessment guidelines *before* starting a research project and *while* writing a paper. After a paper has been written or a model has been built, it may be more difficult (or more costly) to apply information-sharing policies. As a result, we encourage AGI labs to consider how such policies could be integrated into the early and middle stages of the research process.

With these considerations in mind, we believe that our proposals offer useful and tangible ways for AGI labs to reduce catastrophic risks If implemented widely as industry standards, we believe these proposals could supplement and reinforce safety-focused cultures and ultimately reduce the likelihood of catastrophic risks.

Table 2: Recommended AI risk interventions, explanations, and inspirations

Recommended intervention	**Explanation**	**Inspiration from other fields**
Catastrophic Risk Review Board (for publication and model-sharing decisions)	A review board (consisting of individuals with expertise in catastrophic risk reduction) provides recommendations on major publication and model-sharing decisions.	Biological sciences: NSAAB publication review of dual-use research (Casadevall et al., 2014) Cybersecurity: Information-sharing decisions made by ISACs (Ezhei and Tork Ladani, 2017) Nuclear technologies: "Born secret" doctrine (Morland, 2004)
Catastrophic Risk Questionnaire (for publication and model-sharing decisions)	Before major publication and model-sharing decisions, researchers and decision-makers fill out a questionnaire that asks them to consider various ways the information-sharing decision could affect catastrophic risks.	AI & machine learning: AI risk checklist for AI researchers (Hendrycks and Mazeika, 2022) Biological sciences: NSAAB guidelines for dual-use bioresearch (National Science Advisory Board for Biosecurity, 2016, 2023) Checklists and guidelines are also used in other fields (medical, behavioral safety)
Responsive Release Paradigm	Researchers time-stamp their research projects (any outputs, training details, results, etc.) in a private Responsive Release Repository. They release their findings after a set amount of time or after a different team publishes similar research.	Nuclear technologies: "Born secret" doctrine (Morland, 2004) Interdisciplinary: Trusted timestamping methods used in a variety of fields

4.1 Catastrophic risk review boards (inspired by publication review policies in the biological sciences (McLeish and Nightingale, 2007; Malakoff, 2013).

Proposal: For publication decisions, model-sharing decisions and other information-sharing decisions that could be considered dual-use, leadership consults the recommendation of a **catastrophic risk review board**[2]. This board consists of at least one member from a lab's technical safety team, governance team, and security team[3]. Before publishing a

[2]Many industry labs already have an approval process for publications. In the event that a lab already has a process, our recommendation can be incorporated into their existing process (i.e., instead of setting up a new safety board, the existing process can be modified to include the input of a member of the lab's technical safety team, governance team, and security team).

[3]The optimal makeup of the safety board will ultimately depend on a particular lab and its unique context. We believe that our concrete recommendation (having at least one member from a technical team, governance team, and security team) is likely to be feasible for most of the current leading AGI labs. However, we encourage members of labs to think about alternative implementations.

paper or sharing a model, this panel has at least X weeks to review the decision, write a recommendation, and discuss their recommendation with leadership.

We propose that all publication decisions and model-sharing decisions are reviewed by a body consisting of one member from each of these safety teams. Consider a case in which a research team wants to publish paper A. Before publishing, the paper is sent to a group of 5 individuals (from the lab safety teams) who have volunteered to be part of the review process. First, they read the abstract, skim relevant sections of the paper, and determine whether or not the publication is exempt from the review process (in many cases, we expect it to be clear from the abstract that the paper does not need to be reviewed[4]). If they do decide to review the paper, then they would (a) read the paper, (b) consider the short-term and long-term consequences of sharing the information, and (c) write a recommendation sharing their thoughts in a report to leadership.

Catastrophic Risk Review Boards can differ in many ways. We believe that one strength of the proposal is that it can be flexibly applied; labs have some degree of freedom regarding what projects are reviewed by the board, how decisions are made, who is on the board, and a few other considerations.

The review board can choose to recommend publication, recommend against publication, or recommend partial/modified publication (e.g., recommend publication if certain details about the training process are redacted). When making decisions, members of the board would think about potential effects on structural catastrophic risk factors (e.g., race dynamics), safety research, and the company. The board members also consider alternatives to full disclosure (for more details about what this could involve, see the "catastrophic risk questionnaire" proposal).

Several of the major AGI labs have safety teams that work on reducing short-term and long-term risks from AI systems. More specifically, each of the three leading labs (OpenAI, DeepMind, and Anthropic) currently have at least one team working on technical problems (often called the "technical alignment team"), at least one team working on long-term AI governance concerns (often called the "governance team"), and at least one team working on security/cybersecurity risks (often called the "security team"). Individuals on these teams tend to think about potential risks associated with AI systems and generally have competence in various technical and non-technical threats.

Catastrophic risk reviews could ensure that individuals with expertise in safety-relevant domains have a chance to offer input into the information-sharing process. The review board could have a direct effect on improving lab information security by identifying and preventing the spread of sensitive, dangerous, or dual-use information. We also expect some positive indirect effects. For example, catastrophic risk reviews would encourage greater interaction between

Ultimately, it will be important for members of the safety board to have expertise in safety-relevant areas and possess a strong security mindset, but it may not be necessary for them to come from these three particular teams.

[4]In theory, this review board could be overwhelmed with submissions. However, based on current publication rates in AI labs, we don't expect the situation to be overwhelming. For example, OpenAI's research index (https://openai.com/research) lists 18 publications from 2022, and many of these would not be relevant from a catastrophic risk perspective. DeepMind, which is much larger than OpenAI, published papers at a higher rate, though a small fraction of these papers is directly relevant for concerns AGI and catastrophic risks. At the current rate of publishing, we expect that Catastrophic Risk Review Boards would likely only choose to send 3-10 major publication decisions and model-sharing decisions to the full review stage. Furthermore, recall that members of the Catastrophic Risk Review Board have the option to select how many decisions go to full review. Given this, we expect this proposal would not be overly burdensome on Catastrophic Risk Review Board members, especially given the current (manageable) rate of publication decisions and model-sharing decisions.

Table 3: Overview of Catastrophic Risk Review Boards

	Area	Details	Our recommendations
1	What projects are evaluated?	Does the board review publications, models, products, strategy documents, or other projects?	We recommend that the board consider and review the most important information-sharing decisions, regardless of the kind of project.
2	What decisions are made?	Does the board have the ability to make binding decisions about information-sharing decisions, or does it only have the ability to make recommendations?	This can be determined by individual AGI labs.
3	When do researchers interact with the board? At what point in the project life cycle is the board contacted?	When is the Catastrophic Risk Review Board active? (e.g., idea generation, project prioritization after the first draft has been written, when a project has concluded)	Researchers are encouraged to reach out to the board early in the research life cycle (e.g., as a project is being conceptualized or before a paper is written; *particularly before large training runs*). Development teams can work with the board to discuss project prioritization decisions and inform the direction of research *before* a publication decision has to be made.
4	How broad is the scope of the board?	Is the board industry-wide, or does one exist for every AGI lab?	For now, each AGI lab implements its own board. In the future, a third-party external organization could manage the review board.
5	Who is on the board?	Does the board consist of employees at the lab or external parties?	At least one member from a technical safety team, long-term governance team, and security team is on the board. However, they have the resources to involve external people when relevant.
6	How are the decisions made?	How flexible is the board? Are there clear questionnaires?	The board considers the impact of release decisions on catastrophic risks and corporate interests. The board uses the catastrophic risk questionnaire (see subsection) and other internal decision tools.
7	How is the board championed and sustained?	How can incentives be established within the company or industry such that the boards are able to accomplish their goals? How can labs prevent the boards from being "watered down" or dismissed?	Leaders of AGI labs verbally and behaviorally express support for the board. Board decisions are taken seriously, and disagreements are openly discussed. Board members feel like they have buy-in and support from lab leadership, and there are clear ways for board members to express feedback to lab leadership.

researchers of lab safety teams and researchers on other AI research teams. If AI researchers know that members of safety teams may review their papers, this may cause AI researchers across the lab to learn more about the concerns of members of safety teams, supporting a stronger culture of caution and security.

There are also some potential disadvantages to the board that warrant attention. For instance, catastrophic risk reviews could slow down publication and waste the time of safety researchers. These costs can be mitigated if the Catastrophic Risk Review Board members are able to decline to review certain publications (e.g., after reading the abstract and skimming the paper) and if they are expected to review the publication within a certain timeframe of receiving it (e.g., 2 weeks). Another possible disadvantage is that members of the safety review team could be penalized for their decisions. For instance, suppose a member of the review team is aware that lab leaders are excited to publish a paper or share a model. They might (correctly or incorrectly) reason that if they recommend against sharing, they will face retaliation.

Importantly, Catastrophic Risk Review Boards seem compatible with existing norms among AI labs. For example, DeepMind has an Institutional Review Committee (IRC) that meets every two weeks to discuss projects, papers, and collaborations (Kavukcuoglu et al., 2022). Other companies also have ethics boards or review boards that meet periodically to discuss important decisions. Furthermore, researchers have recently called for review boards that inform release decisions for foundation models (Liang et al., 2022).

The key thing that makes Catastrophic Risk Review Boards unique is their focus on *long-term catastrophic risks*. Reasoning about catastrophic risks (rather than more short-term risks can require different considerations, expertise, and processes. Considerations around longer-term risks are often easy to overlook as they are harder to quantify or measure. Hence, the Catastrophic Risk Review Boards could usefully complement AI labs' existing procedures. The success of the boards will be determined, in part, by the people on the boards. We encourage labs to ensure that the board consists of individuals who have a strong understanding of possible catastrophic risks from AI systems and display a strong commitment to preventing such risks. We expect that members of technical alignment teams, governance teams, and security teams are more likely to possess such characteristics.

4.2 Catastrophic risk questionnaires (inspired by checklists in aviation, health care, and ML/AI (Hendrycks and Mazeika, 2022; Thomassen et al., 2014; Treadwell et al., 2014; NeurIPS, 2021).

Proposal: Before major information-sharing decisions, a team of researchers fills out a catastrophic risk questionnaire (see Appendix A). The questionnaire asks researchers to describe potential dual-use considerations, effects on long-term catastrophic risks, effects on race dynamics, alternatives to full disclosure, and general risks and benefits from sharing the information. The questionnaire responses are then reviewed by lab leadership (and potentially a Catastrophic Risk Review Board).

In aviation and health care, the implementation of simple checklists and questionnaires can lead to improved decision-making and reduce the rate of accidents. Safety checklists originated in aviation, and principles from aviation checklists were adapted for use in surgical settings (Weiser et al., 2010). Checklists and questionnaires are thought to work by simplifying decision-making procedures, empowering safety-conscious individuals, ensuring that critical considerations are not overlooked, promoting a standardized decision-making procedure, empowering safety-conscious individuals to speak up, and offering reminders (Bosk et al., 2009; Treadwell et al., 2014). They appear to be effective in various healthcare contexts; review papers have shown that they can increase the detection of safety risks, improve communication among staff, and decrease fatal complications (Thomassen et al., 2014; Treadwell et al., 2014). Note

that in some fields, checklists involve following a set of clear and well-defined steps; navigating risks from AGI may be more complicated and less straightforward. It is also worth noting that the effectiveness of checklists and questionnaires depends on culture; they appear to be most effective when accompanied by cultures that value performance standards, promote strong relationships between safety-conscious individuals and leadership, and generally have a security-focused mindset (Bosk et al., 2009).

AGI labs could deploy questionnaires to ensure that key considerations are examined before major information-sharing decisions. As mentioned in the literature review section, checklists and questionnaires have recently been implemented in ML/AI contexts: NeurIPS released a checklist about societal impacts, and ML researchers recently developed a questionnaire to help scholars consider long-term existential threats from AI systems in their research (NeurIPS, 2021; Hendrycks and Mazeika, 2022). One advantage of checklists is that the items on the checklist do not need to be particularly complex or counterintuitive; checklists can be effective simply by standardizing "common sense" considerations and ensuring that all members of a team are able to participate in the risk assessment process. Thus, checklists can be helpful even when their items include factors that AGI labs already seek to consider. Checklists and questionnaires can also synergize with other recommendations. For example, if a lab has a safety board reviewing publications, members of the safety board could consult the checklist when making decisions. Alternatively, researchers submitting to the safety board could include responses to the checklist in their proposal to the safety board.

On the other hand, questionnaires can lead to a misleading sense of safety. While writing about potential risks and adverse effects is a useful first step, we expect questionnaires to be most impactful when they meaningfully inform the actions that researchers take. If researchers fill out the questionnaire as a formality, but they don't actually consider taking different actions, the impact of the questionnaire is substantially reduced. Moreover, post-hoc rationalization may reduce the quality of such assessment reports.

In Appendix A, we present a Catastrophic Risk Questionnaire for AGI Labs. The questionnaire is intended to focus on information-sharing decisions that relate to AI progress, AI models, and cutting-edge research findings.

4.3 Responsive release paradigm (inspired by trusted timestamping procedures used historically in physics research).

Proposal: When a potentially dual-use finding is discovered, the scientists write a paper describing the finding. However, by default, they do not publish it publicly. Instead, they timestamp the paper, store it in an encrypted private repository, and include a brief "security statement" that describes why they decided not to release the paper publicly. The paper remains in the repository until a different team has published the same finding (or a sufficiently-similar finding). Then, the article is published, along with the security statement. If no sufficiently-similar finding is published within a given amount of time, the researchers have the option to publish proactively.

When Robert Hooke discovered Hooke's law, he wanted to establish that he was the first to discover it, but he didn't want to publish his results immediately. As a result, he published an anagram: *ceiiinosssttuv* (Latin for "*Ut Pondus sic Tensio*", translated as "as the extension, so the weight"). This provided an effective timestamp for Hooke's discovery; he could ensure that he was properly credited for his discovery without publishing his results. In modern times, researchers do not need to rely on anagrams; they can rely on digital timestamping techniques.

One of the benefits of open publication cultures is that they allow scientists to be properly credited for their discoveries. For individual scientists, publications can influence hiring decisions, salary negotiations, and their broader reputation in the field. For AI labs, having favorable publication policies can be useful for attracting talented researchers. At the same time, as mentioned earlier, findings relating to AGI are often dual-use, and open publication norms can increase risks.

The responsive release repository is designed to balance these interests: scientists have a way to take credit for their discoveries, and the rate at which dual-use information is shared is reduced. As mentioned earlier, findings relating to AGI are often dual-use, and open publication norms can increase risks. By limiting or delaying the amount of publicly-available information about AI capabilities, the responsive release repository could reduce the likelihood that incautious actors have access to state-of-the-art capabilities (reducing misuse risks) and increase the lead time of leading labs (reducing race dynamics). Furthermore, by including a security statement along with the publication, scientists can express their commitment to safety, strengthening a norm against the irresponsible disclosure of dual-use information. In the event that a team publishes a finding and another team responds with a responsive release statement, this could lead to a productive dialogue about the information that was shared, whether it should have been shared, and the potential risks from sharing such information. If implemented properly, the responsive release approach could promote a culture of safety, caution, and discourse around dual-use information.

Concretely, responsive release repositories are most useful for information (e.g., training results, algorithmic insights, datasets, or model weights) that most contribute to race dynamics or proliferation. We expect that this will include algorithmic insights (e.g. new architectures) and high-level knowledge about progress, such as the Chinchilla scaling laws (Hoffmann et al., 2022).

Regarding implementation details, scientific results can be stored in repositories with only marginal risks of leaks or espionage. Time-stamped private repositories already exist in other scientific fields. For instance, social scientists can time-stamp their RCT preregistration, the public only gets access to the preregistration files once the data collection and evaluation process is finished and the researchers decide to publish the hidden information (American Economic Association, 2023). To improve information security, timestamps could be stored on the blockchain, and the repository could apply state-of-the-art encryption tools (Gipp et al., 2015; Russell, 2001). It would be especially important to ensure that the repository was protected against potential adversaries.

The proposal also comes with some disadvantages. First, the responsive release repository does not fully allow a scientist to retain credit for the discovery. If two papers are published consecutively, it is likely that citations (and media coverage) will be split between the two papers. Second, there is some subjectivity in evaluating when a "sufficiently-similar" paper has been published. Individual scientists, lab teams, or Catastrophic Risk Review Boards would need to evaluate this on a case-by-case basis. Third, this practice could reduce the rate at which *safety-relevant* information is published. To mitigate this risk, labs could have an appeal process; if a research team can successfully argue that the benefits of publishing (e.g., to safety research) outweigh the costs (e.g., to race dynamics or potential misuse), then they can receive an exemption from the responsive release policy. If the lab has a Catastrophic Risk Review Board, this determination could be made by the review board.

If adopted, the responsive release repository might also impact the overall discourse around AGI risks and the culture of information-sharing. For example, labs that adopt this policy might publish blog posts (or a joint statement) about why they are adopting the policy, raising awareness about dual-use concerns relating to AGI. As in other fields, we expect

that some groups would support the policy (recognizing the ways in which it curtails access to potentially-dangerous information reduces race dynamics) and other would criticize it (potentially arguing that it's unnecessary and that labs are overestimating risks or underestimating benefits from diffusion). On balance, we expect that the adoption of the policy would lead to more discussion and discourse around risks from AGI, and we see this as a beneficial side effect. Just as open discourse about dual-use publication norms seems to have been beneficial in other fields by stimulating discourse about the risks and benefits of information-sharing (e.g., Malakoff, 2013), we expect that greater discourse around information-sharing to be beneficial for the field of AGI development.

5 Future research

We divide our recommendations for future research into two subsections. First, we discuss other proposals that could be investigated. Second, we outline other directions for future research.

5.1 Proposals to explore in future research

Internal information-sharing policies. Future research could examine internal information-sharing policies: how information is shared *within* a given AI lab. Rules and norms could prevent unintentional information leaks. Simply put, if everyone at a company knows a (potentially dangerous) piece of information, the odds of that information leaking to outside groups increase considerably. Consider a case in which each employee has a 99.5% chance of successfully keeping a piece of information secret from outsiders (and these probabilities are independent). If 100 people know this information, the chance of a leak is 60%; if 500 people know this information, the chance of a leak is 92%. Some information leaks can contribute to structural risk factors (e.g., someone from Company A accidentally leaks dual-use information to someone at Company B, accelerating company B's progress and accelerating race dynamics), misuse risks (e.g., a state actor launches cyberattacks against individuals at company A and steals information that help them build powerful AI systems), or accident risks (someone from company A leaks information to someone at an incautious lab, increasing the risk of the incautious actor developing unaligned AI). Future research could explore these plausible failure modes and threat models, understand processes that AI labs currently use to prevent unintentional information leakages, investigate possible disadvantages of internal information-sharing policies, and attempt to design reasonable internal information-sharing policies that can reduce the number of people who have access to dual-use information. Such research could draw from research on security clearance systems in industry and government (Janczewski and Portougal, 2000, 2008), as well as some AI labs (see Conjecture's internal infohazard policy; Leahy et al. 2022).

Coordination across labs to share information relevant to the mitigation of catastrophic risks. Consider a case in which Lab A discovers that a model is exhibiting dangerous capabilities or misaligned tendencies in certain environments. Sharing this information widely could be dangerous, but sharing the information with a select group of vetted trusted safety-conscious actors could be beneficial (e.g., by coordinating to slow down or collaborate on research to address the concerns in Lab A's model). Future research could describe how such collaborations could occur legally and responsibly, how to decide which actors should be included in such collaborations, what kind of information ought to be shared, and what ought to be done when evidence of catastrophic risks is presented. Such work could draw from CERTs from computer security and ISACs/ISAOs that are used in aviation, health care, emergency services sectors, and various other industries (Choucri, 2016).

Anonymous risk reporting. Social pressures can reduce the likelihood that individuals report incidents or potential risks, especially risks from other members of an organization or risks that are speculative in nature. This is especially

Table 4: Future areas of research

Proposed intervention	**Explanation**	**Inspiration**
Internal information sharing	Changing how information is shared *within* a given AI lab can prevent unintentional information leaks.	1. Military and other industries adopted the "need to know" principles and information compartmentalisation. 2. First attempts of such policies within existing AGI organizations (Leahy et al., 2022)
Lab Coordination to Share Risks and Safety Insights	Information about dangerous capabilities or misaligned tendencies perhaps should not be shared publicly but with other AGI labs to improve their systems. Foras should exist where AGI labs can exchange safety-critical information.	1. ISACs are organizations that facilitate information sharing and cooperation in cybersecurity and aviation safety (Clinton, 1998) 2. There are also technical efforts to share cybersecurity threats effectively among organizations without revealing dual-use information and avoiding leaks
Effective Incidence Sharing	AGI Labs creates mechanisms that help individuals share plausible new threats, safety risks or plausible black swan events.	1. Successful whistleblowing policies, e.g., in the finance and pharmaceutical industry (Dasgupta and Kesharwani, 2010) 2. Anonymous reporting to the National UFO Center (National UFO Reporting Center, 2023)
Trainings and role-play	Training, e.g., role-play, can help members of AI labs make more informed information-sharing decisions.	1. The CIA has used tabletop games to help analysts improve their forecasting and decision-making (Larson, 2017). 2. Partnership in AI (PAI) hosted an exercise on publication decisions with members of the AI community (Leibowicz et al., 2019). 3. Security training at tech companies, e.g., Google

concerning given that some risks from AGI systems are often unintuitive or speculative (e.g., An AI system becoming situationally aware and copying its weights to another server; Shlegeris 2022). In order to promote risk reporting despite the social and psychological pressures against such information-sharing, AGI labs could implement (anonymous) reporting systems. Individuals could be encouraged to report any kind of concern: misconduct or risky behavior from colleagues, suspicious findings relating to AI systems (e.g. new capabilities, situational awareness), violations of existing safety regulations, concerns about the potential sentience of AI systems, and other speculative concerns. Future research could examine how such systems could be implemented at AI labs and how responses should be reviewed. Furthermore, future research could examine if an anonymous risk reporting system could be implemented at an industry-wide level, especially for cases when an individual believes their company is not taking the risk seriously.

Such work could draw from whistleblower protection systems (Dasgupta and Kesharwani, 2010) and examples of anonymous reporting systems designed to normalize the reporting of "weird" or "embarrassing" information (e.g., UFO reporting systems; National UFO Reporting Center 2023)

Trainings and role-plays. Future research could explore trainings in which members of AI labs reason through information-sharing decisions and learn techniques to help them make informed decisions. To simulate real-world decisions, such trainings could involve active exercises and role-play scenarios. As an example, in 2019, Partnership in AI (PAI) hosted a dinner with members of the AI community and ran a simulated exercise. Participants imagined that they were part of a review panel considering whether to publish a hypothetical paper (along with code, training data, and a neural network model). This activity allowed participants to practice generating and debating various considerations that occur when making challenging information-sharing decisions (Leibowicz et al., 2019). A team of researchers also offers role-plays to explore different AI futures (Intelligence Rising). Similar activities often occur in other areas where individuals have to make high-stakes decisions. For example, role-play scenarios are common in international relations (Shaw, 2004), and the CIA has used tabletop games to help analysts improve their forecasting and decision-making (Larson, 2017). Future research could examine the kinds of features that make role-play scenarios and serious gaming exercises most effective and develop new scenarios for AI labs. Such exercises, if designed well, could be incorporated into regularly-occurring training exercises or discussion groups for decision-makers at AI labs. More broadly, future research could explore what other kinds of information or exercises might be included in trainings about responsible information-sharing practices: technology foresight exercises, scenario planning, role-play, and worst-case contingency planning.

5.2 Other areas of future research

Future research can also focus on studying best practices in other sectors or provide other information to further inform the generation, refinement and prioritization of information-sharing policy proposals for the AGI industry.

Research to help decision-makers reason about cost-benefit analyses from information-sharing decisions. Future research will be needed to help individuals in groups make more informed decisions about when, how, and with whom to share information. Decision-makers at AI labs are tasked with making difficult cost-benefit analyses with respect to information-sharing decisions. Future work could attempt to help make these cost-benefit analyses more robust by identifying the kind of information that is most dangerous to share, estimating the impact of specific publications or models on AI progress and race dynamics, and examining case studies of high-stakes information-sharing decisions in AI (e.g., a case study that estimates the effects of releasing ChatGPT or the Chinchilla scaling laws).

Identify best practices that relate to domains other than information-sharing policies. For example, future research could examine ways of reducing accidents in high-stakes contexts, methods to reduce structural risk factors in fields where race dynamics are common, or ways of promoting security-focused mindsets in fast-moving organizations. We believe our methodology (reviewing policies across a range of relevant fields, identifying best practices, and considering how those practices could be adapted or implemented in AI labs) could be repeated across a variety of topics, ultimately generating a wider list of concrete proposals to reduce catastrophic risks. This methodology could also be applied to identify recommendations for different stakeholders. While our work focused on suggestions that AI labs could implement, future research could employ similar methods to identify proposals that other actors (e.g., governments, policymakers, hardware companies) could employ in order to reduce catastrophic risks from AI.

Explore other ways the information-sharing platform may be altered. Patent rules and export controls may change what information is owned by whom, who can (legally) replicate what, market monopolization and the proliferation of AGI firms. Future research could study possibilities for patenting AGI-related insights and how export control rules related to AGI may influence what information labs will share and how.

6 Conclusion

We expect AGI labs will continue to face important and challenging decisions about what information to share and with whom. To inform these decisions, AGI labs could adopt policies and norms based on best practices from other fields. Here, we presented three proposals that we believe could help AGI labs manage catastrophic risks from information-sharing decisions. While no single policy will guarantee that such decisions are made well, we believe that each of the proposals has the potential to meaningfully improve or codify lab decision-making procedures.

Catastrophic risk management will require more than well-informed information-decision policies. Ideally, best practices from other fields could be identified and adapted across a wider range of areas (e.g., risk management, accident prevention, institutional decision-making, emergency planning). Notably, however, AGI may also present risks that are more dangerous, more sudden, and more difficult-to-detect than other disciplines. A thorough catastrophic risk management approach would involve a balance between importing useful policies from other fields and creating new ones specifically tailored for AGI risks.

We described our proposals as if they would be implemented by AGI labs, but future work could also explore how these proposals could be implemented by coalitions (e.g., an interlab Catastrophic Risk Review Board) and policymakers. Ultimately, we believe that the most responsible catastrophic risk management approaches will involve a mixture of lab-specific policies, industry-wide standards, national regulations, and international agreements. For now, we hope that AGI labs find this research useful as they revise and augment their information-sharing decisions.

7 Acknowledgements

We are grateful to AI researchers, subject-area experts, and others for helpful feedback on this paper. We would like to thank Darius Meissner, David Manheim, Florian Magin, Gregory Lewis, Jonas Sandbrink, Joshua Monrad, Jide Alaga, Mauricio Baker, Marina Favaro, Markus Anderljung, Nicole Wheeler, Shaun Ee, Toby Shevlane, and Rosie Campbell. We are especially grateful to Daniel Kokotajlo and Jeffrey Ladish for their support and advice throughout the project.

A Catastrophic Risk Questionnaire

A.1 High-level considerations

1. **Dual-use considerations:** What are some potential risks or harmful consequences of this work, its release, or the potential attention it receives? Are there any ways to mitigate these risks?

2. **Consideration of catastrophic risks.** How might our decision (intentionally or unintentionally) increase or decrease the likelihood of catastrophic risks from advanced AI systems?

   2.2a Suppose the release of this information or model caused large amounts of harm, or even a catastrophe. What might have happened? If there are stories that are at least somewhat plausible, alternative release strategies should be considered.

2.2b We may define a catastrophe as and AI incident with 1000 dead or $1bn in economic losses.

A.2 Race dynamics and Industry Effects

3. **Acceleration.** What is the expected impact of this work on AI acceleration and race dynamics? Does the work accelerate the race toward powerful AI systems?

   3.3a If a different group released this work (and we didn't have this work), would this change our thinking, make it easier for us to build AGI, or put pressure on us to move more quickly? If yes, consider if there are ways to release the information in ways that mitigate these risks (e.g., Shevlane, 2022).

4. **Public reaction:** What is the expected impact of this work on the public? Could this contribute to unjustified hype or incautious framings of AI?

   4.4a Are there ways to frame the work in ways that cause the public to decrease hype and emphasize caution? What kinds of cultural memes or unintended messages that might spread?

5. **Industry Norms:** If other groups were to follow the norm of releasing this kind of work, what would the effect be on the overall AI ecosystem?

6. **Effects on incautious actors:** How would sharing this work affect other AI labs? To what extent would releasing or deploying this work benefit incautious actors?

A.3 Consideration of alternatives

7. **Intended benefits of disclosure:** What are the main intended benefits of sharing this work?

8. **Alternatives to full disclosure:** To what extent could the benefits of sharing this work be achieved by alternatives to full disclosure (e.g., structured access (Shevlane, 2022), staged release (Solaiman et al., 2019), release to selected AI safety researchers)? Are there any new release strategies that might be useful for this situation?

   8.8a Are there any specific pieces of information (results, techniques or insights) that would be useful to share with specific actors in order to reduce risks from advanced AI systems? How could you share the information with the appropriate parties (while minimizing the chance that the information leaks further)?

A.4 Procedural Questions

9. **External review and internal red-teaming:** Have we considered consulting external parties (e.g., AI safety researchers) to discuss the pros and cons of sharing this work? Have we internally red-teamed this publication strategy? What were the results?

10. **Miscellaneous:** Are there any other considerations or uncertainties about this work and the form of publication that are worth mentioning?

A.5 Proposal

What do you propose to release, when, how, why, and to whom?

What:

To Whom:

When:

How:

Why:

A.6 Release Strategy

Table 5: Release Strategy Questions

Questions		Response
Who?	What group gets the information?	
What?	What kind of information?	
Where and how?	How is it provided to them?	
Why	Why do they get the information?	
When?	When do they get the information?	
How to prevent leaks?	What procedures, technical solutions, NDAs, monitoring procedures have been taken to avoid that more information is being shared (to a bigger group)?	
How to monitor?	How will the sharing, the benefits, the information security be monitored?	

References

Allen, Gregory C., Benson, Emily, and Reinsch, William Alan. "Improved Export Controls Enforcement Technology Needed for U.S. National Security." Center for Strategic & International Studies. 2022.
URL https://www.csis.org/analysis/improved-export-controls-enforcement-technology-needed-us-national-security

American Economic Association. "AEA RCT Registry." 2023.
URL https://www.socialscienceregistry.org/site/instructions

Anthropic. "Core Views on AI Safety: When, Why, What, and How." 2023.
URL https://www.anthropic.com/index/core-views-on-ai-safety

Bai, (Max) Hui, Voelkel, Jan G., Eichstaedt, Johannes C., and Willer, Robb. "Artificial Intelligence Can Persuade Humans on Political Issues." OSF Preprints. 2023.
URL https://osf.io/stakv/

Barash, Jason R. and Arnon, Stephen S. "A novel strain of Clostridium botulinum that produces type B and type H botulinum toxins." *The Journal of Infectious Diseases* 209.2 (2014): 183–191.

Barrett, Anthony M., Hendrycks, Dan, Newman, Jessica, and Nonnecke, Brandie. "Actionable Guidance for High-Consequence AI Risk Management: Towards Standards Addressing AI Catastrophic Risks." arXiv. 2023.

ArXiv:2206.08966 [cs].
URL http://arxiv.org/abs/2206.08966

Beygelzimer, Alina, Dauphin, Yann, Liang, Percy, and Wortman Vaughan, Jennifer. "Introducing the NeurIPS 2021 Paper Checklist – NeurIPS Blog." 2021.
URL https://blog.neurips.cc/2021/03/26/introducing-the-neurips-2021-paper-checklist/

Bosk, Charles L., Dixon-Woods, Mary, Goeschel, Christine A., and Pronovost, Peter J. "Reality check for checklists." *The Lancet* 374.9688 (2009): 444–445. Publisher: Elsevier.
URL https://www.thelancet.com/journals/lancet/article/PIIS0140673609614409/fulltext

Bostrom, Nick. "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents." *Minds and Machines* 22.2 (2012): 71–85.
URL https://doi.org/10.1007/s11023-012-9281-3

———. "Strategic Implications of Openness in AI Development." *Global Policy* 8.2 (2017): 135–148. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1758-5899.12403.
URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1758-5899.12403

Bostrom, Nick, Douglas, Thomas, and Sandberg, Anders. "The Unilateralist's Curse and the Case for a Principle of Conformity." *Social Epistemology* 30.4 (2016): 350–371. Publisher: Routledge _eprint: https://doi.org/10.1080/02691728.2015.1108373.
URL https://doi.org/10.1080/02691728.2015.1108373

Branwen, Gwern. "It Looks Like You're Trying To Take Over The World." .
URL https://gwern.net/fiction/clippy

Brundage, Miles, Avin, Shahar, Clark, Jack, Toner, Helen, Eckersley, Peter, Garfinkel, Ben, Dafoe, Allan, Scharre, Paul, Zeitzoff, Thomas, Filar, Bobby, Anderson, Hyrum, Roff, Heather, Allen, Gregory C., Steinhardt, Jacob, Flynn, Carrick, hÉigeartaigh, Seán Ó, Beard, Simon, Belfield, Haydn, Farquhar, Sebastian, Lyle, Clare, Crootof, Rebecca, Evans, Owain, Page, Michael, Bryson, Joanna, Yampolskiy, Roman, and Amodei, Dario. "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation." arXiv. 2018. ArXiv:1802.07228 [cs].
URL http://arxiv.org/abs/1802.07228

Burr, William. "A Scheme of 'Control': The United States and the Origins of the Nuclear Suppliers' Group, 1974–1976*." *The International History Review* 36.2 (2014): 252–276. Publisher: Routledge _eprint: https://doi.org/10.1080/07075332.2013.864690.
URL https://doi.org/10.1080/07075332.2013.864690

Carlsmith, Joseph. "Is Power-Seeking AI an Existential Risk?" arXiv. 2022. ArXiv:2206.13353 [cs].
URL http://arxiv.org/abs/2206.13353

Casadevall, Arturo, Dermody, Terence S., Imperiale, Michael J., Sandri-Goldin, Rozanne M., and Shenk, Thomas. "On the Need for a National Board To Assess Dual Use Research of Concern." *Journal of Virology* Publisher: American Society for Microbiology 1752 N St., N.W., Washington, DC.
URL https://journals.asm.org/doi/10.1128/JVI.00875-14

Cheh, Mary M. "Progressive Case and the Atomic Energy Act: Waking to the Dangers of Government Information Controls." *George Washington Law Review* 48: 163.
URL https://heinonline.org/HOL/Page?handle=hein.journals/gwlr48&id=175&div=&collection=

Choucri, Nazli, Madnick, Stuart, and Koepke, Priscilla. "Institutions for Cyber Security: International Responses and Data Sharing Initiatives." Tech. Rep. CISL# 2016-10, Cybersecurity Interdisciplinary Systems Laboratory, Massachusetts Institute of Technology, 2016.
URL https://cams.mit.edu/wp-content/uploads/2017-06.pdf

Christopher, Grant. "3D Printing: A Challenge to Nuclear Export Controls." *Strategic Trade Review* 1.1 (2015): 18–25.
URL https://kclpure.kcl.ac.uk/portal/en/publications/3d-printing-a-challenge-to-nuclear-export-controls(c5bd525f-9dee-4b01-977c-14e642da487d).html

Clinton, Bill. "Presidential Decision Directive/NSC-63." The White House. 1998.
URL https://irp.fas.org/offdocs/pdd/pdd-63.htm

Colgan, Jeff D and Miller, Nicholas L. "Rival Hierarchies and the Origins of Nuclear Technology Sharing." *International Studies Quarterly* 63.2 (2019): 310–321.
URL https://doi.org/10.1093/isq/sqz002

Cottier, Ben. "Understanding the diffusion of large language models: summary." Rethink Priorities. 2022.
URL https://rethinkpriorities.org/publications/understanding-the-diffusion-of-large-language-models-summary

Dafoe, Allan. "AI Governance: A Research Agenda." Center for the Governance of AI, Future of Humanity Institute, University of Oxford. 2018.
URL http://www.fhi.ox.ac.uk/wp-content/uploads/GovAI-Agenda.pdf

Dandurand, Luc and Serrano, Oscar Serrano. "Towards improved cyber security information sharing." *2013 5th International Conference on Cyber Conflict (CYCON 2013)*. 2013, 1–16. ISSN: 2325-5374.

Dasgupta, Siddhartha and Kesharwani, Ankit. "Whistleblowing: A Survey of Literature." *IUP Journal of Corporate Governance* 9.4 (2010): 57–70. Num Pages: 14 Place: Hyderabad, India Publisher: IUP Publications.
URL https://www.proquest.com/docview/759597922/abstract/47FBF38E44DB4AB8PQ/1

Department of Health and Human Services. "Screening Framework Guidance for Providers of Synthetic Double-Stranded DNA." Department of Health and Human Services. 2010.
URL https://aspr.hhs.gov/legal/syndna/Documents/syndna-guidance.pdf

Diggans, James and Leproust, Emily. "Next Steps for Access to Safe, Secure DNA Synthesis." *Frontiers in Bioengineering and Biotechnology* 7.
URL https://www.frontiersin.org/articles/10.3389/fbioe.2019.00086

European Union. "Regulation (EU) 2021/821 of the European Parliament and of the {Council} of 20 May 2021 setting up a Union regime for the control of exports, brokering, technical assistance, transit and transfer of dual-use items (recast)." Official Journal of the European Union. 2021.

URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32021R0821&qid=1631277994952

Ezhei, Mansooreh and Tork Ladani, Behrouz. "Information sharing vs. privacy: A game theoretic analysis." *Expert Systems with Applications* 88: 327–337.
URL https://www.sciencedirect.com/science/article/pii/S0957417417304669

Fergusson, Ian F and Kerr, Paul K. "The U.S. Export Control System and the President's Reform Initiative." *Congressional Research Service* .
URL https://www.everycrsreport.com/files/20130507_R41916_8316468ae7f44bf7d7f0634445bee4bce57fa8ae.pdf

Finnveden, Lukas, Riedel, Jess, and Shulman, Carl. "AGI and Lock-In." Effective Altruism Forum. 2022.
URL https://forum.effectivealtruism.org/posts/KqCybin8rtfP3qztq/agi-and-lock-in

Fischer, Sophie-Charlotte, Leung, Jade, Anderljung, Markus, O'Keefe, Cullen, Torges, Stefan, Khan, Saif, Garfinkel, Ben, and Dafoe, Allan. "AI Policy Levers: A Review of the U.S. Government's Tools to Shape AI Research, Development, and Deployment." Tech. Rep. #2021-10, Center for the Governance of AI, Future of Humanity Institute, University of Oxford, 2021.

Gipp, Bela, Meuschke, Norman, and Gernandt, André. "Decentralized Trusted Timestamping using the Crypto Currency Bitcoin." 2015.
URL https://www.gipp.com/wp-content/papercite-data/pdf/gipp15a.pdf

Gordon, Jeff. "Silence for Sale." *Alabama Law Review* 71.4 (2020): 1109–1184.
URL https://papers.ssrn.com/abstract=3633549

Hendrycks, Dan and Mazeika, Mantas. "X-Risk Analysis for AI Research." arXiv. 2022. ArXiv:2206.05862 [cs].
URL http://arxiv.org/abs/2206.05862

Hoffmann, Jordan, Borgeaud, Sebastian, Mensch, Arthur, Buchatskaya, Elena, Cai, Trevor, Rutherford, Eliza, Casas, Diego de Las, Hendricks, Lisa Anne, Welbl, Johannes, Clark, Aidan, Hennigan, Tom, Noland, Eric, Millican, Katie, Driessche, George van den, Damoc, Bogdan, Guy, Aurelia, Osindero, Simon, Simonyan, Karen, Elsen, Erich, Rae, Jack W., Vinyals, Oriol, and Sifre, Laurent. "Training Compute-Optimal Large Language Models." arXiv. 2022. ArXiv:2203.15556 [cs].
URL http://arxiv.org/abs/2203.15556

Hoffmann, Stefan A., Diggans, James, Densmore, Douglas, Dai, Junbiao, Knight, Tom, Leproust, Emily, Boeke, Jef D., Wheeler, Nicole, and Cai, Yizhi. "Safety by Design: Biosafety and Biosecurity in the Age of Synthetic Genomics." *iScience* .106165 (2023).

Householder, Allen D., Wassermann, Grant, Manion, Art, and King, Christopher. "The CERT Guide to Coordinated Vulnerability Disclosure." Tech. Rep. CMU/SEI-2017-SR-022, Software Engineering Institute, 2017.
URL https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=503330

Hubinger, Evan, van Merwijk, Chris, Mikulik, Vladimir, Skalse, Joar, and Garrabrant, Scott. "Risks from Learned Optimization in Advanced Machine Learning Systems." arXiv. 2021. ArXiv:1906.01820 [cs].
URL http://arxiv.org/abs/1906.01820

Jackson, Sarah. "The CEO of the company behind AI chatbot ChatGPT says the worst-case scenario for artificial intelligence is 'lights out for all of us'." *Business Insider* .
URL https://www.businessinsider.com/chatgpt-openai-ceo-worst-case-ai-lights-out-for-all-2023-1

Janczewski, Lech and Portougal, Victor. "Managing Security Clearances within Government Institutions." *Electronic Government: Concepts, Methodologies, Tools, and Applications*. IGI Global, 2008. 3115–3124. ISBN: 9781599049472 Pages: 3115-3124 Publisher: IGI Global.
URL https://www.igi-global.com/chapter/managing-security-clearances-within-government/www.igi-global.com/chapter/managing-security-clearances-within-government/9917

Janczewski, Lech J. and Portougal, Victor. ""Need-to-know" principle and fuzzy security clearances modelling." *Information Management & Computer Security* 8.5 (2000): 210–217. Publisher: MCB UP Ltd.
URL https://doi.org/10.1108/09685220010356247

Kampanakis, Panos. "Security Automation and Threat Information-Sharing Options." *IEEE Security & Privacy* 12.5 (2014): 42–51. Conference Name: IEEE Security & Privacy.

Kavukcuoglu, Koray, Kohli, Pushmeet, Ibrahim, Lila, Bloxwich, Dawn, and Brown, Sasha. "How our principles helped define AlphaFold's release." 2022.
URL https://www.deepmind.com/blog/how-our-principles-helped-define-alphafolds-release

Kennedy, Donald. "Better Never Than Late." *Science* 310.5746 (2005): 195–195. Publisher: American Association for the Advancement of Science.
URL https://www.science.org/doi/10.1126/science.310.5746.195

Kobza, Deborah. "The Maritime & Port Security Information Sharing and Analysis Organization." Global Situational Awareness Center. 2017.
URL https://downloads.regulations.gov/USCG-2016-1084-0012/attachment_1.pdf

Ladish, Jeffrey and Heim, Lennart. "Information security considerations for AI and the long term future." LessWrong. 2022.
URL https://www.lesswrong.com/posts/2oAxpRuadyjN2ERhe/information-security-considerations-for-ai-and-the-long-term

Larson, Selena. "Why the CIA uses board games to train its officers." *CNN Business* .
URL https://money.cnn.com/2017/03/13/technology/cia-board-games-training/

Leahy, Connor, Black, Sid, Scammell, Chris, and Miotta, Andrea. "Conjecture: Internal Infohazard Policy." LessWrong. 2022.
URL https://www.lesswrong.com/posts/Gs29k3beHiqWFZqnn/conjecture-internal-infohazard-policy

Leibowicz, Claire, Adler, Steven, and Eckersley, Peter. "When Is It Appropriate to Publish High-Stakes AI Research?" 2019.
URL https://partnershiponai.org/resource/when-is-it-appropriate-to-publish-high-stakes-ai-research/

Liang, Percy, Bommasani, Rishi, Creel, Kathleen, and Reich, Rob. "The Time Is Now to Develop Community Norms for the Release of Foundation Models." 2022.
URL https://hai.stanford.edu/news/time-now-develop-community-norms-release-foundation-models

Malakoff, David. "H5N1 Researchers Announce End of Research Moratorium." *Science* .
URL https://www.science.org/content/article/h5n1-researchers-announce-end-research-moratorium

McLeish, Caitríona and Nightingale, Paul. "Biosecurity, bioterrorism and the governance of science: The increasing convergence of science and security policy." *Research Policy* 36.10 (2007): 1635–1654.
URL https://www.sciencedirect.com/science/article/pii/S0048733307002089

Metz, Cade and Weise, Karen. "Microsoft Bets Big on the Creator of ChatGPT in Race to Dominate A.I." *The New York Times* .
URL https://www.nytimes.com/2023/01/12/technology/microsoft-openai-chatgpt.html

Miller, Seumas. *Dual Use Science and Technology, Ethics and Weapons of Mass Destruction*. SpringerBriefs in Ethics. Cham: Springer International Publishing, 2018.
URL http://link.springer.com/10.1007/978-3-319-92606-3

Miller, Seumas and Selgelid, Michael J. "Ethical and Philosophical Consideration of the Dual-use Dilemma in the Biological Sciences." *Science and Engineering Ethics* 13.4 (2007): 523–580.
URL https://doi.org/10.1007/s11948-007-9043-4

Morland, Howard. "Born Secret." *Cardozo Law Review* 26: 1401.
URL https://heinonline.org/HOL/Page?handle=hein.journals/cdozo26&id=1417&div=&collection=

National Research Council. *Biotechnology Research in an Age of Terrorism*. Washington, D.C.: National Academies Press, 2004.
URL http://www.nap.edu/catalog/10827

National Science Advisory Board for Biosecurity. "Proposed Framework for the Oversight of Dual Use Life Sciences Research: Strategies for Minimizing the Potential Misuse of Research Information." National Institutes of Health. 2007.
URL https://osp.od.nih.gov/wp-content/uploads/Proposed-Oversight-Framework-for-Dual-Use-Research.pdf

———. "Addressing Biosecurity Concerns Related to Synthetic Biology." National Institutes of Health. 2010.
URL https://osp.od.nih.gov/wp-content/uploads/NSABB_SynBio_DRAFT_Report-FINAL-2_6-7-10.pdf

———. "Framework for Conducting Risk and Benefit Assessments of Gain-of-Function Research." National Institutes of Health. 2015.
URL https://osp.od.nih.gov/wp-content/uploads/2015/09/NSABB_Framework_for_Risk_and_Benefit_Assessments_of_GOF_Research-APPROVED.pdf

———. "Recommendations for the Evaluation and Oversight of Proposed Gain-of-Function Research." National Institutes of Health. 2016.
URL https://osp.od.nih.gov/wp-content/uploads/2016/06/NSABB_Final_Report_Recommendations_Evaluation_Oversight_Proposed_Gain_of_Function_Research.pdf

———. "Proposed Biosecurity Oversight Framework for the Future of Science." National Institutes of Health. 2023.
URL https://osp.od.nih.gov/wp-content/uploads/2023/03/NSABB-Final-Report-Proposed-Biosecurity-Oversight-Framework-for-the-Future-of-Science.pdf

National UFO Reporting Center. "National UFO Reporting Center." 2023.
URL https://nuforc.org/

NeurIPS. "NeurIPS 2021 Paper Checklist Guidelines." 2021.
URL https://neurips.cc/Conferences/2021/PaperInformation/PaperChecklist

Ngo, Richard, Chan, Lawrence, and Mindermann, Sören. "The alignment problem from a deep learning perspective." arXiv. 2023. ArXiv:2209.00626 [cs].
URL http://arxiv.org/abs/2209.00626

OpenAI. "GPT-4 Technical Report." 2023.
URL https://cdn.openai.com/papers/gpt-4.pdf

Pala, Ali and Zhuang, Jun. "Information Sharing in Cybersecurity: A Review." *Decision Analysis* 16.3 (2019): 172–196. Publisher: INFORMS.
URL https://pubsonline.informs.org/doi/abs/10.1287/deca.2018.0387

Palca, Joe. "Panel Seeks to Safeguard Biological Research." *NPR* .
URL https://www.npr.org/templates/story/story.php?storyId=4727492

Pannu, Jaspreet, Sandbrink, Jonas B., Watson, Matthew, Palmer, Megan J., and Relman, David A. "Protocols and risks: when less is more." *Nature Protocols* 17.1 (2022): 1–2. Number: 1 Publisher: Nature Publishing Group.
URL https://www.nature.com/articles/s41596-021-00655-6

Perrigo, Billy. "Bing's AI Is Threatening Users. That's No Laughing Matter." *TIME* .
URL https://time.com/6256529/bing-openai-chatgpt-danger-alignment/

———. "DeepMind CEO Demis Hassabis Urges Caution on AI." *Time* .
URL https://time.com/6246119/demis-hassabis-deepmind-interview/

Pálya, Hanna and Delaney, Oscar. "Deploying Digital Detection of Dangerous DNA." *Journal of Science Policy & Governance* 21.03 (2023).
URL https://www.sciencepolicyjournal.org/article_1038126_JSPG210306.html

Riebe, Thea and Reuter, Christian. "Dual-Use and Dilemmas for Cybersecurity, Peace and Technology Assessment." *Information Technology for Peace and Security: IT Applications and Infrastructures in Conflicts, Crises, War, and Peace*. ed. Christian Reuter. Wiesbaden: Springer Fachmedien, 2019. 165–183.
URL https://doi.org/10.1007/978-3-658-25652-4_8

Ruefle, Robin, Dorofee, Audrey, Mundie, David, Householder, Allen D., Murray, Michael, and Perl, Samuel J. "Computer Security Incident Response Team Development and Evolution." *IEEE Security & Privacy* 12.5 (2014): 16–26. Conference Name: IEEE Security & Privacy.

Russell, Chris. "Analysis of a Secure Time Stamp Device." SANS Institute. 2001.
URL https://www.sans.org/white-papers/746/

Schuett, Jonas and Anderljung, Markus. "Comments on the Initial Draft of the NIST AI Risk Management Framework." Centre for the Governance of AI. 2022.
URL https://www.nist.gov/system/files/documents/2022/05/19/Centre%20for%20the%20Governance%20of%20AI.pdf

Sedenberg, Elaine M. and Dempsey, James X. "Cybersecurity Information Sharing Governance Structures: An Ecosystem of Diversity, Trust, and Tradeoffs." arXiv. 2018. ArXiv:1805.12266 [cs].
URL http://arxiv.org/abs/1805.12266

Shah, Rohin, Varma, Vikrant, Kumar, Ramana, Phuong, Mary, Krakovna, Victoria, Uesato, Jonathan, and Kenton, Zac. "Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals." arXiv. 2022. ArXiv:2210.01790 [cs].
URL http://arxiv.org/abs/2210.01790

Shaw, Carolyn M. "Using Role-Play Scenarios in the IR Classroom: An Examination of Exercises on Peacekeeping Operations and Foreign Policy Decision Making." *International Studies Perspectives* 5.1 (2004): 1–22. Publisher: Oxford University Press.
URL https://www.jstor.org/stable/44218859

Shevlane, Toby. "Structured access: an emerging paradigm for safe AI deployment." arXiv. 2022. ArXiv:2201.05159 [cs].
URL http://arxiv.org/abs/2201.05159

Shevlane, Toby and Dafoe, Allan. "The Offense-Defense Balance of Scientific Knowledge: Does Publishing AI Research Reduce Misuse?" *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. AIES '20. New York, NY, USA: Association for Computing Machinery, 2020, 173–179.
URL https://doi.org/10.1145/3375627.3375815

Shlegeris, Buck. "The prototypical catastrophic AI action is getting root access to its datacenter." AI Alignment Forum. 2022.

URL https://www.alignmentforum.org/posts/BAzCGCys4BkzGDCWR/the-prototypical-catastrophic-ai-action-is-getting-root

Skopik, Florian, Settanni, Giuseppe, and Fiedler, Roman. "A problem shared is a problem halved: A survey on the dimensions of collective cyber defense through security information sharing." *Computers & Security* 60: 154–176.
URL https://www.sciencedirect.com/science/article/pii/S0167404816300347

Solaiman, Irene. "The Gradient of Generative AI Release: Methods and Considerations." arXiv. 2023. ArXiv:2302.04844 [cs].
URL http://arxiv.org/abs/2302.04844

Solaiman, Irene, Brundage, Miles, Clark, Jack, Askell, Amanda, Herbert-Voss, Ariel, Wu, Jeff, Radford, Alec, Krueger, Gretchen, Kim, Jong Wook, Kreps, Sarah, McCain, Miles, Newhouse, Alex, Blazakis, Jason, McGuffie, Kris, and Wang, Jasmine. "Release Strategies and the Social Impacts of Language Models." arXiv. 2019. ArXiv:1908.09203 [cs].
URL http://arxiv.org/abs/1908.09203

Stewart, Ian. "The Contribution of Intangible Technology Controls in Controlling the Spread of Strategic Technologies." *Strategic Trade Review* 1.1 (2015).

Thomassen, Ø., Storesund, A., Søfand, E., and Brattebø, G. "The effects of safety checklists in medicine: a systematic review." *Acta Anaesthesiologica Scandinavica* 58.1 (2014): 5–18. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/aas.12207.
URL https://onlinelibrary.wiley.com/doi/abs/10.1111/aas.12207

Tosh, Deepak, Sengupta, Shamik, Kamhoua, Charles, Kwiat, Kevin, and Martin, Andrew. "An evolutionary game-theoretic framework for cyber-threat information sharing." *2015 IEEE International Conference on Communications (ICC)*. 2015, 7341–7346. ISSN: 1938-1883.

Treadwell, Jonathan R., Lucas, Scott, and Tsou, Amy Y. "Surgical checklists: a systematic review of impacts and implementation." *BMJ Quality & Safety* 23.4 (2014): 299–318. Publisher: BMJ Publishing Group Ltd Section: Systematic review.
URL https://qualitysafety.bmj.com/content/23/4/299

Vijayan, Jaikumar. "What is an ISAC? How sharing cyber threat information improves security." 2022.
URL https://www.csoonline.com/article/3406505/what-is-an-isac-or-isao-how-these-cyber-threat-information-sharing-organizations-improve-security.html

Weiser, Thomas G., Haynes, Alex B., Lashoher, Angela, Dziekan, Gerald, Boorman, Daniel J., Berry, William R., and Gawande, Atul A. "Perspectives in quality: designing the WHO Surgical Safety Checklist." *International Journal for Quality in Health Care* 22.5 (2010): 365–370.
URL https://doi.org/10.1093/intqhc/mzq039

Zwetsloot, Remco and Dafoe, Allan. "Thinking About Risks From AI: Accidents, Misuse and Structure." Lawfare. 2019.
URL https://www.lawfareblog.com/thinking-about-risks-ai-accidents-misuse-and-structure