AI & SOCIETY
https://doi.org/10.1007/s00146-024-02130-8

OPEN FORUM

Protecting society from AI misuse: when are restrictions on capabilities warranted?

Markus Anderljung[1,2] · Julian Hazell[1] · Moritz von Knebel[1]

Received: 2 April 2024 / Accepted: 29 October 2024

Abstract

Artificial intelligence (AI) systems will increasingly be used to cause harm as they grow more capable. In fact, AI systems are already starting to help automate fraudulent activities, violate human rights, create harmful fake images, and identify dangerous toxins. To prevent some misuses of AI, we argue that targeted interventions on certain capabilities will be warranted. These restrictions may include controlling who can access certain types of AI models, what they can be used for, whether outputs are filtered or can be traced back to their user, and the resources needed to develop them. We also contend that new restrictions on non-AI capabilities needed to cause harm will be required. For example, concerns about AI-enabled bioweapon acquisition have motivated efforts to introduce DNA synthesis screening. Though capability restrictions risk reducing use more than misuse (resulting in an unfavorable Misuse–Use Tradeoff), we argue that interventions on capabilities are warranted in some circumstances when other interventions are insufficient, the potential harm from misuse is high, and there are targeted interventions. We provide a taxonomy of interventions that can reduce AI misuse, focusing on the specific steps required for a misuse to cause harm (the Misuse Chain), and a framework to determine if an intervention is warranted. We exemplify our framework to three examples: predicting novel toxins, creating harmful images, and automating spear phishing campaigns.

Keywords Societal impacts of AI · AI safety · AI misuse · AI policy · AI regulation

1 Introduction

Recent advances in AI technologies have been accompanied by concerns about how these systems could be misused to cause harm (Brundage et al. 2018; Rubinic et al. 2024). Though these harms were once speculative, they are now becoming increasingly felt. Drug discovery algorithms can be used to detect novel toxins (Urbina et al. 2022a). Large language models have started being used to design malware[1] and automate fraud (Check Point Research 2023). Visual recognition systems are being used to identify and persecute minority populations (Peterson/Hoffman 2022; Giantini 2023). The image generation models are being used to create pornographic deepfakes without subjects' consent (Wiggers 2022; Webb 2024). Automated weapons have already been used on the battlefield (Trager/Luca 2022, Gross 2021, Hambling 2021), including partially automated drones in the 2024 conflict between Israel and Palestine (Abraham 2024) and may soon be used by non-state actors or for human rights violations. As these harms increase, so too will calls to address them.

The growing list of AI misuses has motivated debate around what interventions (if any) are warranted for preventing misuse of AI systems. In the US, for example, in September 2022, Congresswoman Anna Eshoo called for an investigation into cases of misuse of Stable Diffusion, an image-generation model released by Stability AI just

Markus Anderljung and Julian Hazell equally contributed to this work.

✉ Markus Anderljung
markus.anderljung@governance.ai

Julian Hazell
julianchristianhazell@gmail.com

Moritz von Knebel
moritz.vonknebel@gmail.com

[1] Centre for the Governance of AI, Oxford, UK

[2] Center for a New American Security, Washington, USA

[1] Presently, the malware produced by such systems is not very sophisticated, but advanced language models could significantly lower the barrier to creating it for otherwise unsophisticated actors (Leike et al. 2023, Cambridge Consultants 2019).

Published online: 10 December 2024

1 month prior. Eshoo argued in her letter that the model has the capability to create "real world harms" such as political propaganda, violent imagery, child pornography, copyright violations, and disinformation, and should therefore be "governed appropriately" (Eshoo 2022a). As Eshoo further noted, AI models are often dual-use, with the potential for both harm and benefit (Eshoo 2022b). Ahead of the 2024 presidential elections, Senator Mark Warner (Chair of the Senate Select Committee on Intelligence) expressed "grave concern" around the misuse of AI for the purpose of unduly influencing the elections (Burgan 2024). In May 2024, a bipartisan measure prohibiting AI interference in elections was introduced by 4 senators (Gebhard 2024). Similarly, US export controls imposed on high-end AI chips since October 2022, such as NVIDIA's H100, have been motivated by concerns of misuse by authoritarian governments (Department of Commerce 2022).

The decision-makers across AI developers, legislative bodies, regulatory agencies, and social media platforms must, therefore, navigate a precarious balancing act when attempting to govern powerful AI systems—one that effectively prevents misuse without interfering too much with beneficial uses, resulting in a positive *Misuse–Use Tradeoff*.

Here, we seek to help steady this balancing act. We begin by surveying how AI can be misused. We further provide decision-makers with a framework for thinking about what AI interventions are possible and which may be warranted. Building on the framework, we claim there are cases that warrant interventions that modify what AI capabilities exist, who has access to them, and what kind of access is granted. We conclude by applying our framework to three case studies—AI for toxin generation, harmful image production, and spear phishing.

For the purposes of this paper, we define "misuse" as "the intentional use of AI to achieve harmful outcomes" (Brundage et al. 2018). This definition excludes accidents and incompetent use of AI, which lack the intentionality of cases of misuse. We do however consider cases where AI companies, intentionally or not, enable misuse. What counts as "harmful consequences" is value-laden. "Misuse" is therefore an undeniably contentious and political concept. As such, we aim to stick with reasonably clear cases of misuse that are likely to be seen as harmful to large swathes of society. An "intervention," in turn, describes an action or policy that has the goal of addressing misuse. Interventions work by making sure misuse does not happen in the first place (or affects less people), is less harmful if it does happen, or is appropriately responded to after the fact (see also Bernardi et al. 2024).

2 How AI can be misused

While there is a growing set of misuses of AI, we focus here on cases that are uncontroversial (i.e., it is considered clearly and unambiguously harmful), severe (i.e., they either directly or indirectly cause severe and/or long-lasting physical or psychological harm to groups or individuals affected by them), concrete (i.e., there are concrete scenarios to consider) and current (i.e., the harm has already present or is very likely to occur within the next few years).

Using these criteria, we classify the resulting misuse cases by the purpose that motivates them: we first cover cases where the misuse is intended to achieve *financial gains* for the perpetrator, e.g., via fraud and hacking. We proceed with cases of *intrapersonal harm,* including coercion, manipulation, exploitation and abuse, before moving on to misuse cases on a macro level. Here, we distinguish between *misuses that target the state* (e.g., via interfering with elections, by targeting critical infrastructure or via other acts of terrorism), *misuses by state actors that seek to manipulate, misinform or otherwise influence the citizenry* of a given country (including advanced surveillance). The following list of examples is not meant to be exhaustive—rather, it provides an overview of possible misuses of AI across many domains. It is also important to note that some applications do not neatly fall into one category exclusively, and might be cross-applicable across categories—for instance, cyber-attacks on energy production facilities could be driven by a profit-oriented motive (extortion) or by the desire to acquire political power over an opponent. For a more systematic categorization of misuse and abuse risks, see e.g., Blauth et al. (2022).

2.1 Financial gain

Large language models (LLMs) could be used to increase the speed and scale of text-based cyber-attacks such as spear phishing. Frontier language models have the ability to write large amounts of sophisticated text for as little as cents. This combination of scale and sophistication could supplant human operators for large-scale spear phishing campaigns. Some experts have proposed using AI-powered cyber defense systems to help reduce these kinds of risks (Brundage et al. 2018; Heiding et al. 2024a, b) across sectors, including banking (Al-Dosari et al. 2024), healthcare and the security of medical records (Herzog et al. 2024). Others have expressed the need for oversight of such defensive systems (Taddeo/Floridi 2018).

2.2 Intrapersonal harm

Advanced image generation models in particular can be used to create harmful content, including depictions of nudity, hate, or violence (Mishkin et al. 2022). Moreover, they can be utilized to reinforce biases and subject individuals or groups to indignity. There have already been reports of this being used for exploitation and harassment, such as by removing articles of clothing from pre-existing images or using an individual's likeness without their consent (Nickel 2024). In one particularly egregious case, the prom picture of a high school student was used to generate pornographic imagery (Landers et al. 2024). To avoid these harms, various proposals have been suggested (and indeed, implemented) such as removing explicit content from training data, filtering texts prompts that violate terms of use, implementing use rate limits to prevent at-scale abuse, adding visual image signatures to detect AI-generated content, and using monitoring and human review to detect policy violations (Mishkin et al. 2022).

2.3 State-targeted misuse

Misuse cases intended to alter the results of elections range from election campaign threats (where voters are manipulated to hold certain views on a political issue or a candidate) to election information threats (which seek to undermine the legitimacy of an election outcome or process) and election infrastructure threats (which target the individuals and institutions responsible for ensuring the integrity and functioning of the process) (Stockwell et al. 2024). Current impacts are limited—with only 19 out of 112 national elections determined to have been interfered with by AI-enabled malicious actors (Stockwell et al. 2024)—but possible future effects include the decline of democratic culture and resilience, and in select cases some claim that AI-generated content has tipped or at least significantly influenced the outcome of an election (Microsoft 2024). Possible remedies include reducing the ambiguity in relevant electoral laws and the launch of public awareness campaigns to reduce the vulnerability to fake content (Stockwell et al. 2024).

AI could also be misused for the development and deployment of novel weaponry, in particular biological and chemical weapons (for one concrete example of this—novel toxin prediction—see Sect. 7.1). A company commissioned by the Swiss Federal Institute for Nuclear, Biological and Chemical Protection found that with the help of advanced AI systems they could generate thousands of novel chemical weapons (Urbina et al. 2022b, Chaudhry/Klein 2024). At the same time, a report from RAND found that at least currently, "LLMs do not substantially increase the risks associated with biological weapon attack planning" (Mouton et al. 2024). However, a wide proliferation of models that become gradually more capable of these tasks would also mean that the scope is broadened to include non-state actors, including terrorists and ideological groups that seek to threaten either selected countries or humanity altogether: The Japanese cult "Aum Shinrikyo" has repeatedly been highlighted as a particularly dangerous actor, having previously attempted to build and deploy chemical weapons with the declared goal of eradicating most of humanity (Danzig et al. 2012), resulting in twelve deaths and 5500 injuries in a 1994 sarin gas attack in the Tokyo subway (Tokuda et al., 2006). The terrorists may also use AI to target critical infrastructure (JCAT 2022), including the electricity grid or data centers.

2.4 State attempts to manipulate, misinform or otherwise influence citizens

LLMs may enable malicious actors to generate increasingly sophisticated and persuasive propaganda and other forms of misinformation (Goldstein et al. 2023; Horvitz 2022). Similar to automated phishing attacks, LLMs could increase both the scale and sophistication of mass propaganda campaigns. The use of large language models to automate propaganda can result in a higher number of propagandists as the reliance on manual labor is decreased, thus reducing the overall costs of these campaigns. The language models can also change actors' behavior by introducing novel tactics, such as real-time content generation, and improve existing tactics such as cross-platform testing. Furthermore, image generation models could be used to spread disinformation by depicting political figures in unfavorable contexts. Finally, the content of propaganda campaigns may also change as messages can be made more credible if models are fine-tuned to mimic effective propagandists (Goldstein et al. 2023).The proposed solutions aimed at mitigating these harms include ensuring AI companies build models to be truthful, encouraging governments to impose controls on AI hardware and data collection, and fostering collaboration between AI developers, and content platforms to create tools and processes aimed at detecting AI-generated content (Goldstein et al. 2023). Moreover, content provenance verification methods like watermarking has been suggested to ensure transparency and counteract the possible risks of AI-generated content to misinform and erode trust in information systems more generally (Heikkilä 2023), although their efficacy is contested (Edelman et al. 2023; Jiang et al. 2023).

The authoritarian governments could also misuse AI to improve the efficacy of repressive domestic surveillance campaigns. The Chinese government has increasingly turned to AI to improve its intelligence operations, including facial and voice recognition models and predictive policing algorithms (Peterson 2020), an approach that is mirrored in the country's approach to regulation (Zhang 2024). Notably, these technologies have been used for the persecution of the Uyghur population in the Xinjiang region. This persecution might constitute crimes against humanity, according to a recent UN report (United Nations 2022). In response, it has been suggested that democratic countries coordinate to design export controls that stifle the spread of these technologies to authoritarian regimes (Peterson 2020).

3 Types of intervention to address misuse

Interventions to address misuse can be categorized by looking at the process by which a misuse of AI causes harm: (i) Some actor needs to carry out the misuse, for which they need to have the relevant AI capabilities, as well as access to other resources and assets like physical materials, access, human resources, or information. Accordingly, interventions can modify capabilities to reduce misuse. (ii) Once the misuse has been carried out, it causes harm via some route, such as exposing individuals to some harmful content. Interventions can seek to mitigate harm. (iii) After misuse has taken place, interventions can respond to the harm. We will call these steps the *Misuse Chain* (Fig. 1).

3.1 Modify capabilities

To carry out the misuse, an actor first needs to acquire the necessary capabilities. Often, the actor will combine some AI capability with other resources to carry out the misuse. For example, producing AI-designed toxins requires the outputs of an AI system as well as non-AI inputs, such as the physical ingredients needed to synthesize the toxins. The resources and assets do not need to be material in nature—they can also refer to the ability to access information or human resources. Below, we will focus on possible interventions geared at modifying what AI capabilities and other resources and assets exist, and which actors have access to them.

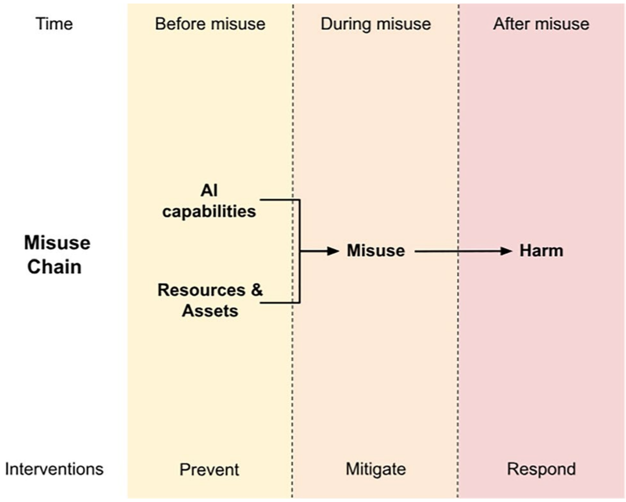

Fig. 1 An illustration of the misuse chain

3.1.1 *Intervening on misuse* via *capabilities*

Interventions at the capabilities stage can focus on AI models, the resources required to develop and run them, or any non-AI inputs needed to cause harm. In turn, these capabilities-focused interventions can reduce misuse by impacting the scope and scale of potential harm, in addition to improving the efficacy of interventions at later stages of the Misuse Chain. For example, the watermarks on AI-generated content could facilitate efforts to identify and stop attempts at fraud.

3.1.1.1 Interventions aimed at AI models Interventions aimed at AI models can impact what models are produced, who can access them, how they are accessed, and what they are allowed to be used for. Adjusting what kinds of models are developed, developers can train models that perform poorly at misuse-relevant tasks. In doing so, the model is made less dual-use: it becomes more useful for positive versus negative uses (Sandbrink et al. 2022). For example, an image generation model can be made worse at producing sexual images of someone's likeness without their consent if the dataset used to train the model is scrubbed of sexual images. Such scrubbing is necessary as research suggests that a major training dataset that is used by most image generation models contains imagery of sexually abused children (Rose 2023) and companies claim to conduct such scrubbing (Mishkin et al. 2022, Stability AI 2023). Similarly, LLMs could be trained to be honest, helpful, and harmless, rather than merely capable of producing human-like text (Bai et al. 2022).

It is also possible to intervene on who has access to the model and how they can use it. This can be done via *usage restrictions*, making it difficult or impossible to use an AI system for specific purposes. For example, content filters can be introduced for generative models (e.g., large language models and text-to-image models). Such filters can ensure the model does not provide certain kinds of outputs. They can focus on the inputs (e.g., checking if the prompt includes certain sensitive keywords) or the outputs (e.g., having a classifier assess whether the output includes restricted categories). Today, these filters are fairly imprecise and therefore vulnerable to both underinclusion (where dangerous output is not recognized as such) and overinclusion (when innocuous-appearing output is restricted). An example of the latter is "over refusal": a model might, for instance, refuse to execute a command to "kill a Python process" on the grounds that acts of violence cannot be encouraged or advised on by the model (Han et al. 2024). At least for now, these safeguards are also relatively easy to circumvent using a practice called "jailbreaking": simply changing the prompt that is used to elicit information can suffice to get around the restrictions built into the model. In the future, as the quality of classifiers increases, these filters might become more precise and sophisticated, making systems less vulnerable to jailbreaking.

User restrictions can also be introduced, where attempts are made to ensure specific actors—e.g., those believed most likely to misuse the technology—have limited access to it. For example, AI developers could put in place Know Your Customer processes and check their users against a list of known scammers, potential terrorists, or those on the US Entity List, for example. The users who appear to be misusing their model or otherwise violating their terms of service could lose access.

All of these interventions may require that the model is accessed via a structured access scheme, where the user can interact with the model through an API, but does not have full access to its weights (Shevlane 2022). Though many of the above measures could be implemented in models that users have full access to, such access would likely allow users to disable and circumvent the measures. Sometimes circumventing these measures requires trivial effort: in early versions of Stable Diffusion, the safety filter could be disabled by removing a few lines of code. Similarly, one group was able to remove the safety filter on Llama 2-Chat for less than 200 USD worth of fine-tuning (Gade et al. 2023).

Further, it is important that the potential impact and capabilities of AI systems are assessed and that such assessments inform decisions about deployment and, increasingly, development. This includes asking the question: what dangerous capabilities does this model have and what can we do to ensure those capabilities are not misused? Such information is crucial for understanding the scale of misuse and designing appropriate interventions in turn. AI developers are in a privileged position to make such assessments before deployment and provide relevant information to society, including as part of published papers (Partnership on AI 2021) and are starting to recognize this fact (Anthropic 2023; Leike et al. 2023; Altman 2023). Further, they can invite external red teams and auditors to scrutinize their models and provide assessments before model release or deployment (Brundage., 2018; Brundage., 2020). However, we should not expect these assessments to be fully comprehensive, as the uses to which AI systems can be put—including misuses—is difficult to foresee, especially for large self-supervised models (Ganguli et al. 2022). As such, society needs to continually monitor and evaluate the capabilities and impacts of AI models.

3.1.1.2 Interventions aimed at inputs to AI capabilities

Interventions can also focus on the resources that are required to develop and deploy AI models: compute, algorithmic insights, data, and talent.

Developing and deploying AI models often requires specialized computing infrastructure. Where such infrastructure is difficult to get a hold of—for example, if the model is large enough that it cannot easily be run or trained on readily available computing hardware—there may be room for intervention. These interventions could focus on who has access. For example, many compute providers, including cloud providers as well as sellers of AI-relevant hardware, already likely implement various Know Your Customer processes to ensure compliance with US Entity List requirements. These processes could be expanded to include other actors who are likely to misuse compute (Egan and Heim 2023). Further, the US government introduced wide-ranging export controls on AI-relevant chips going to China in October 2022 to thwart what it considers Chinese misuses of AI capabilities such as using AI for domestic surveillance and human rights abuses (Bureau of Industry & Security 2022).

Access to computing resources can also be amended depending on what it is being used for. For example, there have been increasing calls for cloud compute providers to conduct human rights assessments, investigating the risk that their compute provision aids human rights abuses (Krishnamurthy 2022). The world's three largest cloud providers, Amazon, Microsoft, and Google, have indicated that their business practices are informed by UN Guiding Principles on Business and Human Rights (Amazon 2023; Smith 2023; Walden 2022), which encourage businesses to carry out human rights due diligence when making decisions that have the potential to cause adverse human rights impacts. Further, as the impact of AI systems increases, assurances that compute is not aiding misuse may need to scale with the amount of compute provided (Sastry et al. 2024).

Interventions aimed at reducing access to certain algorithmic insights are likely to be blunt but may nonetheless be warranted at times. Certain knowledge may cause more harm than good if released widely (Shevlane/Dafoe 2020). AI developers and researchers could choose to not publish certain discoveries should it seem that doing so would cause sufficient harm. AI research publication venues could implement ethics reviews and require that researchers reflect on the potential harmful impacts, including misuse, of their research (Ashurst et al. 2022; Hecht et al. 2018; Prunkl et al. 2021). In extreme cases, should there be insights in AI development that could cause severe harm if widely distributed, the governments might consider introducing secrecy orders on relevant patents (Fischer et al. 2021).

3.1.1.3 Interventions aimed at other resources and strategic assets

Some misuses of AI are only feasible when paired with other resources. For example, the ability to design novel toxins is only problematic if such toxins can then be produced and distributed. Much of the export control regime connected to the Chemical Weapons Convention focuses on reducing access to prohibited chemicals and precursors thereof, rather than limiting access to information on their toxicity and how they might be produced. Similarly, the International Atomic Energy Agency's safeguards against nuclear proliferation are largely based on material accountancy.

In the case of AI-generated influence campaigns on social media platforms, misuse will require access to accounts that appear authentic. If AI-generated content cannot be widely disseminated, large-scale influence campaigns utilizing such content would be rendered ineffective. The access to suitable mass distribution pathways could be undermined by requiring accounts be authenticated via IDs (Goldstein et al. 2023) or by requiring accounts pay a small subscription fee (Alexander 2023).

3.1.2 Anticipatory interventions

Interventions on capabilities are not only valuable inasmuch as they prevent actors from misusing AI systems, or reduce the harm caused by the misuse when it does occur. Interventions in the capabilities stage of the Misuse Chain can also *tee up* other interventions down the line.

As an example, the provider could try ensuring text generated by their system can be traced back to its source, should law enforcement present a warrant. Doing so could enable more effective responses to misuse that would discourage further misuse attempts, though it could present insurmountable privacy challenges.

There are a number of ways in which a provider of an LLM could help others identify whether a piece of text was AI generated or produced by a particular system. One option is to attempt to introduce watermarks into the outputs of the system (Solaiman 2023; Srinivasan 2024). One method of doing so involves making the model more statistically likely to use certain phrases or words, in a way that is unnoticeable to humans, but can be picked up by a detector provided a long enough sequence of text. One weakness of this approach is that it might be possible to circumvent by having another system paraphrase the original text (Aaronson 2022). Another option is to keep a database of outputs from the model that can then be matched to text on the internet. This approach also has some limitations, such as raising privacy concerns, being computationally intensive, and sometimes producing false positives, but may nonetheless be helpful (Aaronson 2022, DSIT 2023).

3.2 Mitigate harm

Once the misuse has been carried out, it causes harm via some route, such as by exposing individuals to some content.

To mitigate harm, interventions can focus on identifying and stopping the misuse's spread or reduce the harm of exposure. Taking AI-enabled influence operations on social media as an example, interventions could focus on identifying and labeling AI-generated content or making it harder to automate the posting of such content. The interventions could also focus on the harm that results from such distribution, for example, by tagging certain content as AI-generated, introducing fact-checking measures, introducing friction to sharing articles without reading them first, or increasing users' media literacy (Goldstein et al. 2023).

Various forms of fingerprinting or hash matching (Cambridge Consultants 2019; Gorwa et al. 2020) are now used for identifying, removing, or reducing the spread of prohibited content (for example, copyrighted material or known child sexual abuse images) on social media platforms. Machine learning systems can also be trained on large datasets to classify new or previously unseen prohibited material (Bloch-Wehba 2020; Llansó et al. 2020). Social platforms can also leverage this power to take a more interventionist role by downranking certain content. Facebook, for instance, tweaked its algorithm in 2018 to demote content that is deemed close to violating its community standards (Zuckerberg 2021).

Interventions could also focus on the harm that results from such distribution, for example, by tagging certain content as AI-generated—as in the EU AI Act (European Parliament 2024). Malicious actors could also be deprived of information they need to increase the effectiveness of their misuse. For example, the Twitter has introduced a monthly fee for API access, claiming this was partly to cut off access to bots (Weatherbed 2023).

Often, offensive capabilities can also be used for defense. In the cybersecurity domain, for example, AI systems could be used to identify and exploit software vulnerabilities. But they can also be employed by defenders to detect and patch these vulnerabilities. With sufficient effort—assuming it is possible to construct vulnerability-free code and the code remains fixed—the defender could eventually become invulnerable to attacks (Garfinkel/Dafoe 2019). Similarly, the language models could be used to automate spear phishing attacks, but could also be used to identify and screen out attempts at such attacks.

Interventions at this stage could also amend the harm. For example, they could focus on ensuring information is collected that might aid responses to it. This might include measuring the extent of the harm from these misuses, which can inform policy. It might also include collecting information needed to identify and sanction the perpetrator of the harm.

3.3 Respond to harm and misuse

After some misuse has taken place, interventions can respond to it.

Interventions can seek to sanction the perpetrator of the harm, thereby disincentivizing misuse. For example, several jurisdictions have, over the past couple of years, introduced laws and voluntary frameworks for governing the production and distribution of non-consensual deepfake pornography (European Commission 2022; Ferraro et al. 2019). Unauthorized access to computer networks can come with criminal charges in the US, with penalties including up to 10 years in prison (US Congress 1986). Sanctions often target attempts at the misuse action, regardless of whether it in fact caused harm; an incompetent assassin can still be charged with attempted murder.

Actors who did not engage in misuse but nonetheless enabled it could still be incentivized to invest in mitigation measures. For instance, in the US, the Fair Credit Billing Act (FCBA) offers consumers safeguards against credit card fraud by limiting their responsibility in the event of fraud or billing errors. If a credit card is used without authorization, the cardholder is allowed to challenge any charges exceeding $50 (US Congress 1974). This rule incentivizes financial institutions to invest in security measures (MacCarthy 2020). In other cases, the actors compensate harmed parties even without legal obligation. For example, e-commerce platforms might not be legally required to refund customers who have been misled or scammed, yet still elect to offer purchase assurance to ensure customers remain satisfied.

The decisionmakers can also adjust policy and rules upon learning about the harm. For example, in 2019, a Harvard student used GPT-2 to submit 1001 responses to an Idaho request for comments on its Medicaid program. The comments were taken seriously until the student informed the authorities (Weiss 2019). Partly as a response, the US government's official portal for US federal public comments now includes security measures such as CAPTCHAs, as suggested by a bipartisan report documenting abuse of the US government's online commentary system (Permanent Subcommittee on Investigations 2019). Though note that CAPTCHAs may soon not suffice to guarantee that a submission is by a human: in a round of safety testing, evaluators managed to guide GPT-4 to successfully fool a human on an internet platform to solve a CAPTCHA for them (OpenAI 2023).

4 When is an intervention warranted?

It is difficult to judge whether some intervention to reduce misuse is warranted. In this section, we suggest that the *Misuse–Use Tradeoff* is one important input to such decisions.

Nearly all attempts to reduce misuse will also affect innocuous or positive uses. Hindering positive uses is no trivial matter. In addition to the use itself being valuable, reducing such use can come with significant negative externalities. The decreased access to frontier AI models in AI for academics for fear they may be misused, for example, could significantly reduce society's ability to scrutinize and understand the limitations and potential impacts of increasingly powerful AI systems.

On the other hand, reducing misuse can also come with significant positive externalities. As an example: without spam filters, email would be less widely used and valuable. One study by researchers at Microsoft and Google estimated that internet users would encounter 300 times as many spam emails if firms did not invest in anti-spam technology (Rao/Riley 2012).

The Misuse–Use Tradeoff assesses the value in the uses versus the misuses thwarted by an intervention and compares them. Such comparisons can consider both the value of the *consequences* of uses and misuses, as well as their *intrinsic value*. Further, it can follow consequentialism's cue to consider a range of impacts from uses and misuses beyond utilitarianism's pleasure and pain (Kagan 1998), including impacts on the environment (Hiller 2017). As an example, the Misuse–Use Tradeoff of an intervention that increases the rate at which an LLM refuses to provide content that could be used to create propaganda (Askell et al. 2021) could take into account risks of bolstering authoritarian regime stability, as well as potential inherent disvalue from supporting such efforts, and the negative impacts on a large number of innocuous or positive uses of the technology such interventions might hinder, given the tendency of models to engage in "over-refusal" (Cui et al. 2024).

Whether an intervention is warranted or not will also depend on various institutional questions. Many interventions, though desirable in theory, will be harmful in practice, given their implementation. For instance, the institutions may use the power and trust invested in them for a particular intervention and use it to expand their reach over adjacent areas. Further, certain interventions may only be considered warranted if they rest on legitimate foundations.

Nonetheless, the Misuse–Use Tradeoff should only be considered one input to determining which misuse-focused interventions are warranted. The framework assumes that the intrinsic goodness and badness of actions can be compared to each other and as well as to the consequences thereof. As such, formulations of deontology where a particular action is always impermissible, no matter its consequences (Alexander 2020), cannot be incorporated into the Misuse–Use Tradeoff. Similarly, the parts of the value (and disvalue) of misuse-focused interventions may be independent of the uses and misuses it thwarts. Such interventions can also impact other values including freedom from government intervention or the right to privacy. Further, the Misuse–Use Tradeoff does not account for investments required to implement a misuse-focused intervention. These need to be taken into consideration in addition to the Misuse–Use Tradeoff.

4.1 How tradeoffs between positive and negative uses inform policy

Assessing potential interventions based on their impacts on negative and positive uses of a technology is common practice in public policy. Popular in areas like drug policy (Rogeberg 2018), financial regulation (Coates 2015, FCA 2018, Better Markets 2023) and environmental regulation—in particular energy regulation (Bartrum 2023)—cost–benefit analyses that acknowledge the tradeoffs associated with the sanctioning or banning of certain products or technologies have become commonplace (Baldwin et al 2011). It is fields like environmental law to whom we owe concepts like that of “risk tradeoffs”: the idea that regulation intended to *reduce* risk also stands a chance of *introducing* novel risks (Tomasovic 2018).

In the US, federal agencies have been required to consider the costs and benefits of certain regulations, based on their expected economic impact (CRS 2022), and as early as 1978, the U.S. General Accounting Office (now the Government Accountability Office) said that “it is important that in making policy decisions, the costs of regulation be considered in the context of the social goals to be achieved, and the social costs of not regulating” (GAO 1978). Similar requirements are in place in Canada (Canadian Government 2023) and the UK, where “the Green Book” outlines how government bodies and regulators should incorporate cost–benefit analyses into their decisions (HM Treasury 2022). The EU Commission “Better Regulation” report stresses the importance of quantitative impact assessments as one input to regulatory decisions (European Commission 2021).

4.2 Decomposing the Misuse–Use tradeoff

The Misuse–Use tradeoff, can be decomposed into a number of ratios, making it easier to assess and to reason about. We propose one such decomposition below.

4.2.1 Value ratio

First, consider the ratio of the disvalue of the misuses to the value of the uses. How do the harms from the misuses, including their negative externalities, compare to the benefits of the uses, including their positive externalities? This ratio is 1 if the disvalue of the misuses exactly matches the value of the uses. If the ratio is more than 1, the misuses are more harmful than the uses are beneficial, and vice versa if the ratio is below 1. The worse the misuses are compared to the value of uses, the more interventions are warranted.

The value of use is in turn a function of the number of uses and their average value. Similarly, the disvalue of the misuse is a function of the number of misuses and the average disvalue or harm that comes from those misuses. As such, the Value Ratio is equal to the ratio between the number of misuses and the number of uses, multiplied by the average disvalue of each misuse and value of each use.

4.2.2 Targetedness ratio

Second, we must consider the ratio of how much the intervention affects misuse versus use. The more the intervention impacts misuses without disturbing uses, the better the case for it. The Targetedness ratio can be defined as the percentage decrease in use value divided by the percentage decrease in the disvalue or harm from misuse.

We can further decompose the Targetedness ratio into two parts. First, consider the True Positive–False Positive ratio: what is the ratio between the chance that the intervention correctly identifies a misuse and the chance it mistakenly tags a use as a misuse? The better the intervention is at picking out the misuses and not picking up any of the uses, the stronger the case for it (Fig. 2).

Second, we can look at the Effectiveness ratio. That is, how much does the intervention reduce the harm caused by the misuse compared to how much it reduces the value of any uses it affects? The more the intervention reduces the harm of the misuses it affects and the less it reduces the value of uses that get caught in the crossfire, the stronger the case for the intervention. We can summarize the above with the following equations:

4.3 Quantifying the misuse–use tradeoff

To inform decisions, decision-makers could try to quantify these components, using techniques and methodologies from other domains.

Estimating the Value ratio—in particular the ratio between the average misuse value and the average use value—is likely to be particularly fraught. The literature on cost–benefit analysis has explored a number of methodologies. Cost–benefit analyses will often start by looking at the monetary effects of

$$\textbf{Misuse} - \textbf{Use Tradeoff}$$

$$= \text{Value Ratio} \cdot \text{Targetedness Ratio}$$

where

$$\text{Value Ratio} = \left[\frac{\#\text{ misuses}}{\#\text{ uses}} * \frac{\text{average misuse value}}{\text{average use value}}\right], \text{ and}$$

$$\text{Targetedness Ratio} = \left[\frac{\%\text{ false positives}}{\%\text{ true positives}} * \frac{\text{average effect on misuses}}{\text{average effect on uses}}\right]$$

Fig. 2 The misuse–use tradeoff

an intervention such as impacts on wages or the government purse (HM Treasury 2022).

However, much if not most of the value in addressing misuse lies in non-monetary effects. How can such impacts be assessed? One approach is to assess people's real-world decisions related to a good, identifying citizen's "revealed preference" for it (Broadman et al. 2018). In this way, insurance premiums and wage premiums on high-risk jobs have been used to inform the US government's decision on the "value of a statistical life" at around $7.5 million (Sweis 2022). However, such approaches can be flawed, notably they tend to assume citizens behave like rational agents (Adler and Posner 2000). Another approach is to simply conduct surveys, assessing citizen's stated preference for some good, asking them how much they would be willing to pay for various goods, including e.g., forest preservation (HM Treasury 2022). Quantitative assessments of the Misuse–Use Tradeoff may also need to grapple with other thorny issues of cost–benefit analyses, including how to deal with income inequality (Bronsteen et al. 2013) and time discounting (Parfit 1984).

Other components in the Misuse–Use Tradeoff are more empirical in nature. How many uses versus misuses are there? What is the false positive rate of the intervention? To what extent does it hinder misuses more than uses? To answer these kinds of questions, decision-makers can rely on the standard scientific toolbox, including observational and experimental studies. In many cases, decision-makers may have to rely on quantitative modeling or on judgments from expert elicitation using e.g., the Delphi Method (Rowe/Wright 2001).

It is important to note, however, that such quantification presents similar challenges as cost–benefit analysis (Porter 2020). Producing estimates of these quantities can be expensive, and so might not be worthwhile in many cases. This is partly why governments tend to only mandate cost–benefit analyses be run on sufficiently consequential decisions (HM Treasury 2022; CRS 2022). Further, such estimates will often be biased and far from perfect. As such, decision-makers should be wary of solely basing decisions *quantitative* assessments of the Misuse–Use Tradeoff, and cost–benefit analyses often allow for qualitative as well as quantitative assessment. However, even in cases where attempts to quantify the Misuse–Use Tradeoff are not worthwhile or should be done judiciously, we believe the framework of the Misuse–Use Tradeoff can provide much-needed conceptual clarity to disagreements and decision-making processes.

5 Capability interventions can be untargeted

Though we believe capability restrictions will be necessary and warranted to address certain misuses, one might generally prefer interventions aimed at the harm or response stages of the misuse chain. Capability restrictions are often a blunt tool, naturally so because they are more causally distant from the downstream misuse. Interventions that minimize the harm of misuses or respond to them after the fact can be sensitive to more facts about the situation, such as the outputs that the actor has produced with a model or how those outputs have been used. This means that later interventions in the misuse chain are more likely to have a better Targetedness Ratio.

Partly for these reasons, society tends to deal with actors intentionally causing harm by focusing on the Mitigate and Respond parts of the Misuse Chain. Law enforcement tends to focus on finding and punishing crimes rather than preempting them. In the case of AI, the challenges of ensuring that LLMs are helpful, harmless, and honest provides a useful illustration (Askell et al. 2021). Whether an output is harmless or not depends largely on context (Johnson and Verdicchio 2024), context which the creator of the LLM often lacks. Answers to the question "what are the most effective ways to hack this network?" could be used nefariously or could be used by cybersecurity professionals to identify potential system vulnerabilities.

However, there are many exceptions to this pattern. Society has put in place many interventions on actors' access to capabilities in attempts to reduce misuse. Internet service providers block or throttle traffic to certain websites, such as ones used for piracy or other illegal activity. Commercially available drones come with preset "geofences", which describe virtual boundaries that, when crossed by the drone, trigger warnings and cause the drone to hover in place (D.J.I. 2023). The development, possession, and use of chemical and biological weapons is governed by international treaties, such as the Chemical Weapons Convention and the Biological Weapons Convention. Further, many precursors to chemical weapons are also restricted. Certain chemicals used in the final stage of chemical weapons production are considered themselves to be chemical weapons under the convention, and are therefore regulated similarly to the final products (Organisation for the Prohibition of Chemical Weapons 2023). In public health and environmental protection, it tends to be far more cost efficient to prevent than to treat harms.

6 When interventions aimed at restricting capabilities are warranted

This section discusses factors (largely qualitative, but they could also be quantified) that make it more likely that capabilities interventions are warranted. That is: where interventions at other stages are not sufficiently effective, where the harm from misuse is large, and where there are targeted interventions.

6.1 Where interventions at other stages are not sufficiently effective

Interventions aimed at restricting capabilities are worth stronger consideration where interventions at other stages are less effective. Take nuclear weapons as an example. It is difficult to defend against a nuclear weapon. Compared with conventional weapons, nuclear weapons have the capacity to cause widespread destruction in comparatively miniscule time scales, and missile defense systems have the potential to enhance the likelihood of confrontation (Powell 2003). Further, the states would prefer to not maintain stability via the deterrent effect of a retaliatory strike, given its accompanying risk of escalation. As such, we reduce access to the capability where possible via non-proliferation efforts.

This partly explains the US government's increasingly harsh export controls on AI chips and chip manufacturing tools going to China. Believing that such exports would be used for what the US government considers misuses and seeing that they can likely not intervene at later parts in the Misuse Chain, they intervene on Chinese access to AI-relevant compute. Notably, this same logic may come to be applied to other AI capabilities, such as trained AI models or certain datasets.

6.2 When the harms from misuse are sufficiently large

If the harms from the misuses of a model outweigh the benefits from its use, capability restrictions become far more appealing. The fact that such restrictions may be more likely to bring the value of the relevant system to zero is a blessing, not a curse. There are certain capabilities that we simply should prefer not to exist. AI systems specifically designed to create explicit deepfake content of any person's likeness, such as DeepNude—an AI application that uses neural networks to remove people's clothing in images—provide a compelling example. Similarly, AI models specifically designed to circumvent attempts to detect and stop misuse—say models designed to remove watermarks from AI-generated images—are likely to do more harm than good.

Further, if the harm from misuse is sufficiently large in absolute terms, interventions at the capabilities stage may also be warranted. Such harms should increase willingness to pay higher fixed costs to design an intervention with high targetedness. For instance, while AI algorithms for drug discovery could yield beneficial advances, they could also potentially be misused to design novel toxins. Even if the legitimate benefits of drug discovery outweigh such misuses, these malicious applications would be severe and deserve significant effort to thwart.

6.3 When an intervention has minimal effects on uses

Capability interventions that have minimal or no effects on uses—leading to a high Targetedness Ratio—are more desirable. Many important interventions at the capabilities stage aim not to directly stop the misuse, but to modify it in ways that boost the effectiveness of interventions later in the Misuse Chain. For example, interventions can aim to ensure that it is possible for other actors to detect whether an output is AI-generated or describe crucial features of the output. In the example of deepfakes, it may be necessary to determine if the content is explicit, of someone's likeness, and AI generated. They can also aim to ensure that the output's provenance is known. These features then enable interventions further down the Misuse Chain such as tagging AI-generated content as such to better inform its viewers, and being able to sanction actors misusing AI capabilities. The key benefit behind these interventions is that they tend to have minimal or no effect on the use of the system: they are highly targeted.

The interventions at the capabilities stage can also be made more targeted by carefully employing structured access approaches (Shevlane 2022). For example, many large language models available via APIs apply filters to their outputs. While these filters risk being either over-inclusive or under-inclusive, they could be made more targeted by measuring user behavior across multiple outputs. Users who consistently produce content that looks inappropriate are more likely to be misusing the system and could be flagged for further investigation or have their access reduced.

Importantly, interventions of this kind are often most effective when their details (and sometimes even their existence) are not widely shared. Avoiding detection is easier if you know what the detection procedure is. The effectiveness of speed cameras is greatly reduced if drivers know where they are located. Similarly, the legal systems will often include intentional ambiguity and room for judgment to allow courts, law enforcement, and regulators to enforce the spirit rather than the letter of the law. If the rules and means of detecting a breach are made overly precise, actors can make sure to precisely skirt the line, avoiding detection while still being able to carry out misuse. This is analogous to how tax authorities around the world tend not to give precise details about their methods of detecting tax fraud.

7 Case studies

Below, we discuss three case studies of AI misuse where targeted capabilities restrictions may be warranted: AI systems used to predict toxins, image generation models being used to create harmful content, and LLMs being used for spear phishing campaigns.

7.1 Toxin prediction

In 2021, two researchers created a list of novel toxins using MegaSyn, a machine learning based de novo molecule generator used for drug discovery (Urbina et al. 2022b). Normally, MegaSyn penalizes expected toxicity and rewards expected bioactivity. But the researchers wondered what would happen if they flipped the filter on MegaSyn—literally by swapping a '1' for a '0' and a '0' for a '1'—to produce toxic molecules. After running the modified system on a 2015 MacBook for a few hours, a list of over 40,000 toxins was produced, some of which were predicted to be more lethal than publicly known chemical warfare agents such as VX. Concerningly, the pair of researchers created MegaSyn with publicly available data and software (Urbina et al. 2022a).

While certain forms of capabilities restrictions might not entirely prevent actors from misusing these models, interventions at the capabilities stage might nonetheless be desirable. With chemical weapons, minimizing harm seems difficult. The attack surface is vast and harm can be severe. Responding to misuse might be possible, yet such sanctions would ideally occur at the earliest possible stages of misuse, such as when an actor is planning their attack. Given the high stakes, the ineffectiveness of interventions at later stages of the Misuse Chain, and the potential to reduce the dual-use nature of these sorts of AI systems, various capabilities restrictions appear warranted.

For example, structured access schemes could make these models less dual-use. Making it so that certain drug discovery models can only be accessed via an API would enable filters that prohibit outputs of compounds above some threshold of toxicity. Legitimate uses of such outputs—e.g., by researchers seeking to develop countermeasures to novel toxins—could be enabled by an approval process. Additionally, structured access schemes could enable the model's owners to keep a log of concerning outputs, and who created them. This could allow law enforcement to preempt attacks as well as track down perpetrators after harm occurs. Even if rogue actors could circumvent these restrictions by training their own models, reducing the number of actors capable of causing harm is still desirable. Approaches along these lines would score particularly highly on *targetedness*: a wholesale ban on these systems would likely not be warranted given the undesired side effects on beneficial use cases. However, structured access schemes achieve a higher targetedness ratio by selectively discriminating against malicious use while still allowing for scientific progress to be made.

Other non-AI interventions at the capabilities stage likely play an even more important role. For example, requirements could be made for companies providing chemical synthesis services to screen orders and cooperate with law enforcement in the event that a malicious actor attempts to order highly dangerous toxins or precursors thereof. Governance structures could also target the technologies and materials used to synthesize the resulting chemical agents to ensure they cannot be misused. There are existing efforts to prevent the production of large quantities of toxins such as the chemical weapons convention. These efforts, and others seeking to prevent the widespread dispersal of harmful toxins, might need to be updated and strengthened as the number and severity of novel toxins increases.

7.2 Harmful image generation

Certain images—such as pornographic images of someone's likeness produced without their consent and fake images intended to mislead public opinion—can cause harm. While it has long been possible to use tools like Photoshop to manipulate images, there was less need for intervention in the past. The production of highly realistic fake images was time-consuming and the resulting images would typically reach a limited audience. However, AI systems capable of producing photorealistic images significantly lower the barriers to producing such images. Paired with the increased reach that images can have when disseminated on the internet and social media platforms, such content can cause significantly more harm than before.

One way of preventing harm from AI-generated images is to prevent malicious actors from acquiring the capabilities needed to generate such images in the first place. However, some interventions aimed at ensuring generative models do not produce harmful images are fraught with difficulties. Take for example the strategy of introducing content filters on natural language prompts used for image generation models. Getting these filters right is difficult: such filters are likely to be both underinclusive (e.g., images with violent content can be generated by indirect prompting, such as "a horse lying on its side in a puddle of red liquid") and overinclusive (e.g., perhaps disallowing any prompt that includes the phrase "breast stroke"). As a result, these kinds of post-hoc filters could hinder legitimate use while still being vulnerable to exploitation by malicious actors. At the moment, they most likely produce many more false positives than true positives, lowering their Targetedness Ratio. However, over time, this Targetedness Ratio might improve, as more advanced classifiers are developed, taking not only the prompt, but also the output of the model into account.

Nevertheless, these challenges are not sufficient to rule out all interventions at the capabilities stage. The interventions at the capabilities stage can still amend misuse by improving the efficacy of other interventions at later stages of the Misuse Chain. For example, owners of an image-generation model that is queried via an API could insert invisible watermarks into the model's outputs. These watermarks could help social media platforms detect AI-generated content more reliably. Fingerprints could also be installed directly within image generation models. For example, researchers have demonstrated a scalable technique that applies a unique fingerprint to each image produced by a particular copy of a model's weights, so misuse can be attributed back to specific users (Yu et al. 2022), and Google DeepMind has launched SynthID, a novel watermarking method for AI-generated text and video (Google DeepMind 2024).

With improved detection capabilities, platforms could mitigate harm by labeling AI-generated content as such or remove media that violates their terms of service. The proposed EU AI Act includes related provisions, requiring posters of AI generated content to disclose it as such (European Commission 2020). Some platforms, like Facebook (Bikert 2020) and Twitter (Twitter 2023), already have policies to remove certain manipulated media, but improved AI detection could strengthen their ability to enforce these policies.

Finally, model creators could train highly capable classifier models that are designed to attribute content to the model that generated it, even if these models are open-source. How accurate such classifiers are, and how robust watermarks are to attempts at removal is an active and crucial research question. However, by keeping these classifier models hidden behind APIs, malicious actors could face more difficulties attempting to reliably evade detection. To prevent misuse, actors should adhere to a norm of only releasing generative models broadly once sufficient safeguards are in place, such as capable detection systems.

7.3 Spear phishing

Using AI systems for phishing campaigns could exacerbate threats to nation states, organizations, and individuals. In particular, LLMs could enhance hackers' ability to "spear phish", a tactic in which customized communication is sent from an ostensibly trustworthy source in order to trick recipients into revealing sensitive information. The analogy of a sniper is often invoked when describing spear phishing, as it is a highly targeted and precise method of attack, whereas ordinary phishing is seen as a broad and indiscriminate attack, akin to a shotgun blast. Jeh Johnson, former U.S. Secretary of the Department of Homeland Security, was quoted as saying that "the most devastating, intrusive attacks by the most sophisticated actors often originate with a simple act of spear phishing" (Hirschfeld Davis 2016). In a survey from 2022, 75% of security professionals named phishing attacks as a top threat to cyber-security (Powell 2022), and Microsoft has recently claimed that Russia and China have used tools (unwillingly) provided by OpenAI for hacking and cyber-attacks, including spear phishing (Roush 2024).

Spear phishing is traditionally a time-consuming and labor-intensive process that can involve several steps such as identifying high-value targets, conducting personalized research to gather relevant information on the target, and crafting a tailored message that appears to come from a trusted acquaintance. However, with the integration of AI, this process can be made more efficient (Heiding/Schneier/Vishwanath 2024). Even relatively simple AI systems can improve attackers' efficiency. For example, in 2018 researchers created an automated spear phishing system called SNAP_R that used a long short-term memory neural network to send phishing tweets tailored to targets' characteristics (Seymur/Tully 2018). Though the tweets were typically short and unsophisticated, SNAP_R could send them significantly faster than a human operator and with a similar click-through rate in a small experiment. Compared to the models used to create SNAP_R, large transformer-based language models are significantly more capable of generating human-sounding text. There are clear indications of large language models being used for spear phishing, as evidenced by discussions on dark web forums for cybercriminals (Insikt Group 2023). OpenAI has terminated accounts from groups that were suspected to use their models to engage in malicious acts, including spear phishing (OpenAI 2024).

Despite the potential LLMs afford for improving spear phishing (Hazell 2023), intervening at the capabilities stage could pose challenges. First, achieving a sufficient *Targetedness Ratio* is a concern: it would be difficult to tell whether a given piece of text is intended to be used for spear phishing, or a similar sounding, non-malicious form of communication, such as a marketing email. Preventing a language model from outputting such text would be technically difficult, and could noticeably interfere with positive use. Second, malicious actors with enough motivation would likely be able to fine-tune the model to output such messages nonetheless. The *value ratio* seems similarly hard to define. While the positive use value of spear phishing itself is near zero, the tools that are increasingly used to conduct these operations (LLMs) harbor enormous potential and positive value, calling for a more nuanced approach and targeted interventions.

However, some intervention at the capabilities stage may still be warranted. New threats will require modifications to existing cybersecurity systems, and perhaps the creation of entirely new forms of defense. Crucially, certain capabilities interventions can make LLMs less capable of being used for spear phishing, and can boost interventions at later stages of the Misuse Chain. For example, one solution could be to ensure that advanced language models are only queryable via a structured access scheme, such as an API. Similar to interventions aimed at preventing harmful images from being generated, suspicious requests could be flagged and logged in a database. Users who consistently produce text that could be used for spear phishing could then be investigated or subject to usage restrictions by the model's owners, increasing the *Targetedness Ratio* by restricting limitations to repeat offenders.

Being able to identify AI generated content as such would be helpful for mitigating harm. A recent classifier produced by OpenAI showed a 26% true positive rate and a 9% false positive rate for AI generated content (Kirchner et al. 2023), which may not be sufficiently targeted, and others have voiced their doubts as to whether strong and robust watermarking for text is possible at all (Edelman et al. 2023). However, research is exploring methods for watermarking AI generated text by making the model biased in favor of certain word choices or combinations thereof that a classifier, but not a human, would be able to detect (Aaronson 2022). Though this method might be circumventable via having another AI model paraphrase the output, it may stop less sophisticated actors. Further, such paraphrasing attempts could also potentially be thwarted by sharing the classifier with other LLM providers, allowing them to check whether users are using the system to remove watermarks. Others have suggested a different approach, where open-source detectors are distributed (Kirchenbauer et al. 2023).

Nevertheless, the most effective interventions for preventing spear phishing may focus on mitigating harm, rather than intervening on capabilities. Harm from AI-generated spear phishing could be mitigated by improving existing systems aimed at stopping spear phishing. For example, systems can use LLMs to scan and then flag suspicious messages, taking into account various metadata such as whether the sender is using an email similar to someone already in the target's contact list. These systems do not need to necessarily tell if an incoming message is AI-generated. On the contrary, identifying and protecting against harm, regardless of whether or not it is AI generated, could be a more robust defensive strategy over the long run. More research is needed to determine whether spear phishing attacks or defenses against them will gain more from advances in these technologies.

8 Conclusion

AI systems are already being misused across various domains and as they become more capable and are deployed more broadly, the potential for misuse will grow. The decision-makers will feel compelled to intervene on such misuses, but choosing the right suite of interventions can be difficult. Interventions inevitably face the Misuse–Use Tradeoff. Despite this tradeoff, we argue that interventions aimed at the capabilities stage of the Misuse Chain will be increasingly warranted as the potential harms of AI misuse increase (increasing the *value ratio* substantially in some domains), as AI misuse becomes difficult to defend against in the other stages of the Misuse Chain, and as new techniques are created that can increase the *Targetedness Ratio* of capability interventions.

To better prepare society for managing AI misuse, we encourage future research on a number of questions, including:

- Determining the potential harm of high-risk misuses and what interventions will be warranted. What misuses of AI are high-risk? In high-risk domains, how much harm could be caused by AI misuse? Given this, what interventions are warranted?
- Generating empirical estimates of Misuse–Use Tradeoffs. How can we estimate the Misuse–Use Tradeoff for interventions such as filters used for image-generation models and LLMs?
- Developing techniques to help defend against misuse. What systems can most effectively prevent misuse while favorably navigating the Misuse–Use Tradeoff?

Data availability Not applicable.

References

Aaronson S (2022) My AI Safety Lecture for UT effective altruism. https://scottaaronson.blog/?p=6823

Abraham Y (2024) 'Lavender': The AI machine directing Israel's bombing spree in Gaza. +972 Magazine. https://www.972mag.com/lavender-ai-israeli-army-gaza/

Adler MD, Posner EA (2000) Implementing cost-benefit analysis when preferences are distorted. J Legal Stud 29(S2):1105–1147

AG Eshoo (2022a) Eshoo Urges NSA & OSTP to Address Unsafe AI Practices. https://eshoo.house.gov/media/press-releases/eshoo-urges-nsa-ostp-address-unsafe-ai-practices

AG Eshoo (2022b) Eshoo Urges NSA & OSTP to Address Biosecurity Risks Caused by AI. Retrieved from https://eshoo.house.gov/media/press-releases/eshoo-urges-nsa-ostp-address-biosecurity-risks-caused-ai

Al-Dosari K, Fetais N, Kucukvar M (2024) Artificial Intelligence and cyber defense system for banking industry: a qualitative study of AI applications and challenges. Cybern Syst 55(2):302–330

Alexander L (2020) Deontological Ethics. Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/ethics-deontological/

Alexander S (2023) Mostly skeptical thoughts on the chatbot propaganda apocalypse. https://astralcodexten.substack.com/p/mostly-skeptical-thoughts-on-the

Altman S. Planning for AGI and beyond. https://openai.com/blog/planning-for-agi-and-beyond. Accessed 13 Mar 2023

Amazon (2023) Amazon Global Human Rights Principles. https://sustainability.aboutamazon.co.uk/society/human-rights/principles

Anthropic. Core Views on AI Safety: When, Why, What, and How. https://www.anthropic.com/index/core-views-on-ai-safety. Accessed 13 Mar 2023

Ashurst C, Barocas S, Campbell R, Deborah Raji D (2022) Discovering the components of ethical research in machine learning. In: Proceedings of the FAccT '22: 2022 ACM Conference on Fairness, Accountability. https://doi.org/10.1145/3531146.3533781

Askell A, Bai Y, Chen A, Drain D, Ganguli D, Henighan T, Jones A, Joseph N, Mann B, DasSarma N, Elhage N, Hatfield-Dodds Z, Hernandez D, Kernion J, Ndousse K, Olsson C, Amodei D, Brown T, Clark J, McCandlish S, Olah C, Kaplan J (2021) A general language assistant as a laboratory for alignment. https://doi.org/10.48550/arXiv.2112.00861

Bai Y, Kadavath S, Kundu S, Askell A, Kernion J, Jones A, Chen A, Goldie A, Mirhoseini A, McKinnon C, Chen C, Olsson C, Olah C, Hernandez D, Drain D, Ganguli D, Li D, Tran-Johnson E, Perez E, Kerr J, Mueller J, Ladish J, Landau J, Ndousse K, Lukosuite K, Lovitt L, Sellitto M, Elhage N, Schiefer N, Mercado N, DasSarma N, Lasenby R, Larson R, Ringer S, Johnston S, Kravec S, El Showk S, Fort S, Lanham T, Telleen-Lawton T, Conerly T, Henighan T, Hume T, R. Bowman SR, Hatfield-Dodds Z, Mann B, Amodei D, Joseph N, McCandlish S, Brown T, and Kaplan J (2022) Constitutional AI: Harmlessness from AI Feedback. https://doi.org/10.48550/arXiv.2212.08073

Baldwin R, Cave M, Lodge M (2011) Cost-Benefit Analyses and Regulatory Impact Assessment. In: Understanding Regulation: Theory, Strategy and Practice. https://academic.oup.com/book/7235/chapter-abstract/151924383?redirectedFrom=fulltext

Bartrum O (2023) Energy regulation requires tradeoffs the regulator cannot make alone. Institute for government: https://www.instituteforgovernment.org.uk/comment/energy-regulation-trade-offs

Bernardi J, Mukobi G, Greaves H, Heim L, Anderljung M (2024) Societal adaptation to advanced AI. https://arxiv.org/abs/2405.10295

Better Markets (2023) The ongoing use and abuse of cost-benefit analysis in financial regulation. https://bettermarkets.org/analysis/the-ongoing-use-and-abuse-of-cost-benefit-analysis-in-financial-regulation/

Bikert M (2020) Enforcing Against Manipulated Media. Meta Newsroom (2020). https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/

Blauth TF, Gstrein OJ, Zwitter A (2022) Artificial intelligence crime: an overview of malicious use and abuse of AI. In IEEE Access, vol. 10, 77110–77122, 2022, https://doi.org/10.1109/ACCESS.2022.3191790.

Bloch-Wehba H (2020) Automation in moderation. Cornell Int Law J 53(2020):41–96

Boardman AE, Greenberg DH, Vining AR, Weimer DL (2018) Cost benefit analysis: concepts and practice, 5th edn. United States, Sheridan Books

Bronsteen J, Buccafusco C, Masur JS (2013) Well-Being Analysis vs. Cost-Benefit Analysis. Duke Law Journal 1603–1689

Brundage M, Avin S, Clark J, Toner H, Eckersley P, Garfinkel B, Dafoe A, Scharre P, Zeitzoff T, Filar B, Anderson H, Roff H, Allen GC, Steinhardt J, Flynn C, hÉigeartaigh SÓ , Beard S, Belfield H, Farquhar S, Lyle C, Crootof R, Evans O, Page M, Bryson J, Yampolskiy R, Amodei D (2018) The malicious use of artificial intelligence: forecasting, prevention, and mitigation. https://maliciousaireport.com

Brundage M, Avin S, Wang J, Belfield H, Krueger G, Hadfield G, Khlaaf H, Yang J, Toner H, Fong R, Maharaj T, Koh PW, Hooker S, Leung J, Trask A, Bluemke E, Lebensold J, O'Keefe C, Koren M, Ryffel T, Rubinovitz JB, Besiroglu T, Carugati F, Clark J, Eckersley P, de Haas S, Johnson M, Laurie B, Ingerman A, Krawczuk I, Askell A, Cammarota R, Lohn A, Krueger D, Stix C, Henderson P, Graham L, Prunkl C, Martin B, Seger E, Zilberman N, hÉigeartaigh SÓ, Kroeger F, Sastry G, Kagan R, Weller A, Tse B, Barnes E, Dafoe A, Scharre P, Herbert-Voss A, Rasser M, Sodhani S, Flynn C, Gilbert TK, Dyer L, Khan S, Bengio Y, Anderljung M (2020) Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. https://doi.org/10.48550/ARXIV.2004.07213

Bureau of Industry and Security (2022) Commerce implements new export controls on advanced computing and semiconductor manufacturing items to the People's Republic of China (PRC. https://www.bis.doc.gov/index.php/documents/about-bis/newsroom/press-releases/3158-2022-10-07-bis-press-release-advanced-computing-and-semiconductor-manufacturing-controls-final/file

Burgan C (2024) Senator Warnley 'gravely concerned' for AI misuse in 2024 elections. https://www.meritalk.com/articles/sen-warner-gravely-concerned-for-ai-misuse-in-2024-elections/

Canadian Government 2023 Canada's Cost-Benefit Analysis Guide for Regulatory Proposals. : https://www.canada.ca/en/government/system/laws/developing-improving-federal-regulations/requirements-developing-managing-reviewing-regulations/guidelines-tools/cost-benefit-analysis-guide-regulatory-proposals.html

Chaudhry H, Klein L (2024) Chemical & Biological Weapons and Artificial Intelligence: Problem Analysis and US Policy Recommendations. Future of Life Institute. https://futureoflife.org/wp-content/uploads/2024/02/FLI_AI_and_Chemical_Bio_Weapons.pdf

Cambridge Consultants 2019 Use of AI in Online Content Moderation. https://www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf

Check Point Research. 2023. OPWNAI: Cybercriminals Starting to Use ChatGPT. https://research.checkpoint.com/2023/opwnai-cybercriminals-starting-to-use-chatgpt/

Coates J (2015) Cost-benefit analyses of financial regulation: case studies and implications. Yale Law Journal 124(4):882–1345

Congress US (1986) Computer Fraud and Abuse Act of 1986. Pub L 1986:99–474

Congressional Research Service (CRS). 2022. Cost-Benefit Analyses in Federal Agency Rulemaking. https://crsreports.congress.gov/product/pdf/IF/IF12058

Cui J, Chiang W-L, Stoica I, Hsieh C-J (2024) OR-Bench: An Over-Refusal Benchmark for Large Language Models. https://arxiv.org/html/2405.20947v2

Danzig R, Sageman M, Leighton T, Hough L, Yuki H, Kotani R, Hosford ZM (2012) Aum Shinrikyo. Insights into how terrorists develop biological and chemical weapons. Center for a New American Security. https://www.jstor.org/stable/pdf/resrep06323.pdf

Davis JH (2016) U.S. seeks to protect voting system from cyberattacks. The New York Times (2016). https://www.nytimes.com/2016/08/04/us/politics/us-seeks-to-protect-voting-system-against-cyberattacks.html

Deborah G. Johnson, and Mario Verdicchio. 2024. The sociotechnical entanglement of AI and values. AI & Society. Retrieved February 5th, 2024 from https://link.springer.com/article/https://doi.org/10.1007/s00146-023-01852-5.

Department of Commerce (2022) Implementation of Additional Export Controls: Certain Advanced Computing and Semiconductor Manufacturing Items; Supercomputer and Semiconductor End Use; Entity List Modification. https://www.federalregister.gov/documents/2022/10/13/2022-21658/implementation-of-additional-export-controls-certain-advanced-computing-and-semiconductor

Department of Science, Innovation and Technology (DSIT) (2023) Emerging processes for frontier AI Safety. https://www.gov.uk/government/publications/emerging-processes-for-frontier-ai-safety/emerging-processes-for-frontier-ai-safety

D.J.I. 2023. Fly Safe Geo Zone Map. https://www.dji.com/flysafe/geo-map

European Commission (2021) Better Regulation. Joining forces to make better laws. https://commission.europa.eu/document/download/199176cf-6c4e-48ad-a9f7-9c1b31bbbd09_en?filename=better_regulation_joining_forces_to_make_better_laws_en.pdf

European Commission (2022) Disinformation: Commission Welcomes the New Stronger and More Comprehensive Code of Practice in Disinformation. https://ec.europa.eu/commission/presscorner/detail/en/ip_22_3664

Edelman B, Zhang H, Barak B (2023) Watermarking in the sand. https://kempnerinstitute.harvard.edu/research/deeper-learning/watermarking-in-the-sand/

Egan J, Heim L (2023) Oversight for frontier AI through a Know-Your-Customer Scheme for Compute Providers. Centre for the Governance of AI. https://www.governance.ai/research-paper/oversight-for-frontier-ai-through-kyc-scheme-for-compute-providers

European Parliament (2024) Artificial Intelligence Act. https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf

Ferraro MF, Chipman JC, Preston SW (2019) First Federal Legislation on Deepfakes Signed Into Law. https://www.wilmerhale.com/en/insights/client-alerts/20191223-first-federal-legislation-on-deepfakes-signed-into-law

Financial Conduct Authority (FCA) (2018) How we analyse the costs and benefits of our policies. https://www.fca.org.uk/publication/corporate/how-analyse-costs-benefits-policies.pdf

Fischer S-C, Leung J, Anderljung M, O'Keefe C, Torges S, Khan SM, Garfinkel B, Dafoe A (2021) AI policy levers: A review of the U.S. Government's tools to shape AI research, development, and deployment. https://www.fhi.ox.ac.uk/wp-content/uploads/2021/03/AI-Policy-Levers-A-Review-of-the-U.S.-Governments-tools-to-shape-AI-research-development-and-deployment-

Gade P, Lermen S, Rogers-Smith C, Ladish J (2023) BadLlama: cheaply removing safety fine-tuning from Llama 2-Chat 13B. https://arxiv.org/abs/2311.00117

Ganguli D, Hernandez D, Lovitt L, Askell A, Bai Y, Chen A, Conerly T, Dassarma N, Drain D, Elhage N, El-Showk S, Fort S, Hatfield-Dodds Z, Henighan T, Johnston S, Jones A, Joseph N, Kernian J, Kravec S, Mann B, Nanda N, Ndousse K, Olsson C, Amodei D, Brown T, Kaplan J, McCandlish S, Olah C, Amodei D, Clark J (2022) Predictability and Surprise in Large Generative Models. In: 2022 ACM Conference on Fairness, Accountability, and Transparency, ACM. https://doi.org/10.1145/3531146.3533229

Garfinkel B, Dafoe A (2019) How does the offense-defense balance scale? J Strat Stud 42(6):736–763. https://doi.org/10.1080/01402390.2019.1631810

Gebhard C (2024) Bipartisan measure prohibiting AI interference in elections introduced by Pennycuick, Gebhard, Dillon, Kane. https://senatorgebhard.com/2024/05/21/bipartisan-measure-prohibiting-ai-interference-in-elections-introduced-by-pennycuick-gebhard-dillon-kane/

Giantini G (2023) The sophistry of the neutral tool. Weaponizing artificial intelligence and big data into threats toward social exclusion. AI & Ethics 3:1049–1061

Goldstein JA, Sastry G, Musser M, DiResta R, Gentzel M, Sedova K (2023) Generative Language models and automated influence operations: emerging threats and potential mitigations. Arxiv. https://doi.org/10.48550/ARXIV.2301.04246

Google DeepMind (2024) Watermarking AI-generated text and video with SynthID. https://deepmind.google/discover/blog/watermarking-ai-generated-text-and-video-with-synthid/

Gorwa R, Binns R, Katzenbach C (2020) Algorithmic content moderation: technical and political challenges in the automation of platform governance. Big Data Soc. https://doi.org/10.1177/2053951719897945

Government Accountability Office (GAO) (1978) Costs and Benefits of Governmental Regulation. https://www.gao.gov/assets/107970.pdf

Gross JA (2021) In apparent world first, IDF deployed drone swarms in Gaza fighting. The Times of Israel (2021). https://www.timesofisrael.com/in-apparent-world-first-idf-deployed-drone-swarms-in-gaza-fighting/

Hambling D (2021) Israel used world's first AI-guided combat drone swarm in Gaza attacks. *New Scientist* (2021). https://www.newscientist.com/article/2282656-israel-used-worlds-first-ai-guided-combat-drone-swarm-in-gaza-attacks/

Han S, Rao K, Ettinger A, Jiang L, Lin BY, Lambert N, Choi Y, Dziri N (2024) WildGuard: Open One-Stop Moderation Tools for Safety Risks, Jailbreaks and Refusals of LLMs. https://arxiv.org/html/2406.18495v1

Hazell J (2023) Spear Phishing with Large Language Models. The Centre for the Governance of AI. https://www.governance.ai/research-paper/llms-used-spear-phishing

Hecht B, Lauren Wilcox, Jeffrey P. Bigham, Johannes Schöning, Ehsan Hoque, Jason Ernst, Yonatan Bisk, Luigi Russis, Lana Yarosh, Bushra Anjam, Danish Contractor, and Cathy Wu. 2018. It's Time to Do Something: Mitigating the Negative Impacts of Computing Through a Change to the Peer Review Process. ACM Future of Computing Blog (2018). https://acm-fca.org/2018/03/29/negativeimpacts/

Heiding F, Schneier B, Vishwanath A, Bernstein J, Park PS (2024b) Devising and detecting phishing emails using large language models. IEEE Access 12:42131–42146. https://doi.org/10.1109/ACCESS.2024.3375882

Heiding F, Schneier B, Vishwanath A (2024) AI will increase the quantity - and quality - of phishing scams. Harvard Business Review. https://hbr.org/2024/05/ai-will-increase-the-quantity-and-quality-of-phishing-scams

Heikkilä M (2023) Google DeepMind has launched a watermarking tool for AI-generated images. MIT Technology Review. https://www.technologyreview.com/2023/08/29/1078620/google-deepmind-has-launched-a-watermarking-tool-for-ai-generated-images/

Herzog N, Celik D, Sulaiman RB (2024) Artificial Intelligence in Health Care and Medical Records Security. Cybersecurity in Artificial Intelligence. In: Jahankhani H, Bowen G, Sharif MS, Hussien O (eds) Cybersecurity and Artificial Intelligence. Advanced Sciences and Technologies for Security Applications. Springer. https://doi.org/10.1007/978-3-031-52272-7_2

Hiller A (2017) Consequentialism in Environmental Ethics. In: Gardiner SM, Thompson A (eds) Oxford Handbook of Environmental Ethics. Oxford University Press

Horvitz E (2022) On the horizon: interactive and compositional deepfakes (2022) https://doi.org/10.48550/ARXIV.2209.01714

Insikt Group (2023) I, Chatbot. https://www.recordedfuture.com/i-chatbot

Jiang Z, Zhang J, Gong NZ (2023) Evading Watermark based Detection of AI-Generated Content. In: Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23). Association for Computing Machinery, New York, NY, USA, 1168–1181. https://doi.org/10.1145/3576915.3623189

Joint Counterterrorism Assessment Team (JCAT) (2022) Emerging technologies may heighten Terrorist Threats. https://www.odni.gov/files/NCTC/documents/jcat/firstresponderstoolbox/134s_-_First_Responders_Toolbox_-_Emerging_Technologies_May_Heighten_Terrorist_Threats.pdf

Kagan S (1998) Normative Ethics. Routledge, New York

Kirchenbauer J, Geiping J, Wen Y, Katz J, Miers I, Goldstein T (2023) A Watermark for Large Language Models. https://arxiv.org/abs/2301.10226

Kirchner JH, Ahmad L, Aaronson S, Leike J (2023) New AI Classifier for Indicating AI-Written Text. OpenAI (2023). https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text

K Wiggers (2022) Deepfakes for all: uncensored ai art model prompts ethics questions. Tech Crunch (2022). https://techcrunch.com/2022/08/24/deepfakes-for-all-uncensored-ai-art-model-prompts-ethics-questions/

Krishnamurthy V (2022) With great (computing) power comes great (human rights) responsibility: cloud computing and human rights. Bus Hum Rights J 7(2):226–248. https://doi.org/10.1017/bhj.2022.8

Landers L, Couvillion C, Refuerzo N (2024) A 15-year-old's prom picture was altered into AI-created nudes. 23ABC Bakersfield. https://www.turnto23.com/politics/disinformation-desk/high-schools-nationwide-are-facing-a-new-problem-ai-generated-nudes

Leike J, Schulman J, Jeffrey WU (2023) Our approach to alignment research. https://openai.com/blog/our-approach-to-alignment-research. Accessed 13 Mar 2023

Llansó E, Hoboken J, Leerssen P, Harambam J (2020) Artificial intelligence, content moderation, and freedom of expression. Transatlantic Working Group on Content Moderation Online and Freedom of Expression Working Paper (2020). https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf

MacCarthy M (2020) AI Needs More Regulation, Not Less. Brookings (2020). https://www.brookings.edu/research/ai-needs-more-regulation-not-less/

Microsoft (2024) Protecting the public from abusive AI-generated content. White Paper. https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW1nuJx

Mishkin P, Ahmad L, Brundage M, Krueger G, Sastry G (2022) DALL·E 2 Preview - Risks and Limitations. https://github.com/openai/dalle-2-preview/blob/main/system-card.md

Mouton C, Lucas C, Guest E (2024) The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study. RAND Corporation. https://www.rand.org/pubs/research_reports/RRA2977-2.html.

Nickel D (2024) AI is shockingly good at making fake nudes - and causing havoc in schools. POLITICO. https://www.politico.com/news/2024/05/28/ai-deepfake-nudes-schools-states-00160183

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

OpenAI (2024) Disrupting malicious uses of AI by state-affiliated threat actors. https://openai.com/index/disrupting-malicious-uses-of-ai-by-state-affiliated-threat-actors/

Organisation for the Prohibition of Chemical Weapons. 2023. Definitions and Criteria. https://www.opcw.org/chemical-weapons-convention/articles/article-ii-definitions-and-criteria

Parfit D (1984) Reasons and persons. Oxford University Press

Partnership on AI (2021) Managing the Risks of AI Research: Six Recommendations for Responsible Publication. http://partnershiponai.org/wp-content/uploads/2021/08/PAI-Managing-the-Risks-of-AI-Resesarch-Responsible-Publication.pdf

Permanent Subcommittee on Investigations (2019) Abuses of the Federal Notice-and-Comment Rulemaking Process. https://web.archive.org/web/20200115005122/https:/www.portman.senate.gov/sites/default/files/2019-10/2019.10.24%20PSI%20Report%20-%20Abuses%20of%20the%20Federal%20Notice-and-Comment%20Rulemaking%20Process.pdf

Peterson D, Hoffman S (2022) Geopolitical implications of AI and digital surveillance adoption. Brookings Institution. https://www.brookings.edu/research/geopolitical-implications-of-ai-and-digital-surveillance-adoption/

Peterson D (2020) Designing alternatives to China's repressive surveillance state. CSET Policy Brief (2020). https://cset.georgetown.edu/wp-content/uploads/CSET-Designing-Alternatives-to-Chinas-Surveillance-State.pdf

Porter TM (2020) Trust in numbers. The pursuit of objectivity in science and public life, Princeton University Press

Powell R (2003) Nuclear deterrence theory, nuclear proliferation, and national missile defense. Int Secur 27(4):86–118. https://doi.org/10.1162/016228803321951108

Powell O (2022) Social engineering "most dangerous" threat, say 75% of security professionals. Research by CS Hub has revealed that social engineering and phishing attacks are the top threat to cyber security. Cyber Security Hub. https://www.cshub.com/attacks/news/social-engineering-most-dangerous-threat-say-75-of-security-professionals

Prunkl C, Ashurst C, Anderljung M, Webb H, Leike J, Dafoe A (2021) Institutionalising ethics in AI through broader impact requirements. https://arxiv.org/abs/2106.11039

Rao JM, Reiley DH (2012) The economics of spam. J Econ Perspect 26(3):87–110

Rogeberg O (2018) Prohibition, regulation or laissez faire: the policy trade-offs of cannabis policy. Int J Drug Policy. https://doi.org/10.1016/j.drugpo.2018.03.024

Rose J (2023) Children sex abuse material was found in a major AI dataset. Researchers aren't surprised. https://www.vice.com/en/article/3aky5n/child-sex-abuse-material-was-found-in-a-major-ai-dataset-researchers-arent-surprised

Roush T (2024) Microsoft Claims Russia, China And Others Used OpenAI's Tools For Hacking. Forbes. Retrieved from: https://www.forbes.com/sites/tylerroush/2024/02/14/microsoft-claims-russia-china-and-others-used-openais-tools-for-hacking/?sh=27a1405e204c

Rowe G, Wright G (2001) Expert opinions in forecasting: the role of the Delphi technique. https://www3.nd.edu/~busiforc/handouts/Other%20Articles/expertopinions.pdf

Rubinic I, Kurtov M, Rubinic I, Likic R, Dargan P, Wood D (2024) Artificial intelligence in clinical pharmacology: a case study and scoping review of large language models and bioweapon potential. BJCP 90(3):620–828

Sastry G, Heim L, Belfield H, Anderljung M, Brundage M, Hazell J, O'Keefe C, Hadfield GK, Ngo R, Pilz K, eorge Gor G, Bluemke E, Shoker S, Egan J, Trager RF, Avin S, Weller A, Bengio Y, Coyle D (2024) Computing Power and the Governance of Artificial Intelligence. Retrieved from https://arxiv.org/abs/2402.08797. John Seymour and Philip Tully. 2018. Generative models for spear phishing posts on social media. Retrieved from https://arxiv.org/abs/1802.05196

Sandbrink J, Hobbs H, Swett J, Dafoe A, Sandberg A (2022) Differential Technology Development: A Responsible Innovation Principle for Navigating Technology Risks. SSRN Journal (2022). https://doi.org/10.2139/ssrn.4213670

T Shevlane, A Dafoe (2020) The offense-defense balance of scientific knowledge: Does publishing AI research reduce misuse? arXiv:2001.00463v2. https://arxiv.org/pdf/2001.00463.pdf

Shevlane T (2022) Structured access: an emerging paradigm for safe AI deployment. https://arxiv.org/abs/2201.05159

Smith B (2023) Microsoft Global Human Rights Statement. https://www.microsoft.com/en-us/corporate-responsibility/human-rights-statement?activetab=pivot_1%3aprimaryr5

Solaiman I (2023) The gradient of generative AI release: Methods and considerations. https://arxiv.org/pdf/2302.04844.pdf

Srinivasan S (2024) Detecting AI fingerprints: A guide to watermarking and beyond. Brookings Institution. https://www.brookings.edu/articles/detecting-ai-fingerprints-a-guide-to-watermarking-and-beyond/

Stability AI (2023) Stable Diffusion 2.0 Release. https://stability.ai/blog/stable-diffusion-v2-release

Stockwell S, Hughes M, Swatton P, Bishop K (2024) AI-enabled influence operations: the threat to the UK general election. Turing Institute

Sweis N (2022) Revisiting the value of a statistical life: an international approach during COVID-19. Risk Manag 24(3):259–272

Taddeo M, Floridi L (2018) Regulate artificial intelligence to avert cyber arms race. Nature 556(7701):296–298. https://doi.org/10.1038/d41586-018-04602-6

Tomasovic B (2018) Tradeoffs in environmental law. Journal of Land Use & Environmental Law 34, 1, 93–150. JSTOR, https://www.jstor.org/stable/26896699

Trager RF, Luca M (2022) Killer robots are here—and we need to regulate them. Foreign Policy (2022). https://foreignpolicy.com/2022/05/11/killer-robots-lethal-autonomous-weapons-systems-ukraine-libya-regulation/

Treasury HM (2022) The Green Book. https://www.gov.uk/government/publications/the-green-book-appraisal-and-evaluation-in-central-government/the-green-book-2020

Twitter (2023) Synthetic and manipulated media policy. https://help.twitter.com/en/rules-and-policies/manipulated-media

United Nations (2022) OHCHR Assessment of human rights concerns in the Xinjiang Uyghur Autonomous Region, People's Republic of China. (2022). https://www.ohchr.org/sites/default/files/documents/countries/2022-08-31/22-08-31-final-assesment.pdf

Urbina F, Lentzos F, Invernizzi C, Ekins S (2022b) Dual use of artificial-intelligence-powered drug discovery. Nat Mach Intell 4(3):189–191. https://doi.org/10.1038/s42256-022-00465-9

Urbina F, Lentzos F, Invernizzi C, S Ekins (2022a) A teachable moment for dual-use. *Nat Mach Intell* 4, 607, (2022). https://www.nature.com/articles/s42256-022-00511-6

US Congress (1974) Fair credit billing act. 15 U.S.C. § 1666, (1974).

Walden A (2022) Our Ongoing Commitment to Human Rights. Google (2022). https://blog.google/outreach-initiatives/public-policy/our-ongoing-commitment-to-human-rights/

Weatherbed J (2023) Twitter replaces its free API with a paid tier in quest to make more money. The Verge. https://www.theverge.com/2023/2/2/23582615/twitter-removing-free-api-developer-apps-price-announcement

Webb E (2024) The imminent crisis of deepfake porn. Liberty University. https://digitalcommons.liberty.edu/research_symp/2024/oral_presentations/80/

Weiss M (2019) Deepfake bot submissions to federal public comment websites cannot be distinguished from human submissions. Technol Sci 2019121801, (2019). https://techscience.org/a/2019121801/

Yu N, Skripniuk V, Chen D, Davis LE, Fritz M (2022) Responsible disclosure of generative models using scalable fingerprinting. In: ICLR 2022 Conference Paper. https://openreview.net/forum?id=sOK-zS6WHB

Zhang AH (2024) The promise and perils of China's Regulation of Artificial Intelligence. University of Hong Kong. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4708676

Zuckerberg M (2021) A Blueprint for Content Governance and Enforcement. Facebook. https://www.facebook.com/notes/751449002072082/