UC BERKELEY

CENTER FOR LONG-TERM CYBERSECURITY

CLTC WHITE PAPER SERIES

# Survey of Search Engine Safeguards and their Applicability for AI

EVAN R. MURPHY, NADA MADKOUR, DEEPIKA RAMAN,
KRYSTAL JACKSON, JESSICA NEWMAN

CLTC WHITE PAPER SERIES

# Survey of Search Engine Safeguards and their Applicability for AI

EVAN R. MURPHY, NADA MADKOUR, DEEPIKA RAMAN,
KRYSTAL JACKSON, JESSICA NEWMAN

**AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley**

*Affiliations listed above are current, or were during the authors' main contributions to this work.*
*Views expressed here are those of the authors, and do not represent views of UC Berkeley or others.*

**May 2025**


## ABSTRACT

This paper reviews existing search engine safeguards and analyzes their potential relevance to AI systems such as conversational agents. We focus on safeguards which address six prominent risk categories that are present for both search engines and AI: misinformation and disinformation, national security risks, risks to the individual, adversarial manipulation, bias and discrimination, and intellectual property infringement. For eight time-tested safeguards, we explore both practical approaches and challenges that could arise in adapting each to an AI context. Among our main findings are that employing human raters at scale and integrated fact-checking are search engine safeguards that appear both highly relevant to and underutilized in today's AI systems. Two other safeguards — removing harmful content and malvertising mitigations — also appear promising, but seem more relevant or ripe for application to future AI systems. While our cross-domain analysis has limitations, our findings suggest that valuable lessons from search engine safeguards could help inform the development of safer and more secure AI systems.

# Contents

# 1. Introduction

Over the past three decades, search engines have encountered a wide range of potential harms and misuse risks, including the spread of misinformation, dissemination of extremist content, privacy violations, and threats to national security. In response, a variety of safeguards to mitigate these risks have been developed for search engines and tested over time.

More recently, artificial intelligence (AI) models have rapidly advanced and are increasingly deployed as services, notably in the form of conversational agents such as ChatGPT, which is among the 10 most-visited websites globally as of April 2025 (Similarweb n.d.). Due to the rapid adoption of this emerging technology, AI developers, governments, academia, and civil society have been racing to devise methods to address the risks posed by AI.

There are notable parallels between current AI and search engine technologies. The leaders in both provide centralized web services with massive user bases, generating content based on user inputs. Consequently, they face analogous misuse risks, including the potential to produce or amplify harmful content, propagate misinformation, infringe on user privacy, and be exploited for malicious purposes. There is also increasing convergence and interplay between search and AI technologies, with Google Search now including AI summaries in its search results page (Reid 2024) and OpenAI developing new AI-based search tools (OpenAI 2024). Hence, as AI developers seek effective risk mitigations, it begs the question: How did search technologies develop successful safeguards, and how might they be adapted for the AI context?

This paper explores the risk-reduction techniques employed by search engines and discusses their potential applicability to improving generative AI[1] safety and reliability. Specifically, we focus on safeguards such as the delisting or purging of dangerous content, the employment of large numbers of human raters, the integration of fact-checking services, protections against malvertising (i.e., advertisements used for malicious purposes), and collaborations with national security agencies, among others.

The value of this cross-domain research lies in addressing the siloed nature of information and expertise across different technological fields. AI researchers often immerse themselves deeply in AI literature and may not fully engage with the development and experiences of other domains such as search engine research. Insights and lessons learned by search engine compa-

1 We use the term "generative AI" to include general-purpose AI systems and large language models (LLMs).

nies in developing safeguards against misuse risks for search may even fail to reach AI developers within the same corporation, if they are located in different reaches of the company.

Table 1 below summarizes the main findings of our research. The following sections discuss the risks we considered and go into more detail on each search engine safeguard. By reviewing the risk mitigation methods that search engines have employed to address various risks, we hope to uncover proven safeguards that may be useful to AI developers and researchers.

### Table 1: Summary of Search Engine Safeguards and their Applicability for AI

| # | Safeguard Name | Safeguard Description | Risks Addressed | Applicability for AI |
|---|---|---|---|---|
| 1 | Removing Harmful Content | Eliminating undesirable content from search indexes or model training data, or otherwise ensuring that harmful content cannot appear in the tool's user-facing output. | - Misinformation and disinformation<br>- National security risks<br>- Safety of the individual<br>- IP infringement<br>- Bias and discrimination | Highly relevant, but current methods for implementation in AI may be very expensive or unreliable. |
| 2 | Content Filters | Filtering out dangerous or unwanted content using keywords, contextual analysis, and media analysis. | - Safety of the individual<br>- Misinformation and disinformation | Highly relevant, though some forms are already used extensively in AI services. |
| 3 | Behavioral Monitoring | Analyzing user interactions and behavior to help identify misuse such as jailbreak attempts and automated abuse. | - Adversarial manipulation<br>- National security risks<br>- Safety of the individual | Highly relevant, but likely already used extensively in AI services. |
| 4 | National Security Reporting | Establishing clear communication channels and protocols between AI developers and national security agencies. | - National security risks | Highly relevant and feasible; there is evidence that this is already done to an extent. |
| 5 | Human Raters at Scale | Employing large numbers of human evaluators to rate content based on established guidelines and criteria. | - Misinformation and disinformation<br>- National security risks<br>- Safety of the individual<br>- Adversarial manipulation<br>- Bias and discrimination<br>- IP infringement | Highly relevant and likely neglected, though expensive. |

| | | | | |
|---|---|---|---|---|
| **6** | Integrated Fact-Checking | Incorporating real-time fact checking mechanisms such as fact checking APIs, knowledge graphs, and providing source citations, to enhance the accuracy and reliability of model output. | - Misinformation and disinformation | Highly relevant and neglected, somewhat feasible to implement though currently fact-checked claims make up a small subset of all claims. |
| **7** | Algorithmic Obfuscation | Withholding technical details, such as model weights, to help protect against adversarial attacks and other risks. | - Adversarial manipulation | Highly relevant and possible, already used extensively in some closed-weight model developers but less by open-weight model developers. |
| **8** | Malvertising Mitigations | Mitigating "malicious advertising" content by integrating security tools that analyze and detect malvertising content, enforcing strict content policies, and developing mechanisms for user warning systems. | - Safety of the individual<br>- Adversarial manipulation<br>- Misinformation and disinformation<br>- National security risks | Low relevance for today's AI systems, could become important if future AI services rely more on advertising. |

# 2. Risks posed by Search Engines and AI[2]

## 2.1 FOCUS RISK CATEGORIES FROM BOTH SEARCH ENGINES AND AI

There is a substantial overlap between the risks posed by search engines and those posed by AI systems. Following is a (non-exhaustive) list of risk categories shared by search engines and AI. These risks are the ones we primarily consider in this paper as we analyze various safeguards:

- **Misinformation and disinformation:** Both search engines and AI systems can facilitate the large-scale unintentional spread of false or misleading information. They can also be used to amplify targeted attempts to deliberately spread deceptive or manipulative information, or to influence public perception, leading to an erosion of trust in information ecosystems and media institutions (Metaxa and Echeverry 2017, Slattery et al. 2024, Solaiman 2024).
- **National security risks:** Search engines and AI tools may enable access to dangerous or sensitive information that can be used to develop new or enhance existing weapons (e.g., Lethal Autonomous Weapons or CBRNE). They can also enable adversarial success in high-value cyber attacks that disrupt critical infrastructure, causing harm to public safety, and to disseminate extremist content (Schmid 2021, Chapter 19).
- **Risks to the individual:** Users of search engines and AI tools may face a variety of risks. These risks include the potential infringement of privacy and lack of consent in data handling (Lukas et al. 2023), increased surveillance, leakage and unauthorized disclosure of sensitive and personally identifiable information (PII), child sexual abuse material (CSAM), phishing (Chiew et al. 2018, p. 3), and overreliance on the system in a way that could compromise autonomy or cause other self-harm (Arora et al. 2016).
- **Adversarial manipulation:** Adversarial actors can manipulate search engines and AI tools through techniques such as SEO spam for search (Spirin et al. 2012) and adversarial attacks for AI (Lin et al. 2021), which allow bad actors to exploit system weaknesses in order to propagate harmful information or invoke unsafe system behaviors.

2 Note that while the listed items in this section are probably best considered as "risk categories", we sometimes refer to them simply as "risks" and use the two terms fairly interchangeably throughout this paper.

- **Bias and discrimination:** Data discrimination in search engine algorithms has been shown to privilege some identities over others (Noble 2018). Similarly, AI system decisions can result in the loss of opportunity or benefits that disparately affect some sections of the population more than others (Ferrer 2021).
- **Intellectual property infringement:** Search engines and AI tools can facilitate creation of copyrighted, trademarked, or licensed content without proper authorization, leading to infringement; increased risk of trade secret disclosure; and unauthorised duplication or plagiarism, raising economic and ethical concerns (Autio et al. 2024, Tanwar et al. 2024).

## 2.2 ADDITIONAL RISK CATEGORIES

This section provides additional examples of risks posed by search engines and/or AI systems, though it is still not exhaustive.

### 2.2.1 Additional Risks from Both Search Engines and AI

Additional examples of risks from both search engines and AI are provided in this subsection. We do not analyze safeguards against these types of risks because we either do not believe meaningful safeguards yet exist to mitigate them for search engines that might be applicable for AI, or because we had to leave them for future research.

- **Cybersecurity risks:** Malicious links, user profiling, and third-party tracking are common risks in using search engines that target individuals and businesses (Hu et al. 2007, Hannak et al. 2013). Similarly, current AI models have already demonstrated the capabilities to assist in cyber attacks (Lin et al. 2024, CPR 2023).
- **Concentration of power risks:** The possession or control of highly successful AI or search-based technologies presents an increasing concentration of economic, military, and political power in the hands of a relatively small number of people or entities (Liu 2018). For example, consider the competition for technological dominance between the US and China (Yeo 2016), search market concentration (Pasquale 2015), and the digital divide (Van Dijk 2006).
- **Filter bubbles and echo chambers:** Personalization of search and AI assistants can create information environments where users primarily encounter content that aligns with their existing beliefs, potentially reinforcing biases and limiting exposure to diverse perspectives (Pariser 2011).

### 2.2.2 Risks Primarily from Search Engines

Examples of risks posed primarily by search engines (and not AI systems) are provided in this subsection. We do not analyze safeguards against these types of risks because we are researching safeguards with applicability to AI, so we only consider safeguards that can help mitigate risks to both search engines and AI.

- **Temporal bias:** Search results often prioritize recent information, potentially undermining historical understanding and creating recency biases in knowledge acquisition (Wouters et al. 2004).
- **Link economy power imbalances:** The ability to rank prominently creates a "PageRank economy," where established websites gain increasing advantage through accumulated linking patterns.[3]
- **Index bias and coverage gaps:** Search engines only index a fraction of the internet, creating "invisible webs" of content that effectively do not exist for most users, regardless of relevance or quality.

### 2.2.3 Risks Primarily from AI

Examples of risks posed primarily by AI systems (and not search engines) are provided in this subsection. We do not analyze safeguards against these types of risks because we are researching search engine safeguards with applicability to AI. Given that these are primarily AI risks, we do not expect to find safeguards for search engines that could help mitigate them. For further discussion on AI risks, see Barrett et al. (2024), Barrett et al. (2025), and Raman et al. (2025).

- **Labor market disruption:** Through the combination of workplace automation and intellectual property theft for model training, workers across a range of industries might experience job displacement and financial losses that could exacerbate inequality (DSIT 2023, Zwetsloot and Dafoe 2019).
- **Persuasion at superhuman levels:** AI models have been effective at producing "personalized persuasion" at scale, which could lead to mass-scale opinion manipulation at a fraction of the cost compared to traditional methods (Chen and Shu 2023, Matz et al. 2024, Wiliams et al. 2024).

3 Link economy power imbalances relate to the natural accumulation of power by successful websites. In contrast to the "adversarial manipulation" risk category listed in section 2.1, sites in this case may be perfectly legitimate and are not necessarily manipulating search engine rankings in any way.

- **Model autonomy:** Without adequate human oversight, AI systems could take actions that do not align with societal values, or pursue critical ML research that could outpace our ability to enact sufficient safeguards (Karnofsky 2024, OpenAI 2023b, Anthropic 2024).
- **Deception:** Models have demonstrated the ability to deceive humans, both during model training (e.g., strategically underperforming during evaluations) and when deployed for human interaction (van der Weij et al. 2024). For instance, GPT-4 convinced a human worker to solve a CAPTCHA on its behalf (Open AI 2023a, pp. 15–16).

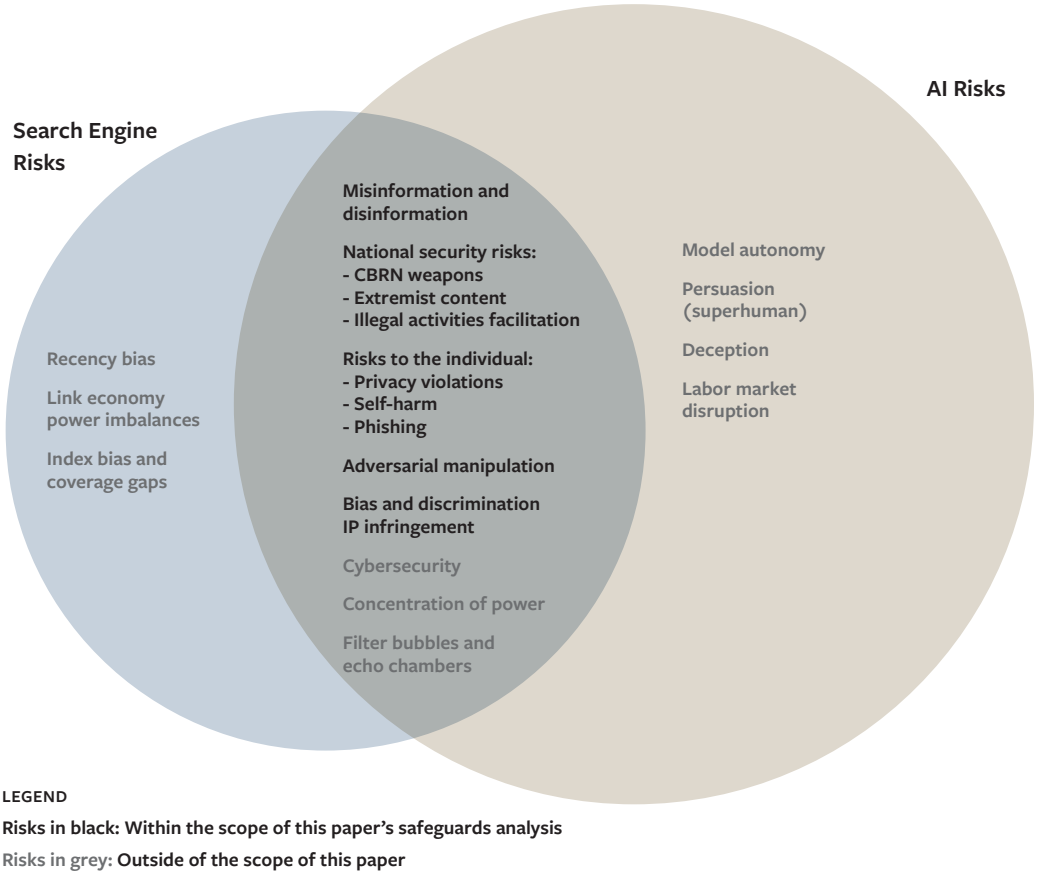


Figure 1: Search engine risks, AI risks, and their relationships (non-exhaustive)

This figure illustrates the substantial overlap between major risks posed by the two technologies, as well as the relatively larger set of risks posed by AI compared to search engines. The risks in black text are within the scope of this paper, and the risks in grey text are outside of this paper's scope for various reasons, though they could still be important. This is not an exhaustive presentation of risks in any of the three categories.

# 3. Analysis of Safeguards

This section expands on the selected subset of search engine safeguards introduced in Table 1 and further examines their relevance/applicability to AI systems. By identifying shared risks between search engines and AI systems, and evaluating the applicability of existing search engine safeguards to AI systems, this analysis aims to contribute to the broader discourse on AI safety. Given the limited historical research on search engine risks and risk mitigations, the selected safeguards are a result of synthesized insights from product reports, industry practices, and the authors' experience with both search engines and AI systems. While the list of selected safeguards is not exhaustive, those included here were chosen based on their apparent success as search engine safeguards and potential applicability to AI systems.

## 3.1 SAFEGUARD #1: REMOVING HARMFUL CONTENT

### 3.1.1 Overview

Search engines may completely delist or remove some content from their indexes, preventing such content from appearing in search results. This approach guarantees that certain sources of content can never be surfaced directly by search engines. For example, individuals in the EU and other European countries with similar data privacy laws may make requests to Google for content about them to be delisted; Google reviews the request and makes a decision on whether or not to delist the requested pages. As of Nov. 30, 2024, 2,934,275 URLs had been delisted (Google Transparency Report 2024). Google also frequently delists content if it is determined to violate copyright laws, if a government requests its removal, or for other reasons (Google Transparency Report 2024a).

**Risks Addressed**

- Misinformation and disinformation
- National security risks
- Safety of the individual
- Intellectual property infringement
- Bias and discrimination

### 3.1.2 Applicability for AI

Removing harmful content from search indexes or model training data (or otherwise ensuring that it cannot appear in the tool's user-facing output) is highly relevant for AI. LLMs without fine-tuning are capable of generating many types of harmful content. Frontier model developers in recent years have typically used techniques like Reinforcement Learning from Human Feedback (RLHF) to make generation of harmful responses less likely. However, such techniques are susceptible to jailbreaks and not very reliable compared to the guarantees provided by delisting webpages from a search engine.

Perhaps the more straightforward analogue to delisting webpages for AI would be purging content from pre-training data[4] and retraining the model from scratch. However, given the cost of training runs for frontier models in the early 2020s, this technique could be extremely expensive. Removing content from pre-training data is also not likely to be as reliable as delisting webpages is for search engines, due to fundamental differences between search and AI. Since a search engine is typically linking users to webpages in a straightforward way, delisting all webpages with certain types of content virtually guarantees that such content will not be presented to users. However, AI models are not simply outputting items verbatim that they have encountered in their pre-training data, and may be capable of inferring the existence of items that they have not specifically encountered before during training.[5]

Machine unlearning (MU) is an alternative approach that could be both practical and effective, allowing for more thorough removal of content from AI systems without the expense of retraining a model from scratch. However, MU is still an emerging field of research and may not be reliable or ready for use by production AI models yet (Ginart et al. 2019; Bourtoule et al. 2021; Liu et al. 2024). Hence, content removal could be a powerful type of safeguard in AI systems and appears underutilized, but it may not be possible yet to perform cheap and effective content removal on AI models.

**Approaches**

- **Reinforcement Learning from Human Feedback (RLHF):** Using human quality review of model outputs to optimize AI performance throughout reinforcement learning.

4 Content may be removed based on the results of data audits and other filtering techniques (Birhane et al. 2021, Dodge et al. 2021).

5 For a contrived example, if a developer decided that the number 13 was harmful because it is too unlucky, and they removed all traces of 13 from a model's pretraining data before training, the model may still be able to infer the existence of 13 from the principles of mathematics it has learned through other parts of the pretraining corpus.

- **Purging pre-training data content:** Removing or deleting data used to train the model prior to model training.
- **Machine unlearning (MU):** Targeted removal of the influence of specific training data points on an already-trained model.

**Challenges**

- **High cost of frontier model training runs:** Frontier models typically require extensive training periods and considerable resources (e.g., computational power, large datasets, energy consumption, and research and development) (Cottier et al. 2024).
- **Lack of robustness:** Models can be exploited by the use of adversarial prompts or other means to circumvent model safeguards and invoke misaligned behavior (Wei et al. 2023). Machine unlearning techniques may also not be sufficiently reliable or appropriate for widespread use.

## 3.2 SAFEGUARD #2: CONTENT FILTERS

### 3.2.1 Overview

Content filtering is a fundamental safeguard used by search engines to prevent the dissemination of inappropriate or harmful material. Features like Google's SafeSearch allow users to filter out explicit content, helping to create a safer browsing experience, especially for children and sensitive audiences (Google n.d.). Content filters utilize keyword detection, image recognition, and other techniques to identify and block content that falls into categories such as pornography, violence, hate speech, or illegal activities.

**Risks Addressed**

- Safety of the individual
- Misinformation and disinformation

### 3.2.2 Applicability for AI

AI models, particularly those that generate text or images, can inadvertently produce content that is inappropriate, offensive, or harmful. Implementing robust content filtering mechanisms is essential to prevent the generation and dissemination of such material.

**Approaches**

- **Keyword and phrase filtering:** Blocking responses that contain disallowed terms or phrases.
- **Contextual analysis:** Using natural language processing to understand the context and ensure that content is appropriate, even if specific keywords are not present.
- **Image and media analysis:** For AI models that generate images, employing computer vision techniques to detect and filter out inappropriate visual content.
- **Dynamic filtering:** Continuously updating filter lists and models to adapt to new slang, code words, or evasion techniques used to bypass filters.

**Challenges**

- **Over-blocking:** Risk of filtering out acceptable content, potentially impacting the usability of the AI service.
- **Under-blocking:** Failure to catch all inappropriate content, allowing harmful material to slip through.
- **Bias and fairness:** Ensuring that filters do not disproportionately affect certain groups or introduce unintended biases.

## 3.3 SAFEGUARD #3: BEHAVIORAL MONITORING

### 3.3.1 Overview

Behavioral monitoring involves analyzing user interactions and patterns to detect malicious activities, such as bot traffic, coordinated misinformation campaigns, or attempts to manipulate search algorithms. Search engines utilize behavioral analytics to:

- **Identify unusual activity:** Detect spikes in traffic or search queries that may indicate the use of automated bots or coordinated efforts to influence search results.
- **Prevent abuse:** Implement rate limiting, CAPTCHA challenges, and other measures to prevent abuse of services.
- **Enhance security:** Monitor for signs of account compromise or fraudulent activities.

**Risks Addressed**

- Adversarial manipulation
- National security risks
- Safety of the individual

### 3.3.2 Applicability for AI

AI models can be subject to misuse through adversarial inputs designed to elicit disallowed content or compromise the system. Behavioral monitoring can help identify:

- **Jailbreak attempts:** Users trying to trick the AI into generating prohibited content.
- **Automated abuse:** Bots or scripts making high volumes of requests to overwhelm the system or extract data.
- **Patterns of misuse:** Repeated attempts to exploit known vulnerabilities.

**Approaches**

- **Logging and analysis:** Collecting data on user interactions to identify patterns indicative of misuse.
- **Anomaly detection:** Using machine learning models to detect deviations from normal usage patterns.
- **Adaptive response mechanisms:** Automatically adjusting system responses or implementing protective measures when misuse is detected.

**Challenges**

- **Privacy concerns:** Ensuring that user data is handled in compliance with privacy laws and ethical standards.
- **False positives:** Avoiding incorrect identification of legitimate user behavior as malicious.
- **Resource allocation:** Monitoring can require significant computational resources, especially at scale.

## 3.4 SAFEGUARD #4: NATIONAL SECURITY REPORTING

### 3.4.1 Overview

Search engine companies collaborate closely with government agencies to address national security concerns, such as the potential dissemination of terrorist content, cyber threats, and other activities that could compromise public safety. For instance, Google complies with legal requests from government entities to remove content that violates laws or poses national security risks. According to the Google Transparency Report (2024b), Google received numerous requests from governments worldwide to remove content or provide user data for law

enforcement or national security purposes. These collaborations involve proactive content monitoring and mechanisms to respond rapidly to emerging threats.

In addition to legal compliance, search engines have developed internal policies and technologies to detect and remove content related to terrorism, violent extremism, and other national security threats. Automated systems, such as machine learning algorithms, are used to identify and flag such content, complemented by human reviewers who assess and take appropriate action.

**Risks Addressed**

- National security risks

### 3.4.2 Applicability for AI

AI models, particularly large language models deployed as conversational agents, may inadvertently generate content that poses national security risks. For example, an AI model might provide detailed instructions for illicit activities or generate propaganda that promotes extremist ideologies. There may also be cases where users are engaged in activity on an AI platform that could aid in a national security investigation.

Enhancing collaboration between AI developers and national security agencies is crucial for identifying and mitigating potential risks. Establishing clear communication channels and protocols can help AI developers receive guidance on content that should be restricted and enable them to report concerning patterns of user interactions.

AI models can integrate monitoring systems similar to those used by search engines to detect and prevent the generation of content that could compromise national security. This may involve implementing advanced filtering mechanisms, leveraging threat intelligence feeds, and employing human reviewers with appropriate security clearances to oversee sensitive areas. Challenges include balancing national security considerations with user privacy and free speech rights, as well as the technical difficulty of controlling AI-generated content compared to filtering existing web content.

Overall, national security reporting and monitoring are important for AI and should be achievable for AI models deployed as a service (less so for open-weight models). OpenAI publishes reports on government requests for user data, including anonymized figures around how many such requests they have received and fulfilled in a given time period (OpenAI n.d.). So while the

full extent of collaborations between AI developers and national security authorities may not be publicly known, there is evidence that national security reporting mechanisms have already been established to some extent at frontier AI companies.

**Approaches**

- **Establishing clear communication channels:** Establishing communication protocols with authorities can help AI developers receive guidance and report concerns.
- **Integration of monitoring systems and filtering mechanisms:** Automated monitoring of generated content can help identify and report or block the generation of dangerous content.
- **Leveraging threat intelligence feeds:** Utilizing existing threat intelligence feeds can help increase awareness and improve alignment with the current threat landscape.

**Challenges**

- **Privacy and surveillance concerns:** Collaboration with national security agencies raises legitimate concerns about surveillance overreach and potential infringement on civil liberties. AI systems collecting user data and communications for security purposes should balance legitimate security needs against privacy rights. These systems may disproportionately impact marginalized communities or political dissidents if implemented without robust oversight mechanisms.
- **Compliance:** Collaborative efforts between AI developers and national security agents are contingent upon the voluntary compliance of both parties, which is not guaranteed.

## 3.5 SAFEGUARD #5: HUMAN RATERS AT SCALE

### 3.5.1 Overview

Search engines like Google employ extensive networks of human evaluators, known as Search Quality Raters, to assess the quality, relevance, and safety of search results. These raters use detailed guidelines — the Search Quality Evaluator Guidelines (Google 2020) — to rate webpages based on criteria such as experience, expertise, authoritativeness, and trustworthiness (E-E-A-T), as well as to identify harmful or inappropriate content. As of 2020, Google reportedly had over 16,000 such raters worldwide (Google 2020).

Human raters provide critical insights that automated systems may miss, including nuanced judgments about content quality, cultural sensitivities, and emerging types of harmful content.

Their evaluations help improve search algorithms and ensure that search results meet users' needs while adhering to safety standards.

**Risks Addressed**

- Misinformation and disinformation
- National security risks
- Safety of the individual
- Adversarial manipulation
- Bias and discrimination
- Infringement of intellectual property

### 3.5.2 Applicability for AI

AI models, especially those deployed at scale, interact with millions of users and generate vast amounts of content. Relying solely on small teams for training and moderation may not be sufficient to address the diversity and complexity of potential issues. Incorporating large numbers of human evaluators can enhance the ability of AI developers to detect and mitigate harmful outputs.

AI model developers are already relying on some human raters, but compared to the 16,000 raters that Google claimed to employ for search, the scale seems to be much smaller. OpenAI credited 40 human labelers[6] in its paper introducing InstructGPT (Ouyang et al. 2022). TIME reported on documents detailing OpenAI's contracts with outsourcing firm Sama in late 2021 to hire labelers to reduce toxicity in ChatGPT, estimating that "[a]round three dozen workers" were hired (TIME 2023). If those figures are at all representative, then leading AI model developers are using between two and three orders of magnitude fewer human raters than their counterparts in search.

**Approaches**

By employing a diverse pool of human reviewers, AI developers can gain insights into cultural nuances, ethical considerations, and subtle forms of misuse that automated systems might overlook. These human evaluators can participate in various stages of AI development, including:

- **Data curation:** Reviewing and annotating training data to ensure quality and appropriateness.

6 There are 40 names in the list, but one name appears twice, so there are only 39 unique names and there may only have been that many hired data labelers for InstructGPT.

- **Model evaluation:** Assessing AI outputs for compliance with safety guidelines and identifying areas for improvement.
- **Policy development:** Providing feedback to inform content moderation policies and ethical guidelines.

**Challenges**

Implementing this safeguard poses challenges, such as:

- **Scaling the workforce:** Employing human raters would require a substantial number of workers, resulting in increased need for management, additional costs, and other resources.
- **Ensuring the well-being of human evaluators:** Evaluators may be exposed to disturbing content, which can have undesirable effects on the evaluators' well-being.
- **Maintaining consistency in evaluations:** Inconsistent evaluation methods by human evaluators can lead to unreliable results, particularly if part of the evaluation process includes comparison to previous evaluation results.
- **Confidentiality and intellectual property (IP):** Considerations related to confidentiality and intellectual property can arise, particularly when involving external contractors.

Despite these challenges, the long-term reliance of search engines on large numbers of human evaluators underscores the value of human judgment in content moderation and guidance of information systems. AI developers could adopt similar practices to enhance the safety and reliability of AI services. While this technique is expensive, it could still be a worthwhile investment.

## 3.6 SAFEGUARD #6: INTEGRATED FACT-CHECKING

### 3.6.1 Overview

Search engines have implemented integrated fact-checking mechanisms to combat the spread of misinformation and enhance the credibility of information presented to users. Google, for instance, introduced the "Fact Check" label in search results and Google News, highlighting articles that include fact-checked claims (Conger 2017). This feature relies on partnerships with reputable fact-checking organizations that use the ClaimReview markup to indicate fact-checked content (Google Developers 2021). By December 2019, fact checks were displayed over 11 million times daily across search results globally, amounting to approximately four billion impressions annually (Google Blog 2019).

These efforts aim to provide users with accurate information and help them make informed decisions. Fact-checking labels and summaries offer context and verification, reducing the spread of false or misleading content.

**Risks Addressed**

- Misinformation and disinformation

### 3.6.2 Applicability for AI

AI models, particularly large language models used in conversational agents, may generate incorrect or misleading information due to limitations in their training data or understanding of context. Without integrated fact-checking, AI-generated content could inadvertently contribute to the spread of misinformation.

**Approaches**

Incorporating real-time fact-checking mechanisms into AI models can enhance the accuracy and reliability of their outputs. This could involve:

- **Integrating fact-checking APIs:** Connecting AI models with external fact-checking services to verify information before presenting it to users.
- **Implementing knowledge graphs:** Using structured databases of verified information to cross-reference and validate generated content.
- **Providing high-quality source citations:** Encouraging AI models to include references or links to reputable sources, allowing users to verify the information independently. A recent experiment found that many AI chatbots are providing incorrect citations (Jaźwińska and Chandrasekar 2025).

Some AI services have begun to address this issue. For example, OpenAI's ChatGPT has experimented with providing source links in its responses to improve transparency and allow users to verify information.

**Challenges**

- **Computational overhead:** Real-time fact-checking can increase response times and require additional computational resources.
- **Data availability:** Accessing up-to-date and comprehensive verified data can be challenging.
- **Handling discrepancies:** Dealing with conflicting information from different reputable sources requires sophisticated reasoning capabilities.

## 3.7 SAFEGUARD #7: ALGORITHMIC OBFUSCATION

### 3.7.1 Overview

Algorithmic obfuscation refers to the practice of keeping the specifics of algorithms, such as search ranking or recommendation systems, confidential to prevent manipulation by bad actors. Search engines like Google regularly update and adjust their algorithms and are deliberately vague about the details to:

- **Prevent gaming:** Obfuscating algorithms can help avoid situations where content creators manipulate content to unfairly improve their rankings (e.g., black-hat SEO techniques).
- **Enhance security:** Hiding the specifics of algorithms can help reduce the risk that adversaries will exploit known algorithmic weaknesses.
- **Maintain integrity:** Algorithmic obfuscation can help ensure that search results remain relevant and trustworthy.

**Risks Addressed**

- Adversarial manipulation

### 3.7.2 Applicability for AI

**Approaches**

AI models are susceptible to exploitation if adversaries understand their inner workings in detail. By withholding certain technical details, AI developers can:

- **Reduce vulnerability to adversarial examples:** Withholding details can limit the ability of attackers to craft inputs that fool the AI into producing undesired outputs.
- **Prevent data poisoning:** Obfuscation can make it harder for bad actors to introduce malicious data into training sets.

**Challenges**

- **Balancing transparency and security:** Developers need to provide sufficient information to build user trust without enabling exploitation of the services by malicious users.
- **Regulatory compliance:** Makers of AI models must adhere to laws and guidelines that may require disclosure of certain algorithmic processes.
- **Community trust:** It is important to manage perceptions, as excessive secrecy may lead to skepticism or mistrust among users and stakeholders (Pasquale 2015).

Algorithmic obfuscation is highly relevant to AI and is already being employed by closed-weight AI developers. Open-weight model developers, however, provide much more transparency into their algorithms, which can provide many benefits but also is subject to risks such as adversarial manipulation.

## 3.8 SAFEGUARD #8: MALVERTISING MITIGATIONS

### 3.8.1 Overview

Malvertising, or malicious advertising, refers to the use of online advertising to distribute malware, phishing scams, or other harmful content. Search engines and advertising platforms have developed sophisticated systems to detect and prevent malvertising, protecting users from these threats. These safeguards involve:

- **Real-time scanning:** Automated systems analyze ads for malicious code or suspicious behavior before they are displayed.
- **Strict advertising policies:** Platforms enforce policies that prohibit deceptive or harmful content, with penalties for violations.
- **Publisher verification:** Publishers vet advertisers and publishers to ensure they are legitimate and trustworthy.

For example, Google reported that in 2021, it blocked or removed over 3.4 billion ads for violating policies and suspended over 5.6 million advertiser accounts (Spencer 2022).

**Risks Addressed**

- Safety of the individual
- Adversarial manipulation
- Misinformation and disinformation
- National security risks

### 3.8.2 Applicability for AI

Some AI services already generate content that includes external links or references, and as AI services evolve, they may begin to include more advertising. Without appropriate safeguards, AI models could be manipulated to generate content that directs users to malicious websites or promotes fraudulent activities.

**Approaches**

To mitigate these risks, AI developers can:

- **Implement malicious content detection:** Developers can integrate security tools that analyze AI outputs for potential malvertising or harmful content.
- **Enforce content policies:** AI companies can establish and enforce strict policies that prohibit the generation of content that includes malicious links or promotes illegal activities.
- **User warning systems:** AI tools can include mechanisms to alert users when content may be suspicious or harmful.

**Challenges**

- **Emerging threats:** Malicious actors continually develop new tactics to evade detection, requiring ongoing updates to security measures.
- **False positives/negatives:** Developers must balance the detection accuracy to minimize both false alarms and missed threats.
- **User experience:** It can be a challenge to ensure that security measures do not unduly hinder the user experience.

# 4. Discussion

This survey of search engine safeguards reveals some potential opportunities for AI developers. Notably, there are several important risk categories where search engines and AI overlap. We can also see that while some protections widely used by search engines have already found their way into AI services, others appear to be underutilized.

## 4.1 STRUCTURAL DIFFERENCES BETWEEN SEARCH ENGINES AND AI

One key observation is that the fundamental differences between search engines and AI models shape the effectiveness and ease of implementation of certain safeguards. Search engines primarily index and retrieve existing content from the web, making it relatively straightforward to block or remove problematic material at the source.

In contrast, AI models do not retrieve content, but generate novel content based on statistical patterns learned in training. This generative nature of AI complicates the removal of harmful information. For example, removing harmful training data after deployment may not fully eliminate its influence, as models may be able to infer or recreate the learned harmful content. Mitigation techniques such as RLHF attempt to steer models toward safe behavior, but are not robust and are often vulnerable to jailbreaks. This highlights a core challenge: while *removing harmful content* (Safeguard #1) is straightforward for search engines, doing so for AI systems remains far more complex. Emerging techniques like machine unlearning could offer potential for data removal, but such methods remain underdeveloped and computationally expensive, and are not yet reliable enough for widespread deployment.

Additionally, certain safeguards would be more challenging or impossible to implement in open-source AI contexts, where model weights, or potentially code or other details are public. For instance, algorithmic obfuscation can be a useful strategy in closed-weight environments, but it is scarcely applicable when model weights are freely available. Similarly, behavioral monitoring would likely be very challenging or infeasible to implement in an open-source context where there is no centralized host.

## 4.2 PROMISING SAFEGUARDS FOR AI

At the same time, some safeguards have stood out as potentially well suited for direct application to AI. These measures, refined over years of implementation in the context of search engines, have been used to help mitigate many risks for search engines that overlap with risks for AI (e.g., the spread of misinformation or other harmful outputs).

For instance, ***large-scale use of human raters*** **(Safeguard #5)** could significantly improve the nuanced judgment required to identify subtle forms of harmful or misleading content, compared to the relatively small-scale use of human raters typically employed by AI developers today. Although this approach entails substantial cost and logistical effort, it has proven instrumental for search engines over decades and could similarly help AI developers ensure their models better align with human values and expectations. Given the tremendous scale and impact of AI services, a substantial investment in human oversight may be justified.

Another safeguard which appears promising is ***fact-checking partnerships and data verification tools*** **(Safeguard #6).** This measure directly targets misinformation and disinformation, which are among the most pervasive issues facing both search engines and AI. Incorporating real-time verification mechanisms, utilizing knowledge graphs, or sourcing reputable fact-checking APIs could substantially improve the quality and reliability of AI outputs, and help to mitigate the pervasive problem of hallucinations in many current AI models. While the complexity of implementing these systems and the limited coverage of fact-checked claims pose constraints, the potential benefits for reducing the generation and spread of misinformation and disinformation could be great.

Some safeguards that appear underused in AI could become more pressing as AI models evolve and their commercial models change. ***Malvertising mitigations*** **(Safeguard #8)** are currently less relevant for AI systems that do not rely heavily on advertising. However, as generative AI increasingly begins to deliver sponsored content, malvertising safeguards will become more important. Proactive consideration of such techniques — ranging from real-time scanning to robust advertiser vetting — can prevent bad actors from exploiting the AI ecosystem in ways analogous to how they have targeted search engines.

With ***national security reporting*** **(Safeguard #4),** search engines have set precedents for working with authorities to identify, track, and remove content that poses threats to public safety or national interests, as well as supporting data requests from authorities on user accounts that are suspected of illegal activities. There is some evidence that leading AI devel-

opers are already engaged in similar national security collaborations to a degree, though the full extent is not publicly known. While balancing national security considerations with civil liberties — such as privacy and free speech — remains a challenge, mechanisms to identify and report dangerous content to the appropriate authorities could enhance national security risks, including the spread of extremist content or sensitive information that could be used to build dangerous weapons.

# 5. Limitations and Further Research

Some search engine safeguards may involve trade secrets and non-public information, limiting our ability to analyze them fully. Additionally, the authors of this paper are primarily AI researchers rather than search engine experts, hence our understanding may not capture all nuances of search engine technologies. When studying specific search technology, we primarily considered Google, and we may have overlooked some safeguards that are only provided by other search engines. The landscape of risks for both search engines and AI is complicated. As discussed in Section 2, we were only able to investigate safeguards for a subset of the risks that are present for both search engines and AI systems (see section 5.4 on additional safeguards for further research).

The limitations emerging from our choice to narrow our focus to Google Search are most apparent when considering some of the risks discussed in Section 2.2.1. Google dominates the search engine market, holding an overwhelming ~90% of worldwide market share, which grants it a near-monopoly.[7] This concentration of power raises concerns about how information is curated and presented, particularly in relation to echo chambers, filter bubbles, and potential biases in search results.

Google's vast index database of trillions of web pages, universality of resources, personalized results, and integration into the suite of Google services make it the premier (and often default) choice for most users of the internet. However, this precision and personalization are often at the expense of harvesting user data across Google products, sometimes surreptitiously,[8] to enable targeted advertising. This tradeoff between privacy and efficiency may be better handled by privacy-centric services like DuckDuckGo. We did not extensively explore alternatives such as DuckDuckGo, Bing, or other search engines, which may employ different indexing strategies, ranking algorithms, or design choices that might address these risks differently. Expanding this investigation to other search engines could offer valuable insights for developing more robust and transparent safeguards. Just as crucial would be the analysis of user awareness and behaviors to illuminate the accompanying efforts required in managing these risks effectively.

7 Source: https://gs.statcounter.com/search-engine-market-share. These market shares are often as high as in geographic regions like India and Africa.

8 https://time.com/6962521/google-incognito-lawsuit-data-settlement/

There are also many important novel risks posed by AI systems that are not relevant for search engines, and hence are not likely to find any parallel solutions to draw from in search technology.

## 5.1 AI SUMMARIES IN SEARCH

We have not conducted an in-depth investigation into the integration of AI into search engines or the unique ways this might exacerbate some of the risks illustrated in this paper. Google Search has increasingly promoted AI summaries as a feature in their search results. The opaque determination about which queries warrant an AI overview, the absence of straightforward options to disable this feature, and the accompanying concerns, such as bias amplification, adversarial manipulation, and possible misinformation propagation at scale, result in a complex interaction of search and AI risks. These risks may require specialized safeguards that are distinct from those traditionally used in search, including enhanced adversarial robustness, transparency in ranking algorithms, and mitigations against AI-generated hallucinations.

## 5.2 OPEN-SOURCE AI

The dynamics between hosted AI services and open-source AI models differ, affecting how safeguards can be implemented. There are also different considerations regarding free speech and censorship concerns for search engines versus AI models, which may impact the applicability of certain safeguards.

## 5.3 EVALUATION OF SAFEGUARDS

While this paper surveys existing search engine safeguards and explores their applicability to AI risk mitigation, it does not typically provide robust evidence of each safeguard's effectiveness. As elaborated in Section 3, search engines have developed various mechanisms to address risks that seem successful, but detailed assessment of data related to their performance in risk mitigation is not readily available. In the absence of this information, we reached out to external reviewers, including individuals currently and formerly employed at a major search engine company, to evaluate existing safeguards, but we received limited input on the matter.

Therefore, we employed the following heuristic in assessing these safeguards: if a safeguard has been maintained for several years, it likely demonstrates some level of effectiveness, as a for-profit company would not allocate resources to an ineffective measure. However, this assumption is not foolproof: some safeguards may be retained for reasons beyond efficacy, such as public relations benefits or alignment with market expectations of perceived responsible AI practices.

Future work should include systematic evaluation of these safeguards through empirical studies to measure their impact on user behavior, content quality, and risk reduction. For instance, fact-checking integrations in search results may reduce the spread of misinformation and disinformation, but their effectiveness in altering user trust or engagement with false information requires deeper analysis. Empirical evaluations of safeguards will not only illustrate their efficacy in mitigating search-related risks, but can also be an important element of assessing their potential adaptability to AI systems. For example, a safeguard that performs well in structured search environments may not necessarily function as effectively in generative AI systems, where outputs are dynamically generated rather than retrieved from indexed content.

## 5.4 ADDITIONAL SAFEGUARDS FOR FURTHER RESEARCH

Further research is needed on additional search safeguards we did not analyze which could be applicable to AI. These likely include privacy and data protection safeguards, misinformation and disinformation controls, adversarial manipulation protections, phased deployment, adversarial manipulation protections, bias and fairness mitigations, and news/information recency checks. Additional notes on some of these safeguards are provided below.

**Privacy and data protection safeguards**

In accordance with GDPR and CCPA, search engines respond to data subject access requests, which span information such as the purpose of the data, how it will be used, and where the data originated (Mantelero 2013). Google, for example, honors "right to be forgotten" removal requests for personal data and will de-index sensitive personal information (like medical or financial records) upon request. AI systems pose privacy risks: models can memorize and regurgitate personal data as well as be prompted to verbatim reproduce information from their training set, such as phone numbers, names (Nasr et al. 2023). Further research on how safeguards could be applied to AI could be valuable, however some have noted that imple-

menting these sorts of privacy and protections could be very challenging or impossible for AI (Gucluturk 2024).

**Misinformation and disinformation controls**

Search engines have implemented measures to counter false or misleading content. For example, Google prioritizes authoritative sources and down-ranks "borderline content" to curb misinformation and disinformation (Google n.d.). It also applies extra scrutiny to sensitive "Your Money or Your Life" (YMYL) queries (e.g., topics relating to health or finance), often avoiding direct answers or providing disclaimers for unverified info.

Using ranking algorithms to favor high-quality information is a key part of how search engines attempt to maintain information quality (Google 2019). It seems worthwhile to further investigate these practices and how parallel measures (e.g., more sophisticated eligibility criteria around which data can be included in a training corpus) could help address misinformation and disinformation risks in AI systems.

**Bias and fairness safeguards**

Search engines employ certain bias and fairness safeguards, such as filtering and adjusting auto-complete suggestions, to avoid reflecting harmful stereotypes or falsehoods. They also provide users the ability to report poor results, which search engines use to improve rankings (Google Search Central n.d.). Further research could explore how such measures might be applied in the context of AI systems as well.

# 6. Conclusion

This paper has explored whether the decades of experience search engines have had in safeguard development and risk mitigation might offer valuable lessons for the developers of AI foundation models. We cataloged eight different safeguards used by search engines, each addressing a subset of the six risk categories we focused on (section 2.1): misinformation, national security risks, risks to the individual, adversarial manipulation, bias and discrimination, and intellectual property infringement, all of which are shared by both search and AI systems.

We found that some of the safeguards we analyzed — such as *content filters* (Safeguard #2), *behavioral monitoring* (Safeguard #3), *national security reporting* (Safeguard #4), and *algorithmic obfuscation* (Safeguard #7) — appear highly relevant to AI. Although developers of closed-weight, hosted AI models often use these techniques already, opportunities may exist to refine and strengthen them further by applying approaches from their more mature counterparts in search.

However, other safeguards employed by search engines — such as the use of *human raters at scale* (Safeguard #5) and *integrated fact-checking* (Safeguard #6) — appear to remain underutilized in AI systems. Both are at least moderately feasible, if potentially costly, to implement. Together, they could help mitigate all six of the shared risks we have focused on.

Finally, two safeguards we studied — *removing harmful content* (Safeguard #1) and *malvertising mitigations* (Safeguard #8) — may hold greater promise for the future than for immediate adoption. Techniques like machine unlearning (MU) must mature further before removing harmful content from AI models can be done reliably and economically. Likewise, before adding advertising features to their AI products, developers would be wise to study the history of malvertising exploits affecting search engines, as well as the responses and mitigations that proved effective. Both of these safeguards together also could help address all six of the shared risks we have focused on once they are fully developed and integrated.

We hope our analysis raises awareness among AI developers, researchers, and other stakeholders about risk mitigation strategies they might not have otherwise considered. Since this paper only examined a subset of search engine safeguards and the overlapping risks between search and AI, further cross-domain research — whether between AI and search, or AI and other parallel domains — could uncover even more mitigation strategies. Such efforts can guide the development of AI systems to more closely align with societal values and prioritize user safety and security.

# Acknowledgments

We presented an earlier version of this research to Task Force 5.1 of the U.S. Artificial Intelligence Safety Institute Consortium (AISIC) in December 2024. Thanks to the other task force members, especially Danaé Metaxa, for their thoughtful discussions which both provided the initial inspiration for this research and helped to improve it. Thanks also to Fili Wiese and Tej Kalianda for their thoughtful comments on a draft of this paper. Special thanks to Ann Cleaveland both for her thoughtful comments and for providing a home and support for this work at the UC Berkeley Center for Long-Term Cybersecurity (CLTC). Thanks to Chuck Kapelke for his careful editing. Thanks to Nicole Hayward for her skilled design work. Thanks to Rachel Wesen for her support with the publication of this paper through CLTC, and to other CLTC staff who contributed. Finally, thanks to our families and pets for all of their patience and support as we toiled away on this research. Any errors are our own.

# References


Anthropic. (2024). *Responsible scaling policy*. https://assets.anthropic.com/m/24a47b00f10301cd/original/Anthropic-Responsible-Scaling-Policy-2024-10-15.pdf

Arora, V. S., Stuckler, D., & McKee, M. (2016). Tracking search engine queries for suicide in the United Kingdom, 2004–2013. *Public Health, 137*, 147–153. https://doi.org/10.1016/j.puhe.2015.10.015

Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., & Roberts, K. (2024). *Artificial intelligence risk management framework: Generative artificial intelligence profile* [NIST Trustworthy and Responsible AI]. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.600-1

Barrett, A. M., Jackson, K., Murphy, E. R., Madkour, N., & Newman, J. (2024). *Benchmark Early and Red Team Often: A Framework for Assessing and Managing Dual-Use Hazards of AI Foundation Models.* arXiv. https://arxiv.org/abs/2405.10986

Barrett, A. M., Newman, J., Nonnecke, B., Madkour, N., Hendrycks, D., Murphy, E. R., Jackson, K., & Raman, D. (2025). *AI risk-management standards profile for general-purpose AI (GPAI) and foundation models*. https://cltc.berkeley.edu/wp-content/uploads/2025/01/Berkeley-AI-Risk-Management-Standards-Profile-for-General-Purpose-AI-and-Foundation-Models-v1-1.pdf

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., & Papernot, N. (2021). Machine unlearning. *2021 IEEE Symposium on Security and Privacy (SP)*, 141–159.

Chen, C., & Shu, K. (2024). Can LLM-generated misinformation be detected? *arXiv*. https://arxiv.org/abs/2309.13788

Chiew, K. L., Sheng, K., & Tan, C. L. (2018). A survey of phishing attacks: Their types, vectors and technical approaches. *Expert Systems with Applications, 106*, 1–20. https://doi.org/10.1016/j.eswa.2018.03.050

Conger, K. (2017). *Expanding Fact Checking in Google Search and News*. Google Blog. https://blog.google/outreach-initiatives/google-news-initiative/expanding-fact-checking-google/

Cottier, B., Rahman, R., Fattorini, L., Maslej, N., Besiroglu, T., & Owen, D. (2024). The rising costs of training frontier AI models. *arXiv*. https://arxiv.org/abs/2405.21015

CPR. (2023). *OPWNAI: Cybercriminals starting to use ChatGPT*. Check Point Research. https://research.checkpoint.com/2023/opwnai-cybercriminals-starting-to-use-chatgpt/

DSIT. (2023). *Capabilities and risks from frontier AI*. UK Department for Science, Innovation & Technology. https://assets.publishing.service.gov.uk/media/65395abae6c968000daa9b25/frontier-ai-capabilities-risks-report.pdf

Ferrer, X., Van Nuenen, T., Such, J. M., Coté, M., & Criado, N. (2021). Bias and discrimination in AI: A cross-disciplinary perspective. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. https://doi.org/10.1002/widm.1356

Ginart, A., Kaung, M. M., Valiant, G., & Zou, J. (2019). Making AI forget you: Data deletion in machine learning. In H. Wallach et al. (Eds.), *Advances in Neural Information Processing Systems 32* (pp. 3513–3526). Curran Associates, Inc.

Google. (n.d.). *Google's approach to fighting misinformation online*. Google Safety Center. Retrieved April 2025, from https://safety.google/intl/en_us/stories/fighting-misinformation-online/

Google. (n.d.). *SafeSearch*. Retrieved April 2025, from https://www.google.com/safesearch

Google. (2019). *How Google fights disinformation*. https://kstatic.googleusercontent.com/files/388aa7d18189665e5f5579aef18e181c2d4283fb7b0d4691689dfd1bf92f7ac2ea6816e09c02eb98d5501b8e5705ead65af653cdf94071c47361821e362da55b

Google. (2019). *How we highlight fact checks in Search and Google News.* Google Blog. https://blog.google/outreach-initiatives/google-news-initiative/how-we-highlight-fact-checks-search-and-google-news/

Google. (2020). *Search quality evaluator guidelines*. https://static.googleusercontent.com/media/guidelines.raterhub.com/en//searchqualityevaluatorguidelines.pdf

Google. (2024). *Generative AI in search: Let Google do the searching for you*. Google Blog. https://blog.google/products/search/generative-ai-google-search-may-2024/

Google Developers. (2021). *Fact check*. https://developers.google.com/search/docs/data-types/factcheck

Google Search Central. (n.d.). *Spam Policies for Google Web Search*. Google for Developers. Retrieved April 2025, from https://developers.google.com/search/docs/essentials/spam-policies

Google Transparency Report. (2024a). *European privacy requests for search removals*. https://transparencyreport.google.com/eu-privacy/overview

Google Transparency Report. (2024b). *Removal requests*. https://transparencyreport.google.com/government-removals/overview

Google Transparency Report. (2024c). *U.S. national security*. https://transparencyreport.google.com/user-data/us-national-security

Gucluturk, O. (2024). *How to handle GDPR data access requests in AI-driven personal data processing*. OECD.AI. https://oecd.ai/en/wonk/gdpr-data-access-requests

Hannak, A., Sapiezynski, P., Kakhki, A. M., Krishnamurthy, B., Lazer, D., Mislove, A., & Wilson, C. (2013). Measuring personalization of web search. *Proceedings of the 22nd International Conference on World Wide Web*, 527–538. https://doi.org/10.1145/2488388.2488435

Hu, J., Zeng, H.-J., Li, H., Niu, C., & Chen, Z. (2007). Demographic prediction based on user's browsing behavior. *Proceedings of the 16th International Conference on World Wide Web*, 151–160. https://doi.org/10.1145/1242572.1242594

Jaźwińska, K., & Chandrasekar, A. (2025). AI search has a citation problem. *Columbia Journalism Review*. https://www.cjr.org/tow_center/we-compared-eight-ai-search-engines-theyre-all-bad-at-citing-news.php

Karnofsky, H. (2024). *A sketch of potential tripwire capabilities for AI*. Carnegie Endowment for International Peace. https://carnegieendowment.org/research/2024/12/a-sketch-of-potential-tripwire-capabilities-for-ai?lang=en

Lin, J., Dang, L., Rahouti, M., & Xiong, K. (2021). ML attack models: Adversarial attacks and data poisoning attacks. *arXiv*. https://arxiv.org/abs/2112.02797

Lin, Z., Cui, J., Liao, X., & Wang, X. (2024). Malla: Demystifying real-world large language model integrated malicious services. *arXiv*. https://arxiv.org/abs/2401.03315

Liu, H.-Y. (2018). The power structure of artificial intelligence. *Law, Innovation and Technology, 10*(2), 197–229. https://doi.org/10.1080/17579961.2018.1527480

Liu, S., Yao, Y., Jia, J., Casper, S., Baracaldo, N., Hase, P., Yao, Y., Liu, C. Y., Xu, X., Li, H., Varshney, K. R., Bansal, M., Koyejo, S., & Liu, Y. (2024). Rethinking machine unlearning for large language models. *arXiv*. https://arxiv.org/abs/2402.08787

Lukas, N., Salem, A., Sim, R., Tople, S., Wutschitz, L., & Zanella-Béguelin, S. (2023). Analyzing leakage of personally identifiable information in language models. *arXiv*. https://arxiv.org/abs/2302.00539

Mantelero, A. (2013). The EU proposal for a general data protection regulation and the roots of the 'right to be forgotten'. *Computer Law & Security Review, 29*(3), 229-235. https://www.sciencedirect.com/science/article/abs/pii/S0267364913000654?via%3Dihub

Matz, S., Teeny, J., Vaid, S. S., Peters, H., Harari, G. M., & Cerf, M. (2024). The potential of generative AI for personalized persuasion at scale. *Nature Scientific Reports*. https://www.nature.com/articles/s41598-024-53755-0

Metaxa, D., & Torres-Echeverry, N. (2017). Google's role in spreading fake news and misinformation. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3062984

Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo, C. A., Wallace, E., Tramèr, F., & Lee, K. (2023). Scalable extraction of training data from (production) language models. *arXiv*. https://arxiv.org/abs/2311.17035

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. NYU Press. https://doi.org/10.18574/nyu/9781479833641.001.0001

OpenAI. (n.d.). *Trust and transparency*. Retrieved April 2025, from https://openai.com/trust-and-transparency/

OpenAI. (2023a). *GPT-4 system card*. https://cdn.openai.com/papers/gpt-4-system-card.pdf

OpenAI. (2023b). *Preparedness framework (Beta)*. https://cdn.openai.com/openai-preparedness-framework-beta.pdf

OpenAI. (2024). *Introducing ChatGPT search*. https://openai.com/index/introducing-chatgpt-search/

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *arXiv*. https://arxiv.org/abs/2203.02155

Pariser, E. (2011). *The filter bubble: What the internet is hiding from you*. Penguin Press.

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press. http://www.jstor.org/stable/j.ctt13x0hch

Perrigo, B. (2023, January 18). Exclusive: OpenAI used Kenyan workers on less than $2 per hour to make ChatGPT less toxic. *TIME*. https://time.com/6247678/openai-chatgpt-kenya-workers/

Raman, D., Madkour, N., Murphy, E. R., Jackson, K., & Newman, J. (2025). *Intolerable risk threshold recommendations for artificial intelligence*. *arXiv*. https://arxiv.org/abs/2503.05812

Reid, L. (2024, May 14). Generative AI in Search: Let Google do the searching for you. Google. https://blog.google/products/search/generative-ai-google-search-may-2024/

Schmid, A. P. (Ed.). (2021). *Handbook of terrorism prevention and preparedness*. International Centre for Counter-Terrorism — ICCT. https://icct.nl/handbook-terrorism-prevention-and-preparedness

Similarweb. (n.d.). *Top websites ranking*. Retrieved April 15, 2025, from https://www.similarweb.com/top-websites/

Slattery, P., Saeri, A. K., Grundy, E. A., Graham, J., Noetel, M., Uuk, R., & Thompson, N. (2024). The AI risk repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence. *arXiv*. https://arxiv.org/abs/2408.12622

Solaiman, I., Talat, Z., Agnew, W., Ahmad, L., Baker, D., Blodgett, S. L., & Subramonian, A. (2023). Evaluating the social impact of generative AI systems in systems and society. *arXiv*. https://arxiv.org/abs/2306.05949

Spencer, S. (2022, May 4). Our 2021 Ads Safety Report. Google. https://blog.google/products/ads-commerce/ads-safety-report-2021/

Spirin, N., & Han, J. (2012). Survey on web spam detection. *ACM SIGKDD Explorations Newsletter, 13*(2), 50–64. https://doi.org/10.1145/2207243.2207252

Tanwar, P., & Poply, J. (2024). Navigating the AI IP nexus: Legal complexities and forward paths for intellectual property in the age of artificial intelligence. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4804599

van der Weij, T., Hofstätter, F., Jaffe, O., Brown, S. F., & Ward, F. R. (2024). AI sandbagging: Language models can strategically underperform on evaluations. *arXiv*. https://arxiv.org/abs/2406.07358

van Dijk, J. A. (2006). Digital divide research, achievements and shortcomings. *Poetics, 34*(4-5), 221–235.

Wei, A., Haghtalab, N., & Steinhardt, J. (2023). Jailbroken: How does LLM safety training fail? *arXiv*. https://arxiv.org/abs/2307.02483

Williams, A. R., Burke-Moore, L., Chan, R. S.-Y., Enock, F. E., Nanni, F., Sippy, T., Chung, Y.-L., Gabasova, E., Hackenburg, K., & Bright, J. (2024). Large language models can consistently generate high-quality content for election disinformation operations. *arXiv*. https://www.arxiv.org/abs/2408.06731

Wouters, P., Hellsten, I., & Leydesdorff, L. (2004). Internet time and the reliability of search engines. *First Monday, 9*(10).

Yeo, S. (2016). Geopolitics of search: Google versus China? *Media, Culture & Society, 38*(4), 591–605. https://doi.org/10.1177/0163443716643014

Zwetsloot, R., & Dafoe, A. (2019, February 11). Thinking about risks from AI: Accidents, misuse and structure. *Lawfare*. https://www.lawfareblog.com/thinking-about-risks-ai-accidents-misuse-and-structure


CLTC
Center for Long-Term
Cybersecurity
UC Berkeley