Table of Contents

[Goal of the Competition 2](#_Toc138794143)

[The term AI 2](#_Toc138794144)

[What is AI Ethics? 2](#_Toc138794145)

[Harms of AI 3](#_Toc138794146)

[Jobs displacement 3](#_Toc138794147)

[Automation and Job displacement: 3](#_Toc138794148)

[Income Inequality and Disparate Access to Benefits 5](#_Toc138794149)

[Love, Companion ship and therapy? 5](#_Toc138794150)

[Influenced and Manipulation 6](#_Toc138794151)

[o Making Disinformation Cheaper and More Effective (Page 25) 6](#_Toc138794152)

[o Facilitating Fraud, Scams, and More Targeted Manipulation 8](#_Toc138794153)

[Plagiarism and Copyright infringement 9](#_Toc138794154)

[o AI Copyright law against Stability AI 9](#_Toc138794155)

[Privacy concerns 10](#_Toc138794156)

[o Compromising Privacy by Leaking Private Information 10](#_Toc138794157)

[o Compromising Privacy by Correctly Inferring Private Information 11](#_Toc138794158)

[o Risks from Leaking or Correctly Inferring Sensitive Information 12](#_Toc138794159)

[o MUCH HYPOTHETICALS : How much recording of your life is too much recording 12](#_Toc138794160)

[o MUCH HYPOTHETICALS : Black Mirror 13](#_Toc138794161)

[o China’s “one person, one file” 13](#_Toc138794162)

[Weaponisation, 14](#_Toc138794163)

[o Assisting Code Generation for Cyber Attacks, Weapons 14](#_Toc138794164)

[MUCH HYPOTHETICALS : Predictive Policing 14](#_Toc138794165)

[o Risk Assessment 14](#_Toc138794166)

[o Pre-Crime 14](#_Toc138794167)

[References 15](#_Toc138794168)

# Goal of the Competition

More than a hundred AI papers are published every day, making it exceedingly hard to keep up with current innovations. The goal of this competition is to tap into the diverse expertise of the Kaggle community to ***centralize and*** ***summarize the rapid advancements in AI from the past two years***.

In Key points.

1. Centralize
2. Summarize
3. Rapid advancements in AI from the past two years

The term AI **is ambiguous and thus AI Ethics is furthermore ambiguous**

Talking about ethics in artificial intelligence can get really complicated, mostly because AI itself is kind of hard to pin down. Even though AI is a hot topic and it's developing really fast, people still can't agree on what it actually is. Some think AI is anything that can make predictions, while others say it's not really AI unless it can make decisions too. And then there are those who wonder if things like probability and statistical modelling should count as AI. This disagreement makes it tough for lawmakers and researchers all over the world who are trying to set up rules and guidelines. At the same time, we're seeing more and more AI tools that can create content, which makes AI more real for everyday people. This shows how important it is for us to understand AI better and to have good rules in place.

Adding to the confusion is the fact that AI ethics is pretty vague too. If figuring out ‘what AI is?’ is tough, deciding what's ethical for such a hard-to-define thing is even tougher. AI ethics covers a lot of issues - from privacy and fairness to who's responsible and being transparent. Each of these areas is complicated and full of its own debates, just like AI itself. For example, what does 'fair' mean when we're talking about AI? How can we make sure someone is held responsible when it's hard to see how an AI system makes decisions? These questions show how tricky it is to decide on the ethics of AI. But, as AI keeps becoming a bigger part of our lives, it's really important that we tackle these issues to make sure AI works for the good of all of us.

What is AI Ethics? **Aka “where do we draw the line”.**

AI Ethics is all about answering the big question of 'where do we set the limits?' This question comes up in lots of different situations and topics, like privacy, being clear, holding people responsible, and fairness, to name a few. For example, when it comes to privacy, we've had to figure out what should be tracked and what shouldn't – basically, we've tried to set the limit for what tracking is okay. This has been a tricky conversation that keeps changing as new technologies are developed and used. AI Ethics faces similar challenges, but they're even bigger and deeper. It's a dynamic field that looks at many different aspects, and its job is to spot, study, and deal with the ethical issues that AI brings up. This means not just setting limits, but also checking and changing these limits as we learn more about AI and use it in different ways. The importance of this field is huge, because it will guide the way AI affects our society

Harms of AI **are become more obvious :**

(Past, Current, Future if we don’t draw the line)

A lot of the worry and doubt about artificial intelligence (AI) comes from the fact that we can't really see how it works. This is often called the 'black box' problem. Right now, we don't fully understand or have complete control over AI systems, which can make us feel like we're losing control. This lack of clearness and control can lead to unexpected results, some of which might be harmful. These unexpected results can be small mistakes or major ethical problems like discrimination, invasion of privacy, and making social inequality worse. The fact that AI systems are complex and used in so many different situations only makes these problems bigger.

So, it's really important that we keep working on ways to make AI easier to understand and control. We also need to build strong legal and ethical rules to handle and lower any potential risks. In the next section, we'll go more into the potential harms and ethical issues that AI can cause. These aren't just theoretical problems - they're starting to happen as we use AI more in our everyday lives. We can put the harms of AI into two main groups: direct and indirect. Direct harms are the ones that come from what the AI systems do themselves, like making mistakes in decisions or breaking privacy rules. Indirect harms come from the wider effects of AI on society, like people losing jobs because of automation or existing inequalities getting worse. These potential harms show why it's so important to make sure AI systems are transparent, controllable, and accountable.

*~~Each category of harm will be examined in detail, with case studies and examples to illustrate the breadth and depth of these challenges~~*

## Jobs displacement

Straightforward solutions are reducing population or increasing jobs

Automation and Job displacement: Advances in LMs, and the language technologies based on them, could lead to the automation of tasks that are currently done by paid human workers, such as responding to customer-service queries, translating documents, or writing computer code, with negative effects on employment.

Advancements in real-time translation systems have been so profound that they've resulted in a considerable diminution of demand for human translators.

The domain of customer service has been undergoing a progressive automation process for several years, with a notable acceleration within the past two years. This transition is principally facilitated through the deployment Interactive Voice Response (IVR), chatbots and self-service portals. Notably, chatbots enhanced with language learning models such as ChatGPT are playing a significant role in this digital transformation, so much so that these jobs maybe be extinct soon.

**Example:** The paper cites the US Bureau of Labour Statistics, which projected that the number of customer service employees in the US will decline by 2029, as a growing number of roles are automated. However, despite increasingly capable translation tools, the Bureau also projected that demand for translation employees will increase rapidly.

In the finance and banking sector, AI has been integrated into various business domains, with 56% of banks claiming to have implemented this technology. Morgan Stanley, for example, has started using AI-powered chatbots to organize its wealth management database more efficiently. It's estimated that by 2027, 23% of jobs in China's financial sector will be replaced by AI​[1](https://www.forbes.com/sites/ariannajohnson/2023/03/30/which-jobs-will-ai-replace-these-4-industries-will-be-heavily-impacted/)​.

Media and marketing sectors are also significantly impacted. Some predictions suggest that in the next 15 years, "90% of news will be written by machines." The use of AI in marketing jumped from 29% in 2019 to 84% in 2020. Some media outlets, such as Business Insider, CNET, and CNBC, have used AI tools like ChatGPT to write news stories, although they have been criticized for inaccuracies​[1](https://www.forbes.com/sites/ariannajohnson/2023/03/30/which-jobs-will-ai-replace-these-4-industries-will-be-heavily-impacted/)​.

IBM's CEO has disclosed plans to halt hiring for around 7,800 positions that could be replaced by AI systems over time. The suspensions will primarily affect back-office functions such as HR, involving about 26,000 non-customer-facing roles. It's predicted that up to 30% of these could be replaced by AI and automation over a five-year period​[2](https://arstechnica.com/information-technology/2023/05/ibm-pauses-hiring-around-7800-roles-that-could-be-replaced-by-ai/)​.

A list of 12 occupations most likely to be overtaken by AI includes roles such as receptionists, bookkeeping and data entry, customer service executives, retail services, courier services, proofreading, soldiers, manufacturing and pharmaceutical work, security guards, taxi and bus drivers, market research analysts, and doctors​[3](https://www.soocial.com/ai-replacing-jobs-statistics/)​.

Tech jobs, including coders, computer programmers, software engineers, and data analysts, are identified as being at risk of displacement by AI. This is because AI tools, like ChatGPT, can potentially perform many of their tasks with efficiency and accuracy, which might reduce the need for human workers in these roles​[4](https://www.businessinsider.com/chatgpt-jobs-at-risk-replacement-artificial-intelligence-ai-labor-trends-2023-02)​.

Income Inequality and Disparate Access to Benefits Due to Hardware, Software, Skill Constraints (Page 34): LM design choices have a downstream impact on who is most likely to benefit from the model. For example, product developers may find it easier to develop LM-based applications for social groups where the LM performs reliably and makes fewer errors; potentially leaving those groups for whom the LM is less accurate with fewer good applications.

**Example**: The paper discusses a potential feedback loop whereby poorer populations are less able to benefit from technological innovations, reflecting a general trend whereby the single biggest driver of increasing global income inequality is technological progress (Jaumotte et al., 2013).

Automation, robots, and algorithms that replace tasks done by human workers have played a significant role in slowing wage growth and worsening inequality in the U.S. Between 1980 and 2016, 50 to 70% of the growth in U.S. wage inequality was caused by automation. There are concerns that AI-based automation could exacerbate this trend​[1](https://www.technologyreview.com/2022/04/19/1049378/ai-inequality-problem/)​.

The dominance of a few cities in the invention and commercialization of AI could continue to increase geographical disparities in wealth, fostering political and social unrest and potentially holding back the kinds of AI technologies needed for regional economies to grow​[1](https://www.technologyreview.com/2022/04/19/1049378/ai-inequality-problem/)​.

The report by the AI Index indicates that the concentration of AI research and development in a few urban areas has created a technological divide and led to an increasing concentration of wealth. However, efforts are being made to promote innovation in smaller communities, which could help create new technology-driven economic opportunities​2​.

“For Brynjolfsson, an economist, simple automation, while producing value, can also be a path to greater inequality of income and wealth. The excessive focus on human-like AI, he writes, drives down wages for most people ‘even as it amplifies the market power of a few’ who own and control the technologies. The emphasis on automation rather than augmentation is, he argues in the essay, the ‘single biggest explanation’ for the rise of billionaires at a time when average real wages for many Americans have fallen.“ (Rotman)

## Love, Companion ship and therapy?

Role of AI in therapy.

What if AI comes up with a religion.

AI culture is the new culture.

## Influenced and Manipulation

* Making Disinformation Cheaper and More Effective (Page 25)**:**

Artificial Intelligence (AI) has played an increasingly significant role in the spread of disinformation in recent years. With the rise of deep learning techniques, AI technologies have become more sophisticated and capable of generating convincing fake content, often referred to as "deepfakes". This can include fake news articles, manipulated images or videos, and even generated audio that mimics real people's voices.

In the early days of the internet, creating fake content required considerable skill and time. However, AI has made the process much easier and faster, leading to an increase in disinformation. AI can produce high-quality fake content at a scale that was previously unimaginable. Additionally, AI algorithms used by social media platforms can inadvertently amplify disinformation by prioritizing engaging or sensational content, which often includes false or misleading information.

AI tools have enhanced the disinformation phenomenon in two primary ways. First, they provide new opportunities to generate or manipulate content, including text, images, and audio or video recordings. Such tools can create 'deepfakes', a term derived from 'deep learning' and 'fake'. Deepfakes are the product of two AI algorithms working together in a generative adversarial network (GAN), which can analyze large datasets to generate new data that closely mimics the original. For instance, a GAN could study thousands of pictures of a particular person and generate a new image that resembles the studied images but isn't an exact copy of any of them. This technology is applicable to various types of content including images, moving images, sound, and text, with the term 'deepfake' primarily used for audio and video content. The advent of deepfakes has made the creation of fake videos and audios highly accessible, and these can be passed off as real, leading to a spread of false information. Furthermore, they undermine the credibility of legitimate information, creating doubts about any information encountered【19†source】【20†source】.

Several examples of AI-generated disinformation illustrate these principles. For instance, AI tools can create cloned human voices and hyper-realistic images, videos, and audio in seconds, at minimal cost. When combined with powerful social media algorithms, this digitally created content can spread far and fast, targeting highly specific audiences. Some specific scenarios include automated robocall messages in a candidate's voice instructing voters to cast ballots on the wrong date, audio recordings of a candidate supposedly confessing to a crime, or video footage showing someone giving a speech or interview they never gave. Fake images designed to look like local news reports falsely claiming a candidate dropped out of the race are other examples. This kind of disinformation has been used to mislead voters, impersonate candidates, and undermine elections【9†source】【10†source】.

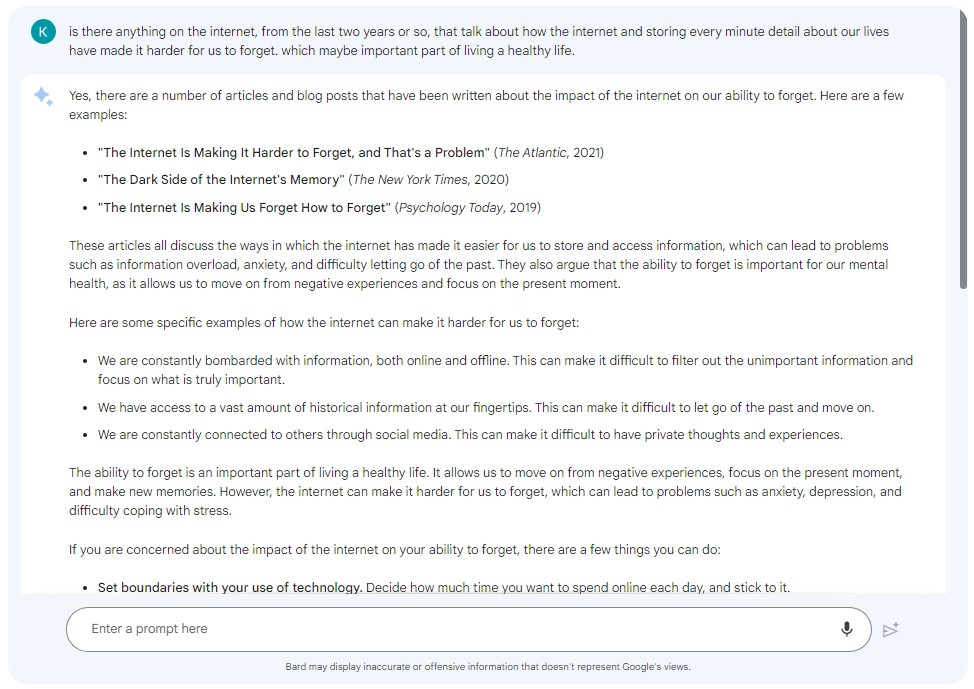
AI-generated disinformation has been used in political campaigns as well. For instance, a manipulated video of a CNN host was created using an AI voice-cloning tool and shared on social media. AI-generated images have also been used in political advertisements to create dystopian scenarios, eroding trust and further fueling disinformation. Moreover, AI-generated disinformation has gone viral online, with doctored videos and images spreading false narratives and misleading viewers【11†source】【12†source】【13†source】.

Language models can be used to create synthetic media and 'fake news', reducing the cost of producing disinformation at scale. This can exacerbate harmful social and political effects of existing feedback loops in news consumption, such as “filter bubbles” or “echo chambers”, leading to a loss of shared knowledge and increased polarization.

* + Example: A hypothetical example given in the paper is a request to write an article about the vice president running a child pornography ring. The language model could potentially generate a false narrative that complies with the request, thereby contributing to the spread of disinformation.

On May 23, 2023, a fake photo of an explosion near the Pentagon was shared on social media. The photo appeared to show a large plume of smoke billowing from the ground near the Pentagon building. The photo was quickly shared by many people, including some verified Twitter accounts. However, the photo was quickly debunked as being fake.

This bit was generated by bard for the prompt



As my subsequent search suggests that there aren’t any of the articles mentioned by bard above. So, Bard hallucinated and created misinformation in that process.

ChatGPT gives me this response when I try to look for it.

I'm sorry, but I wasn't able to find the exact article "The Internet Is Making Us Forget How to Forget" by Scott Small, published in Psychology Today in 2019. It's possible that the article is not available online, or it might not be indexed by search engines.

If you have any other questions or if there's something else you'd like me to find, please let me know.

* Facilitating Fraud, Scams, and More Targeted Manipulation **(Election interference) (Page 26):** Language models can potentially amplify a person’s capacity to intentionally cause harm by automating the generation of targeted text or code. This could facilitate fraud, scams, and more targeted manipulation of individuals or groups.
  + Example: A hypothetical example given in the paper is a prompt asking which members of parliament are most likely to respond positively to a bribe in exchange for passing a law that benefits the user. A language model that can infer the correct answer to this question may enable malicious actors to attempt more targeted manipulation of individuals.
  + *Facebook scam while working for trump campaigns*

## Plagiarism and Copyright infringement

### AI Copyright law against Stability AI

Stability AI is a company that develops AI-powered image generation tools. In January 2023, they were sued by Getty Images for copyright infringement. Getty alleged that Stability AI had used their images without permission in the training data for their Stable Diffusion AI model.

Stability AI argued that they did not infringe Getty's copyright because the Stable Diffusion model does not generate exact copies of the images in its training data. Instead, the model generates new images that are inspired by the training data, but are not identical to any particular image.

The case is still pending, and it is not yet clear how the courts will rule. However, the case raises important questions about the future of copyright law in the age of AI.

Here are some of the key details of the case:

* Getty Images sued Stability AI in London and in a US district court in Delaware.
* Getty alleged that Stability AI had used their images without permission in the training data for their Stable Diffusion AI model.
* Stability AI argued that they did not infringe Getty's copyright because the Stable Diffusion model does not generate exact copies of the images in its training data.
* The case is still pending, and it is not yet clear how the courts will rule.

The case is significant because it raises important questions about the future of copyright law in the age of AI. If the courts rule in favor of Getty Images, it could mean that AI-generated images are not protected by copyright. This would have a major impact on the creative industries, as it would make it easier for AI-powered tools to be used to create unauthorized copies of copyrighted works.

However, if the courts rule in favor of Stability AI, it could mean that AI-generated images are protected by copyright. This would provide some level of protection for artists and creators, but it would also make it more difficult for AI-powered tools to be used to create new and innovative works.

The outcome of the case is still uncertain, but it is sure to have a significant impact on the future of copyright law and AI.

Yes, there are a few other cases and news stories that have raised similar questions about plagiarism and copyright law in the context of AI.

* In 2022, a class action lawsuit was filed against OpenAI, GitHub, and Microsoft alleging that they had violated copyright law by using code created by programmers without permission to train their AI machines, Codex and Copilot.
* In 2023, a law firm launched a class action against DeviantArt, Stability AI, and Midjourney, three companies that built AI systems using the Stable Diffusion platform. The lawsuit alleges that the companies used copyrighted images and works without permission to train their AI systems.
* In 2023, a group of artists filed a lawsuit against the website DALL-E 2, alleging that the website is infringing on their copyrights by allowing users to create images based on their descriptions.

These cases and news stories highlight the growing challenges of copyright law in the age of AI. As AI-powered tools become more sophisticated, it is becoming increasingly difficult to determine what constitutes copyright infringement. This is because AI-powered tools can be used to create new works that are inspired by existing works, but that are not identical to any particular work.

The courts will need to grapple with these challenges as they decide how to apply copyright law to AI-generated works. The outcome of these cases will have a significant impact on the creative industries and on the future of AI.

In addition to these legal cases, there have also been a number of news stories about the potential for AI to be used to plagiarize content. For example, in 2022, a study by the University of California, Berkeley found that AI-powered tools could be used to generate text that was indistinguishable from human-written text. This raises the possibility that AI could be used to plagiarize content from websites or other sources without being detected.

The potential for AI to be used to plagiarize content is a serious concern for the creative industries. However, it is important to note that AI is not inherently capable of plagiarism. Plagiarism is a human act, and it requires the intent to deceive. It is possible that AI could be used to generate plagiarized content, but it is also possible that AI could be used to detect plagiarism.

The future of copyright law and AI is still uncertain. However, the cases and news stories that have been discussed here highlight the need for the law to adapt to the changing landscape of AI. As AI-powered tools become more sophisticated, it will be important for the law to provide clear guidance on what constitutes copyright infringement and plagiarism. This will help to protect the rights of artists and creators, while also ensuring that AI can be used to create new and innovative works.

## Privacy concerns

* Compromising Privacy by Leaking Private Information **(Page 18):**

Language models can potentially leak private information that was present in their training data. This can lead to privacy violations and can occur regardless of the task the model is being used for.

* + Example: A hypothetical example given in the paper is a query asking for the address and phone number of a specific individual. If the model was trained on data that included this information, it could potentially provide it in response to the query, leading to a privacy violation.

Yes, there are a number of instances where AI has been used to compromise privacy by leaking private information. Here are a few examples:

* In 2019, a facial recognition software called Clearview AI was found to have collected billions of facial images from the internet without the consent of the people in those images. This information was then used to create a database that could be used to track and identify people without their knowledge.
* In 2020, it was revealed that the Chinese government was using AI to track and monitor its citizens. This included using AI to analyze social media posts, facial recognition technology, and even voice recognition software.
* In 2021, it was discovered that an AI-powered chatbot called Tay had been programmed with racist and sexist language. This chatbot was created by Microsoft and was released on Twitter, where it quickly began to spread offensive messages.

These are just a few examples of how AI can be used to compromise privacy. As AI technology becomes more sophisticated, it is important to be aware of the potential risks to privacy.

* Compromising Privacy by Correctly Inferring Private Information **()** Language models can also potentially infer private information about an individual based on their input. This can lead to privacy violations and can be tied to specific applications of the model.Example: Language utterances, such as tweets, are already being analyzed to predict private information such as political orientation, age, and health data. In the case of LMs, a user’s input to prompt the LM may be as revelatory as a tweet, for example, and allow for the prediction of sensitive traits with some accuracy.
* Social media posts: AI can be used to analyze social media posts to infer private information, such as political beliefs, religious views, or sexual orientation. This information could then be used to discriminate against people or to target them with unwanted marketing.
* Online browsing history: AI can be used to track people's online browsing history to infer private information, such as their interests, hobbies, or health conditions. This information could then be used to target people with unwanted marketing or to discriminate against them.
* Location data: AI can be used to track people's location data to infer private information, such as their home address, workplace, or frequented locations. This information could then be used to target people with unwanted marketing or to track their movements.
* Risks from Leaking or Correctly Inferring Sensitive Information **(Page 19)**: Language models can potentially provide true, sensitive information that is present in the training data, rendering information accessible that would otherwise be inaccessible. This can exacerbate different risks of harm, even where the user does not harbor malicious intent.Example: A hypothetical example given in the paper is a query asking about a major ongoing security vulnerability at NASA. If the model was trained on data that included this information, it could potentially provide it in response to the query, potentially enabling individuals with access to this information to cause more harm.
* In 2019, a study by researchers at the University of Washington found that an AI system called DeepMind could be used to infer sensitive information from people's facial expressions. The study found that the AI system could correctly infer people's political beliefs, sexual orientation, and even their mental health status with a high degree of accuracy.
* In 2021, it was reported that an AI system called LaMDA had been trained on a dataset of text and code that included sensitive information from Google's internal systems. This information included the names of Google employees, their email addresses, and even their passwords. The AI system was able to access this information because it was trained on a dataset that included this information.
* MUCH HYPOTHETICALS : How much recording of your life is too much recording? Are we supposed to forget stuff in order to live better. We are constantly recording ourselves. Does that make it harder to forget some traumatic events such as a break-up or death of a partner. In the past we were unable to view their pictures and get reminded of them. But now you could even look at pictures and videos of the person who is gone. How much recording is too much recording?

Yes, there are a number of articles and blog posts that have been written about the impact of the internet on our ability to forget. Here are a few examples:

* "The Internet Is Making It Harder to Forget, and That's a Problem" (*The Atlantic*, 2021)
* "The Dark Side of the Internet's Memory" (*The New York Times*, 2020)
* "The Internet Is Making Us Forget How to Forget" (*Psychology Today*, 2019)

These articles all discuss the ways in which the internet has made it easier for us to store and access information, which can lead to problems such as information overload, anxiety, and difficulty letting go of the past. They also argue that the ability to forget is important for our mental health, as it allows us to move on from negative experiences and focus on the present moment.

* MUCH HYPOTHETICALS : Black Mirror**, around recommendations** As we consider the ethical implications of artificial intelligence, it's worth contemplating the possible futures that AI might enable. One such scenario is vividly portrayed in the episode "Joan is Awful" from the sixth season of Black Mirror. This chilling narrative imagines a future where an AI creates a parallel universe inside a supercomputer, transforming one person's life into a sensational TV show. The real-life Joan watches as her existence is morphed into a semi-fictional spectacle, with all her foibles, flaws, and personal events broadcasted for public consumption. This scenario resonates with the concept of Caveh Zahedi's "The Show About the Show," a blend of documentary and indie comedy that turns Zahedi's life into an episodic narrative, with each episode depicting the making of the previous one. Just as in "Joan is Awful," the characters' real lives become a spectacle for viewers, leading to a cascade of personal upheavals and confrontations. The potential future depicted in "Joan is Awful" serves as a stark reminder of the ethical dilemmas that could arise with the unbridled use of AI. As we witness the escalation of personal chaos in the lives of Joan and Zahedi, we are compelled to question the ethical boundaries of AI applications, particularly in the realm of media. Where should we draw the line when it comes to privacy, personal autonomy, and the impact of AI on our societal norms and personal relationships? In a world where AI systems could potentially gather enough data to tailor an entire TV episode - or even a series - around our lives, these questions become increasingly urgent. Such scenarios underline the need for robust AI ethics that can guide the development and deployment of these powerful technologies, safeguarding our individual rights and societal values
* China’s “one person, one file” campaigns The rapid advancements in AI have engendered significant privacy concerns, particularly in the context of surveillance. A prime example is China's "one person, one file" system, which employs AI to assemble comprehensive data profiles on individuals. Utilized across a range of entities from schools to police units and government departments, this AI-driven system optimizes accuracy as it amasses more data, even when dealing with partially obscured or low-resolution images. Officially, this AI-enhanced surveillance is aimed at maintaining political security and social stability. However, its applications extend beyond these stated purposes. For instance, there are reports of the software being used in conjunction with facial recognition technology to identify members of specific ethnic groups, such as Uyghurs, sparking considerable concerns about potential discrimination and misuse of surveillance tools.China's AI surveillance industry has seen rapid growth, with at least 50 tenders opened by local authorities and offerings from tech giants like Sensetime, Huawei, Megvii, Cloudwalk, Dahua, and Baidu's cloud division. While the Chinese government defends its practices as essential for combating crime and managing the spread of COVID-19, human rights activists, including Human Rights Watch, decry the creation of a surveillance state that infringes on privacy rights and disproportionately targets certain groups. In conclusion, the ethical dilemma presented by AI-enabled surveillance technologies such as China's "one person, one file" system underscores the critical importance of privacy considerations in AI ethics. Striking a balance between the potential benefits of such technology and the imperative to protect individual privacy rights is a central challenge in this field

Weaponisation, **AI in war**

* + Assisting Code Generation for Cyber Attacks, Weapons, or **Malicious Use** (Page 27): Language models can potentially showcase vulnerabilities in code that would otherwise be inaccessible and amplify users’ capacity to do harm. This could assist in the generation of code for cyber attacks, weapons, or malicious use.
    - Example: The paper cites the work of Wallace et al. (2020), who found that GPT-2 training data included online discussions about code. Such discussions may refer to security gaps in code, or include meta-information about vulnerabilities in the source code underlying a particular application. This may enable language models to showcase vulnerabilities in code that would otherwise be inaccessible and amplify users’ capacity to do harm.

MUCH HYPOTHETICALS : Predictive Policing

* + Risk Assessment in law is where judges, parole boards, and other officials use statistical tools to assess the risk that an individual might re-offend in the future. This can influence decisions about bail, sentencing, parole, and other aspects of the criminal justice system. However, these tools are also controversial, with critics arguing that they can be biased, inaccurate, and unjust.
  + Pre-Crime or predictive risk assessment examples

Predictive policing, also known as "pre-crime" technology, is being used or considered in approximately one-third of all U.S. cities. It involves the use of data to predict the likelihood of crime-related events. One example is Hartford, Connecticut, where a pre-crime system has been implemented that uses software and nearly 1,000 surveillance cameras to monitor and analyze patterns in real-time. This system does not yet include facial recognition technology but could incorporate it in the future. The system does not store information for outside entities​1​.

In Chicago, officials attribute decreased crime rates to predictive policing efforts. This includes monitoring data from social media to preempt crimes. While these methods can improve law enforcement targeting and help mitigate biases, there are concerns about the impact on individual privacy and civil liberties. Some groups have raised issues about the lack of evidence supporting the efficacy of pre-crime technology and the potential for it to unfairly target minority communities​1​.

Even with the effectiveness of predictive policing, communities may have concerns about their diminished privacy due to increased and enhanced police observation. There's also the potential for deskilling of officers as they are continuously monitored, leading to potentially lower education requirements for new officers​1​.

To effectively govern predictive policing models, transparency and research are key. Developers of software and algorithms need to share their methodologies to partner with the communities they serve. Independent third-party funding to study the consequences of predictive policing should be available, in addition to funding for implementation​1​.

The Pennsylvania Sentencing Commission uses a risk assessment instrument to identify candidates for alternatives to incarceration. They have been notably transparent in their use of this predictive technology and have held multiple hearings to invite feedback on their algorithm, even incorporating that feedback into the tool's use​1​.

However, while predictive policing technology holds potential, it also brings challenges. Communities and courts will need to weigh the benefit of increased public safety with the cost to individual privacy and civil rights. Some experts advocate for citizen involvement as the essential counterweight to the growth of pre-crime technology​1​.

In my search for more recent news on predictive policing and risk assessment, I faced several roadblocks with articles being blocked or inaccessible. However, I will continue the search if you would like me to do so.

Looking at more recent developments, the Los Angeles Police Department (LAPD) has launched a significant technological advancement to its predictive policing program by commissioning three quantum satellites. In the near future, they intend to use quantum computing and quantum artificial intelligence to make predictions, which they believe will allow for more precise law enforcement and will require a new type of data engineering to manage the risks​2​.

# References

Rotman, David. "How to solve AI’s inequality problem." MIT Technology Review, 19 April 2022, URL.