Table of Contents

[The term AI 1](#_Toc138691165)

[What is AI Ethics? 2](#_Toc138691166)

[Harms of AI 2](#_Toc138691167)

[Jobs displacement 3](#_Toc138691168)

[Increasing Inequality and Negative Effects on Job Quality 3](#_Toc138691169)

[Disparate Access to Benefits 3](#_Toc138691170)

[Love, Companion ship and therapy? 4](#_Toc138691171)

[Influenced and Manipulation 4](#_Toc138691172)

[o Making Disinformation Cheaper and More Effective (Page 25) 4](#_Toc138691173)

[o Facilitating Fraud, Scams, and More Targeted Manipulation 4](#_Toc138691174)

[Plagiarism and Copyright infringement 4](#_Toc138691175)

[o AI Copyright law against Stability AI 4](#_Toc138691176)

[Privacy concerns 4](#_Toc138691177)

[o Compromising Privacy by Leaking Private Information 4](#_Toc138691178)

[o Compromising Privacy by Correctly Inferring Private Information 5](#_Toc138691179)

[o Risks from Leaking or Correctly Inferring Sensitive Information 5](#_Toc138691180)

[o MUCH HYPOTHETICALS : How much recording of your life is too much recording 5](#_Toc138691181)

[o MUCH HYPOTHETICALS : Black Mirror 5](#_Toc138691182)

[o China’s “one person, one file” 6](#_Toc138691183)

[Weaponisation, 6](#_Toc138691184)

[o Assisting Code Generation for Cyber Attacks, Weapons 6](#_Toc138691185)

[MUCH HYPOTHETICALS : Predictive Policing 6](#_Toc138691186)

[o Risk Assessment 6](#_Toc138691187)

[o Pre-Crime 7](#_Toc138691188)

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

The discourse on ethics in artificial intelligence is both abstract and complex, owing largely to the nebulous nature of AI itself. Despite the buzz and rapid advancements surrounding AI, the term remains open to interpretation and debate. Some view AI as systems capable of prediction, while others argue that true AI must encompass both prediction and decision-making abilities. Yet, others question if probability and statistical modelling should be considered part of AI. This lack of consensus presents a significant challenge for lawmakers and researchers globally as they strive to establish comprehensive rules and guidelines. Simultaneously, we are witnessing a swift uptake of AI tools capable of creating content, making the concept of AI more tangible to the general public. This growing ubiquity of AI underscores the urgent need for better understanding and regulation

Further complicating matters is the inherent ambiguity of AI ethics itself. If defining AI is a challenge, defining ethical boundaries for a nebulous concept is doubly so. AI ethics encompasses a wide range of concerns - from privacy and fairness to accountability and transparency. Each of these areas is complex and fraught with its own controversies and debates, mirroring the complexity of AI itself. For instance, what constitutes 'fair' in an AI context? How can we ensure 'accountability' when the decision-making process of an AI system is opaque? These questions highlight the ambiguities and challenges we face in shaping the ethical framework of AI. Yet, as AI continues to permeate every aspect of our lives, it is imperative that we confront and navigate these ambiguities to ensure that AI serves the best interests of humanity.

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

AI Ethics, at its core, grapples with the crucial question of 'where do we draw the line?' This question resonates across a myriad of contexts and issues, including privacy, transparency, accountability, and fairness, among others. For instance, in the realm of privacy, we have wrestled with determining what should be tracked and what should not – essentially, we've sought to define the boundary of acceptable tracking. This has been a complex, evolving conversation that mirrors the rapid development and deployment of new technologies. AI Ethics, as a field, is tasked with similar challenges, but with an even broader and more profound scope. It is a dynamic, multi-faceted discipline that seeks to identify, analyze, and address the ethical implications of AI. This involves not only drawing lines but also constantly re-evaluating and adjusting these lines as our understanding and use of AI evolves. The enormity and significance of this field cannot be overstated, as it will shape the trajectory of AI's impact on society.

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

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

Much of the skepticism and concern around artificial intelligence (AI) emanates from the opacity of its inner workings. This is often referred to as the 'black box' phenomenon. Our current understanding and technological capabilities limit our ability to fully control and predict the behavior of AI systems, leading to an unsettling sense of a loss of control. This lack of transparency and control can result in unintended consequences, some of which may be detrimental. These unintended outcomes can range from minor errors to major ethical issues such as discrimination, privacy invasion, and increased social inequality. The complexity of AI systems, coupled with the vast and diverse contexts in which they are deployed, amplifies these challenges.

It is essential, therefore, to continue developing strategies for making AI more explainable and controllable, as well as creating robust legal and ethical frameworks to manage and mitigate potential risks.

In the following section, we delve deeper into the potential harms and ethical dilemmas posed by AI. These issues are not merely speculative, but have real-world implications that are beginning to materialize as AI becomes more integrated into our daily lives.

The harms of AI can be broadly categorized into two types: direct and indirect. Direct harms are those that result from the actions of AI systems themselves, such as errors in decision-making or the violation of privacy. Indirect harms, on the other hand, are those that stem from the broader societal implications of AI, such as job displacement due to automation or exacerbation of existing inequalities.

These potential harms underscore the urgency of ensuring transparency, control, and accountability in AI systems.

*~~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

Increasing Inequality and Negative Effects on Job Quality (Page 32-33): 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.

**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.

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).

## 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)**:** 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.
  + Latest leaked photo of blast in the pentagon
* 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

## 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.
* 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.
* 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.
* 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?
* 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​.