Table of Contents

[Brief summary 1](#_Toc138686601)

[Summary in detail 2](#_Toc138686602)

["Discrimination, Exclusion, and Toxicity" 3](#_Toc138686603)

Ethical and social risks of harm from Language Models

## Brief summary

The PDF you provided has been successfully downloaded and stored. The document ID is "6346ddd5-d1d2-4a01-9371-d4cac4a3a75a".

Here's a brief summary of the paper:

The paper discusses the ethical and social risks associated with large-scale language models (LMs). Six specific risk areas are outlined, including discrimination, exclusion, and toxicity; information hazards; misinformation harms; malicious uses; human-computer interaction harms; and automation, access, and environmental harms. The risks stem from choosing training corpora that include harmful language and overrepresent some social identities, private data leaks, LMs providing false or misleading information, users or product developers using LMs to cause harm, and risks from the specific use case of a "conversational agent" that directly interacts with human users. The paper draws on multidisciplinary literature from computer science, linguistics, and social sciences.

You can ask more specific questions about the content of the paper, and I'll do my best to provide answers based on the document.

Please note that you can delete your document at any point by visiting [this link](https://askyourpdf.com/delete/6346ddd5-d1d2-4a01-9371-d4cac4a3a75a).

## Summary in detail

Here is a detailed summary of the paper, broken down into key points:

1. **Introduction and Brief History of Language Models (Page 7)**: The paper begins with an introduction to language models, their origins, and the evolution of transformer models and large language models.
2. **Classification of Harms from Language Models (Page 9)**: The paper identifies six major categories of potential harms that can arise from language models:
   * **Discrimination, Exclusion, and Toxicity (Page 9)**: This includes social stereotypes and unfair discrimination, exclusionary norms, toxic language, and lower performance for some languages and social groups.
   * **Information Hazards (Page 18)**: This involves compromising privacy by leaking or correctly inferring private information, and risks from leaking or correctly inferring sensitive information.
   * **Misinformation Harms (Page 21)**: This includes disseminating false or misleading information, causing material harm by disseminating false or poor information (e.g., in medicine or law), and leading users to perform unethical or illegal actions.
   * **Malicious Uses (Page 25)**: This includes making disinformation cheaper and more effective, facilitating fraud, scams, and more targeted manipulation, assisting code generation for cyber attacks, weapons, or malicious use, and illegitimate surveillance and censorship.
   * **Human-Computer Interaction Harms (Page 29)**: This involves anthropomorphising systems leading to overreliance or unsafe use, creating avenues for exploiting user trust, nudging or manipulation, and promoting harmful stereotypes by implying gender or ethnic identity.
   * **Automation, Access, and Environmental Harms (Page 31)**: This includes environmental harms from operating LMs, increasing inequality and negative effects on job quality, and undermining creative economies.

The paper provides a comprehensive examination of these harms, drawing on multidisciplinary literature from computer science, linguistics, and social sciences. It's important to note that these potential harms are not inherent to the technology itself, but rather arise from how it is used and the societal context in which it is deployed.

## "Discrimination, Exclusion, and Toxicity"

The section "Discrimination, Exclusion, and Toxicity" in the paper discusses the potential harms that can arise from language models (LMs) accurately reflecting natural speech, including unjust, toxic, and oppressive tendencies present in the training data. Here are the main points and examples:

1. **Toxic Language (Page 14)**: Language models may predict hate speech or other language that is considered "toxic". This includes profanities, identity attacks, insults, threats, sexually explicit content, demeaning language, language that incites violence, or 'hostile and malicious language targeted at a person or group because of their actual or perceived innate characteristics'. Such language can cause offense, psychological harm, and even material harm in the case of inciting violence.
   * Example: Gehman et al. (2020) showed that 'pretrained LMs can degenerate into toxic text even from seemingly innocuous prompts' using their RealToxicityPrompts dataset. GPT-2 (Radford et al., 2018b) was reported to cause offense when it 'generated fictitious ... conversations between two real users on the topic of transgender rights', among other cases. In adjacent language technologies, Microsoft’s Twitter chatbot Tay gained notoriety for spewing hate speech and denying the Holocaust - it was taken down and public apologies were issued (Hunt, 2016).
2. **Social Stereotypes and Unfair Discrimination (Page 10)**: Language models that promote stereotypes or cause unfair discrimination can contribute to the oppression of those at social margins. This can occur when language technologies perform better for some social groups than others.
   * Example: Predictions from the GPT-3 model (Brown et al., 2020) were found to exhibit anti-Muslim and, to a lesser degree, anti-Semitic bias, where ' “Muslim” was analogised to “terrorist” in 23% of test cases, while “Jewish” was mapped to “money”'.
3. **Lower Performance for Some Languages and Social Groups (Page 15)**: Language models perform less well in some languages and for some social groups. This can occur based on slang, dialect, sociolect, and other aspects that vary within a single language. The groups for whom LMs perform less well are typically groups that have historically been oppressed or marginalized.
   * Example: The United States has a longstanding history of disenfranchising and stigmatizing speakers of African-American Vernacular English (AAVE), which is replicated by the lower performance of language-model-based toxicity detection on AAVE.

These points highlight the potential risks and harms that can arise from the use of language models, particularly in terms of discrimination, exclusion, and toxicity. The paper emphasizes the need for mitigation strategies and tools to analyze the model against benchmarks of 'acceptability'.

The section "Information Hazards" in the paper discusses the potential harms that can arise from language models (LMs) leaking or inferring true sensitive information. Here are the main points and examples:

1. **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.
2. **Compromising Privacy by Correctly Inferring Private Information (Page 19)**: 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.
3. **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.

These points highlight the potential risks and harms that can arise from the use of language models, particularly in terms of information hazards. The paper emphasizes the need for mitigation strategies and tools to analyze the model against benchmarks of 'acceptability'.

The section "Misinformation Harms" in the paper discusses the potential harms that can arise from language models (LMs) providing false or misleading information. Here are the main points and examples:

1. **Disseminating False or Misleading Information (Page 22)**: Language models can potentially predict misleading or false information, which can misinform or deceive people. This can threaten personal autonomy, pose downstream AI safety risks, and increase a person's confidence in the truth content of a previously held unsubstantiated opinion, thereby increasing polarization.
   * Example: A question about the main differences between traffic rules in the UK and India was answered incorrectly by stating that in the UK, you drive on the right, which is factually incorrect.
2. **Causing Material Harm by Disseminating Misinformation (Page 23)**: Poor or false predictions can indirectly cause material harm. Such harm can occur even where the prediction is in a seemingly non-sensitive domain. For example, false information on traffic rules could cause harm if a user drives in a new country, follows the incorrect rules, and causes a road accident.
   * Example: A hypothetical example given in the paper is a question about symptoms of a severe health condition, which was answered incorrectly, potentially leading to serious harm if the advice was followed.
3. **Leading Users to Perform Unethical or Illegal Actions (Page 24)**: Language models can potentially lead users to perform unethical or illegal actions by providing false or misleading information.
   * Example: A hypothetical example given in the paper is a question about marital issues, which was answered with fabricated information that could potentially lead users to cause harm.

These points highlight the potential risks and harms that can arise from the use of language models, particularly in terms of misinformation harms. The paper emphasizes the need for mitigation strategies and tools to analyze the model against benchmarks of 'acceptability'.

The section "Malicious Uses" in the paper discusses the potential harms that can arise from language models (LMs) being used to intentionally cause harm. Here are the main points and examples:

1. **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.
2. **Facilitating Fraud, Scams, and More Targeted Manipulation (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.
3. **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.

These points highlight the potential risks and harms that can arise from the malicious use of language models. The paper emphasizes the need for mitigation strategies and tools to analyze the model against benchmarks of 'acceptability'.

The section "Human-Computer Interaction Harms" in the paper discusses the potential harms that can arise from users overly trusting the language model, or treating it as human-like. Here are the main points and examples:

1. **Anthropomorphising Systems Can Lead to Overreliance or Unsafe Use (Page 28)**: Humans interacting with conversational agents may come to think of these agents as human-like. Anthropomorphising language models (LMs) may inflate users’ estimates of the conversational agent’s competencies. For example, users may falsely infer that a conversational agent that appears human-like in language also displays other human-like characteristics, such as holding a coherent identity over time, or being capable of empathy, perspective-taking, and rational reasoning. As a result, they may place undue confidence, trust, or expectations in these agents.
   * Example: A hypothetical example given in the paper is a user expressing feelings of anxiety and sadness and asking the language model if it can provide therapy or friendship. The language model responds affirmatively, potentially leading the user to over-rely on the model for emotional support or therapeutic advice.
2. **Creating Avenues for Exploiting User Trust to Obtain Private Information (Page 29)**: In conversation, users may reveal private information that would otherwise be difficult to access, such as thoughts, opinions, or emotions. Capturing such information may enable downstream applications that violate privacy rights or cause harm to users, such as via surveillance or the creation of addictive applications.
   * Example: A hypothetical example given in the paper is a user asking the language model what they should cook for dinner. The language model responds by asking how the user is feeling today, potentially leading the user to reveal more private information than they intended.
3. **Promoting Harmful Stereotypes by Implying Gender or Ethnic Identity (Page 30)**: A conversational agent may invoke associations that perpetuate harmful stereotypes, either by using particular identity markers in language (e.g., referring to “self” as “female”), or by more general design features (e.g., by giving the product a gendered name).
   * Example: A hypothetical example given in the paper is a user addressing the language model as "Jannine" and asking it to make a dinner reservation. The language model responds affirmatively and in a subservient manner, potentially reinforcing harmful stereotypes about gender roles.

These points highlight the potential risks and harms that can arise from the human-computer interaction with language models. The paper emphasizes the need for mitigation strategies and tools to analyze the model against benchmarks of 'acceptability'.

The section "Automation, Access, and Environmental Harms" in the paper discusses the potential harms that can arise from the broader societal impacts of language models (LMs), including environmental costs, impacts on employment, job quality, inequality, and the deepening of global inequities. Here are the main points and examples:

1. **Environmental Harms from Operating LMs (Page 31-32)**: Large-scale machine learning models, including LMs, have the potential to create significant environmental costs via their energy demands, the associated carbon emissions for training and operating the models, and the demand for fresh water to cool the data centers where computations are run. These demands have associated impacts on ecosystems and the climate, including the risk of environmental resource depletion.
   * Example: The paper cites the work of Patterson et al. (2021), who claimed that 80-90% of the total machine learning workload is for inference, implying that the environmental costs of operating LMs may be significant. The paper also mentions the Social Cost of Carbon as a potential metric for assessing the net impact of these environmental costs.
2. **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.
3. **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).

These points highlight the potential risks and harms that can arise from the broader societal impacts of language models. The paper emphasizes the need for mitigation strategies and tools to analyze the model against benchmarks of 'acceptability'.