The Only Way is Ethics: A Guide to Ethical Research with Large Language Models

Eddie L. Ungless[1], Nikolas Vitsakis[2], Zeerak Talat[1], James Garforth[1], Björn Ross[1], Arno Onken[1], Atoosa Kasirzadeh[1], Alexandra Birch[1]
[1]University of Edinburgh [2]Heriot-Watt University
a.birch@ed.ac.uk

Abstract

There is a significant body of work looking at the ethical considerations of large language models (LLMs): critiquing tools to measure performance and harms; proposing toolkits to aid in ideation; discussing the risks to workers; considering legislation around privacy and security etc. As yet there is no work that integrates these resources into a single practical guide that focuses on LLMs; we attempt this ambitious goal. We introduce LLM ETHICS WHITEPAPER, which we provide as an open and living resource for NLP practitioners, and those tasked with evaluating the ethical implications of others' work. Our goal is to translate ethics literature into concrete recommendations and provocations for thinking with clear first steps, aimed at computer scientists. LLM ETHICS WHITEPAPER distils a thorough literature review into clear Do's and Don'ts, which we present also in this paper. We likewise identify useful toolkits to support ethical work. We refer the interested reader to the full LLM ETHICS WHITEPAPER, which provides a succinct discussion of ethical considerations at each stage in a project lifecycle, as well as citations for the hundreds of papers from which we drew our recommendations. The present paper can be thought of as a pocket guide to conducting ethical research with LLMs.

1 Introduction

As LLMs grow increasingly powerful, their advancements in natural language understanding and generation are impressive (Min et al., 2023). However, mitigating the risks they present remains a complex challenge, and categorising these risks is a crucial aspect of ethical research related to LLMs (Weidinger et al., 2022). Key concerns include the potential to perpetuate and even amplify existing biases present in training data (Gallegos et al., 2024), challenges in safeguarding user privacy (Yao et al., 2024), hallucination or incorrect responses (Abercrombie et al., 2023; Xu et al., 2024), malicious use of their capabilities (Cuthbertson, 2023), and infringement of copyright (Lucchi, 2023). Given that many of these ethical challenges remain unresolved, it is essential for those involved in developing LLMs and LLM-based applications to consider potential harms, particularly as these models see broader adoption.

Several frameworks have already been developed to address AI ethics and safety. For example The U.S. National Institute of Standards and Technology (NIST) has an AI Risk Management Framework (RMF)[1], which provides broad guidelines for managing AI-related risks. NIST has also recently released a document outlining specific risks and recommended actions for Generative AI[2]. Whilst widely adopted, the NIST guidelines are voluntary. In contrast, the EU AI Act[3] represents a legally binding regulatory framework designed to ensure the safe and ethical use of AI within the European Union. It emphasises transparency, human oversight, and the prevention of discriminatory outcomes, with the goal of protecting fundamental rights and promoting trustworthy AI.

The NIST AI RMF and EU AI Act are broad, focusing on AI deployment and risk management across industries. There are other frameworks which are more research-focused, guiding ethical considerations in academic AI work. For example the Conference on Neural Information Processing Systems (NeurIPS) Ethics Guidelines[4] evaluates AI research for ethical concerns as part of the paper submission process. A similar effort from the Association of Computational Linguistics (ACL) has created an Ethics Checklist which guides authors

[1] https://www.nist.gov/itl/ai-risk-management-framework
[2] https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf
[3] https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai
[4] https://neurips.cc/public/EthicsGuidelines

Proceedings of the 31st International Conference on Computational Linguistics, pages 8992–9005
January 19–24, 2025.

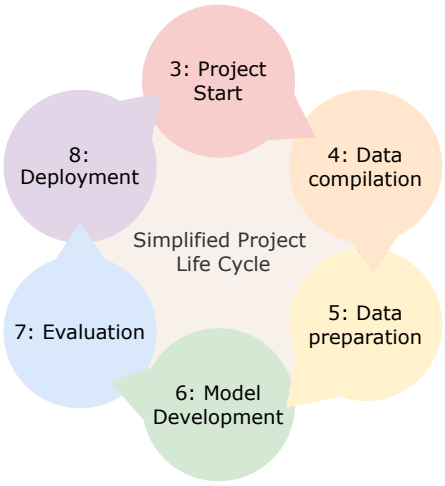

Figure 1: Diagram showing simplified project lifecycle that forms the structure of this paper. This reflects section numbers in the current paper.

in addressing ethical implications, including limitations, and correct treatment of human annotators.[5]

Despite there being a number of frameworks for the ethical development of AI, we believe that there is still a need for a practical whitepaper focused on the needs of a practitioner working with LLMs. To meet this need, we have created our Ethics Whitepaper[6] (Ungless et al., 2024), henceforth LLM ETHICS WHITEPAPER.

LLM ETHICS WHITEPAPER presents insight and pointers to the most relevant ethical research, as it relates to each of the steps in the project lifecycle. It provides more detail than the guidelines of NeurIPS and ACL, but is more "digestible" and directly applicable to research with LLMs than the NIST frameworks or the EU AI act. We hope our LLM ETHICS WHITEPAPER, and this Overview paper, will prove valuable to all practitioners, whether you are looking for succinct best practice recommendations, a directory of relevant literature, or an introduction to some of the controversies in the field.

LLM ETHICS WHITEPAPER and this Overview are both structured around a (simplified) project lifecycle, as depicted in Figure 1. Our aim for these documents is for them to be used as a reference guide throughout a project, rather than for post-hoc reflection. We begin in Section 3 by outlining the importance of ethics and discuss themes relevant to the entire development life cycle. An extended version of this Section can also be found in LLM ETHICS WHITEPAPER. In each of the following sections of the present paper we summarise the main topics covered in LLM ETHICS WHITEPAPER and present key resources. Specifically, Do's and Don'ts which encompass concrete steps or clear provocations for thinking, which were directly drawn from our extensive literature review, plus tools to guide ethical work. Interested readers should refer to LLM ETHICS WHITEPAPER for more details plus full references, which is hosted on Github[7] to facilitate continual feedback from NLP practitioners. We conclude with Limitations and plans for future developments of LLM ETHICS WHITEPAPER.

2 Methodology

A primary goal of LLM ETHICS WHITEPAPER is to provide a comprehensive directory of resources for ethical research related to LLMs. As such, a systematic literature review of the ACL Anthology was conducted (note that LLM ETHICS WHITEPAPER is not itself strictly a systematic literature review paper, see below). The Anthology was searched for paper abstracts containing at least one term from the following key term lists: related to the type of resource = `tool[a-z]*`, `toolkit`, `[A-za-z]*sheets*`, `guidelines*`, `principles`, `framework`, `approach`; related to ethics = `ethic[a-z]*`, `harms*`, `fair[a-z]*`, `risks*`. These lists were determined by first using more comprehensive lists then eliminating terms to improve the precision of the search. The resulting papers were manually reviewed to determine which were relevant to the scope of LLM ETHICS WHITEPAPER. During the search we identified a 2023 EACL tutorial titled "Understanding Ethics in NLP Authoring and Reviewing" (Benotti et al., 2023). The references for this tutorial were manually reviewed and where relevant included in LLM ETHICS WHITEPAPER.

A second literature review was conducted using Semantic Scholar using the search terms: `toolkit OR sheets OR guideline OR principles OR framework OR approach ethics OR ethical OR harms OR fair OR fairness OR risk AND "language models"`. These were likewise manually reviewed for inclusion.

The resulting resources were categorised by their

[5] https://aclrollingreview.org/responsibleNLPresearch/

[6] https://doi.org/10.48550/arXiv.2410.19812

[7] https://github.com/MxEddie/Ethics-Whitepaper

relevance to different stages in a project's lifespan (from ideation to deployment). Primary themes in the literature were identified and used to structure each section; themes were identified by the first author using a bottom-up approach based on the ethical issue(s) each identified paper addressed. Themes were then discussed with all authors and refined in the context of further papers familiar to the authors.

As LLM ETHICS WHITEPAPER progressed, additional resources familiar to the authors were added *ad hoc*. Additionally, papers identified during the research review papers were removed for a number of reasons, either because they were deemed to have limited relevance, or because authors deemed the focus to be too narrow, the recommendations covered by other papers etc. Thus LLM ETHICS WHITEPAPER does not represent our systematic literature review in its entirety, but rather is primarily intended as a practical resource for conducting ethical research related to LLMs, informed by our expertise as practitioners. Combining a literature review with our own expertise ensures broad coverage whilst maintaining a pragmatic focus.

3 Project Start

The social risk of generative AI including LLMs, can have wide-reaching effects from representational harms to safety concerns, which has been widely recognised (see e.g., Weidinger et al., 2021; Bender et al., 2021; Uzun, 2023; Wei and Zhou, 2022). This recognition has given rise to a large number of efforts seeking to evaluate their risks and actualised harms, each effort presenting its own limitations (Solaiman et al., 2024; Goldfarb-Tarrant et al., 2023; Blodgett et al., 2020). Nevertheless, efforts towards developing technologies that minimise harms, in particular to marginalised communities, are vital. Best efforts require considering a wide range of topics and questions that must be adapted to each individual application and deployment context. In this Section we explain why ethics is relevant to all practitioners (Section 3.1), then highlight resources to aid in the initial discovery process (Section 3.2). We also lay out best practice that will be valuable to all those working with language technologies, namely related to working with stakeholders, and environmental considerations (Section 3.3 and Section 3.4).

3.1 Who needs ethics?

As computer science becomes pervasive in modern lives, so too does it become intertwined with the experience of those lives. Decisions made by researchers and developers compound together to influence every aspect of the technical systems which ultimately govern how we all live (Winner, 1980). This is often at a scale, or level of complexity, which makes it impossible to seek clear resolutions when outcomes are harmful (Van de Poel, 2020; Kasirzadeh, 2021; Miller, 2021; Birhane et al., 2022; Santurkar et al., 2023; Pistilli et al., 2024).

Techno-cultural artefacts, such as LLMs, have political dimensions (Winner, 1980), because they further entrench certain kinds of power e.g. marginalised peoples' data is often used without consent or compensation; technology typically works best for language varieties associated with whiteness (Blodgett and O'Connor, 2017); benchmarks are published which are biased against minorities (Buolamwini and Gebru, 2018). Unfortunately, the training and work cultures of computer scientists can condition us to believe we can ignore power relations (Malazita and Resetar, 2019), because the "objective" nature of our work seems to absolve us of having to consider issues of our technologies in the world (Talat et al., 2021) – when dealing with code and numbers it becomes easier to forget about the real humans who are impacted by our design choices. LLMs are no exception (Leidner and Plachouras, 2017), though their recent rise in prevalence has made their ethical dimensions more salient (and more vital to address).

The design of techno-cultural artefacts like LLMs should be considered interdisciplinary by its very nature, as it requires an understanding of the physical and social systems that they must interact with in order to achieve their function. Experts exist in all of these other areas of study, as well as their intersections, but very often our lack of appreciation for their expertise, or lack of shared language, impede us from seeking them out. This is especially true for expertise in the social sciences and philosophy (Raji et al., 2021; Inie and Derczynski, 2021; Danks, 2022).

There is a tendency to assume that social and ethical issues are not designers' problems but someone else's (Widder and Nafus, 2023), but this is not the case. If you do not reflect on your design decisions as you make them then you are complicit

in the avoidable consequences of those decisions (Talat et al., 2021). The decision to follow a code of ethics (McNamara et al., 2018) or employ a pre-packaged ethical toolkit, does not immediately solve the problem because these decisions require a level of ethical reflection to be effective (Wong et al., 2023).

3.2 Laying the Groundwork

It is important to think about ethics from the very beginning, in order to be able to question all aspects of the project, including if specific tasks should even be undertaken. One way of doing this is by using ethics sheets (Mohammad, 2022), which are sets of questions to ask and answer before starting an AI project. It includes questions like "Why should we automate this task?", and "How can the automated system be abused?". An alternative is using the Assessment List for Trustworthy Artificial Intelligence (ALTAI)[8], which is a tool that helps business and organisations to self-assess the trustworthiness of their AI systems under development. The European Commission appointed a group of experts to provide advice on its artificial intelligence strategy and they translated these requirements into a detailed Assessment List, taking into account feedback from a six month long piloting process within the European AI community. You could use question sets such as these to ensure ethical considerations are present from the start of your project.

Regulated industries such as aerospace, medicine and finance have critical safety issues to address, and a primary way these have been addressed is using auditable processes throughout a project. Audits are tools for interrogating complex processes, to determine whether they comply with company policy and industry standards or regulations (Liu et al., 2012). Raji et al. (2020) introduce a framework for algorithmic auditing that supports artificial intelligence system development, which is intended to contribute to closing the gap between principles and practice. A formal process such as this can help by raising awareness, assigning responsibility, and improving consistency in both procedures and outcomes (Leidner and Plachouras, 2017). At the very least, your organisation should establish an ethics review board to evaluate new products, services, or research plans.

[8] https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=68342

3.3 Stakeholders

Given the vast amounts of training data required and the wide-reaching applications of LLMs, every project will have many stakeholders e.g. those who provide the data (Havens et al., 2020), end-users of the application (Yang et al., 2023), or those a model will be used *on*, who are often given limited power to influence design decisions e.g. migrants (Nalbandian, 2022). A vital early step is identifying key direct stakeholders and establishing the best ways to work with them in order to build systems that are widely beneficial – which will be highly context dependent. The ideation toolkit (Sloane et al., 2022) will help you to identify stakeholders, and can be used alongside existing taxonomies e.g. (Lewis et al., 2020; Langer et al., 2021; Bird et al., 2023; Havens et al., 2020, i.a.). Crucially, stakeholders should be identified before development, so they can (if they wish) be involved in co-production – and object to proposed technologies. Kawakami et al. (2024) present a toolkit for early stage deliberation with stakeholders which includes question prompts, while Caselli et al. (2021) provide 9 guiding principles for effective participatory design – design which involves mutual learning between designer and participant – in the context of NLP research.

You must consider power relations between stakeholders (Havens et al., 2020), and also between yourself and the stakeholders. Reflexive considerations about a researcher's own power are rare in computer science research (Ovalle et al., 2023; Devinney et al., 2022) but can help establish the limitations of your work (Liang et al., 2021; Liang, 2021).

When working on technologies for indigenous and endangered languages, sensitive stakeholder collaboration is particularly important (Bird, 2020; Liu et al., 2022; Mahelona et al., 2023). Work on stakeholder engagement in NLP can learn much from the Indigenous Data Sovereignty movement (Sloane et al., 2022).

3.4 Energy Consumption

Throughout the life cycle of a project, you should consider the energy consumption of your model, which relates to data sourcing practices, model design, choice of hardware, and use at production. Strubell et al. (2019) suggest that model development likely contributes a "substantial proportion of the... emissions attributed to many NLP

researchers". Strubell et al. (2019) call for more research on computationally efficient hardware and algorithms, and the standardised calculation and reporting of finetuning cost-benefit assessments, so researchers can select efficient models (models that are responsive to finetuning) to work with. Similar recommendations are made by Henderson et al. (2020), who also provide a framework for tracking energy, compute and carbon impacts. Patterson et al. (2022) provide best practice for reducing the carbon footprint, including the development of sparse over dense model architectures and the use of cloud computing that relies on renewable energy sources. Bannour et al. (2021) provide a taxonomy of tools available to measure the impact of NLP technologies. Sasha Luccioni and colleagues have in particular championed the accurate reporting of the carbon emissions of ML systems including LLMs (Luccioni et al., 2023; Luccioni and Hernandez-Garcia, 2023; Wang et al., 2023; Luccioni et al., 2024; Dodge et al., 2022; Lacoste et al., 2019).

3.5 Key Resources

Do's and Don'ts

- **Do** engage with affected communities from the beginning – **don't** just ask for their feedback
- **Do** allow for flexibility in project direction as informed by stakeholder input – **don't** assume what communities want and need
- **Do** consider the power relations between stakeholders – **don't** forget about the relationships with yourself
- **Do** engage with ethics review boards to ensure oversight, or set one up if necessary – **don't** assume because it's computer science that moral and political values are out of scope
- **Do** create an internal audit procedure to ensure ethical processes are developed and followed – **don't** just leave it to a post-hoc review
- **Do** consider use of compressed models and cloud resources to minimise energy impact – **don't** assume you need energy intensive models for the best performance

Useful Tool(kit)s:

- Ethics sheets to discover harms and mitigation strategies – Mohammad (2022)
- The Assessment List for Trustworthy Artificial Intelligence (ALTAI)[9]
- Internal audit framework to ensure that ethical processes are implemented and followed – Raji et al. (2020)
- Value Scenarios framework to identify likely impact of technology – Nathan et al. (2007)
- Guiding principles for effective participatory design – Caselli et al. (2021)
- Best practice for reducing carbon footprint during training – Patterson et al. (2022)
- Taxonomy of tools available to measure environmental impact of NLP technologies – Bannour et al. (2021)
- Software package to estimate carbon dioxide required to execute Python codebase – `https://github.com/mlco2/codecarbon`

4 Data compilation

In this Section of LLM ETHICS WHITEPAPER we discuss best practice for compiling original data sets, and critique typical practices such as the position of data as a raw resource rather than something that is transformed by the act of collection. We discuss best practice for addressing issues of consent and safety which includes distinguishing those who produce data and those featured in the data ("data subjects") and respecting their potentially distinct rights. We also discuss best practice for sharing or using shared resources such as thorough documentation and using API such as Elazar et al. (2024) to explore large data sets. For a full discussion of this section and following sections, please see the LLM ETHICS WHITEPAPER.

Key Resources
Do's and Don'ts

- **Do** reflect on and document the decisions you make when collecting data – **don't** forget that *how* you collect data transforms it
- **Do** consider if it is ethical to scrape web content, even for content that is publicly available (e.g., by relying on frameworks of ethical data scraping such as Mancosu and Vegetti (2020)) – **don't** crawl content that website creators have indicated should not be crawled (e.g. via `robots.txt` files)

[9] `https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=68342`

- **Do** consider the subjects of the data – **don't** just think about the rights of data producers
- **Do** respect copyright and privacy from the beginning – **don't** expect the public to do the work of requesting removal (but give them the option!)
- **Do** provide a datasheet for any data set you produce – **don't** forget to document intended use and limitations

Useful Tool(kit)s:

- Case study structure to identify who is missing from collected data – Markl (2022)
- Best practice from Indigenous data sovereignty movement – Walter et al. (2021)
- Checklist for responsible data collection and reuse - Rogers et al. (2021)
- API to explore content of popular massive data sets - Elazar et al. (2024)
- Guidelines to create datasheets - Gebru et al. (2020)

5 Data preparation

In this Section of LLM ETHICS WHITEPAPER we discuss how attempts to clean and filter data can cause harm, even where the intention was to prevent harm! For example, toxicity detection systems, be it word lists or ML models, are typically biased in flagging sentences containing marginalised identity terms as toxic (Bender et al., 2021; Röttger et al., 2021; Calabrese et al., 2021). Each cleaning step should be carefully justified. We also discuss best practice for working with crowd workers, who we recommend be treated as human participants (e.g. research is subject to approval from ethical review board where possible). We offer guidance for those designing target label taxonomies, such as carefully consider what is assumed and what is lost through your choice of proxy (Guerdan et al., 2023). We discuss how to handle disagreement between annotators, which are common in subjective tasks and can reflect ideological differences.

Key Resources
Do's and Don'ts

- **Do** carefully reflect on *whose* data you are excluding when cleaning – **don't** rely on popular tools to give you fair results
- **Do** make explicit what information you are trying to record with your choice of proxy – **don't** forget that labels and proxies are simplifications
- **Do** work with affected communities to define labels and annotate your data – **don't** forget that harm is subjective, and a spectrum
- **Don't** release low quality data that may be repurposed for evaluation
- **Do** gather information about your annotators – **don't** assume annotators with similar identities will give similar annotations
- **Do** treat crowdworkers as human participants and follow best practice for human participant research, such as collecting informed consent; seek formal ethics (e.g. Institutional review board) approval where applicable – **don't** assume that when using paid annotators you do not need to follow typical ethics procedures

Useful Tool(kit)s:

- Recommendations for those conducting data filtering – Hong et al. (2024)
- Taxonomy of personal information and best practice for privacy – Subramani et al. (2023)
- Guidance of selecting proxy labels – Guerdan et al. (2023)
- Best practice when using identity terms as labels – Larson (2017)
- Detailed overview of risks of using crowdworkers – Shmueli et al. (2021)

6 Model Development + Selection

In this Section of LLM ETHICS WHITEPAPER we discuss the ethical ramifications of model design and training, and pre-trained model selection decisions. We echo Hooker (2021) in arguing against the belief that all bias issues stem from data imbalance and explain how subtle model design changes can have big impacts on fairness. We also discuss the limitations of debiasing, as current techniques often make superficial changes (Gonen and Goldberg, 2019), fail to relate to downstream improvements (Steed et al., 2022), and can in fact exacerbate harm (Xu et al., 2021). We also briefly touch on alignment techniques, exploring the difficulty of defining human values (Kasirzadeh and Gabriel, 2023; Kasirzadeh, 2024) and of maintaining alignment throughout a project.

Key Resources
Do's and Don'ts

- **Do** consider how subtle changes can improve performance for marginalised people – **don't** assume that all bias comes from data imbalance
- **Do** use and create model cards for documenting correct and indended uses of models – **don't** assume that a model will be reliable for all populations you might care about
- **Do** test for harm on the deployed model – **don't** test on larger versions before compression as harms can be amplified by this process
- **Do** explore techniques such as finetuning to mitigate harm – but **don't** forget this can introduce new harms, and may not be effective
- **Do** maintain vigilance to ensure alignment is maintained throughout a pipeline – **don't** assume there is only one fixed set of human values

Useful Tool(kit)s:

- Very clear explanation of how model design choices impact fairness – Hooker (2021)
- Templates to document ML models including intended use context – Mitchell et al. (2019)
- Overview of techniques to mitigate LLM harms – Kumar et al. (2022)

7 Evaluation

Here we discuss some of the ethical problems that can arise during performance evaluation, due for example to evaluation not being robust. We offer best practice and cautions for effective performance evaluation. For example, we caution that benchmarks are not objective and can encourage chasing scores which do not relate to real world improvements. We also discuss in detail the benefits and limitations of many harm evaluation strategies. Despite the ubiquitous nature of the harms of LLMs (Rauh et al., 2022; Weidinger et al., 2021), the study of such harms has yet to be standardised. Attempts often lack rigour (Blodgett et al., 2020, 2021; Goldfarb-Tarrant et al., 2023). We briefly present some popular methods for evaluating harms in LLMs, discuss ethical implications and make recommendations.

Key Resources
Do's and Don'ts

- **Do** pair bias metrics that relate to real world (downstream) harms with human evaluation – **don't** rely on quick, quantitative metrics alone, as evaluation in language generation can be unreliable
- **Do** develop and use benchmarks to evaluate concrete, well-scoped and contextualised tasks – **don't** present benchmarks as markers of progress towards general-purpose capabilities
- **Do** carefully reflect on what specific harm you are trying to measure and why the methodology you have created or borrowed is appropriate – **don't** assume bias metrics can be re-used in all new contexts
- **Do** use alternatives to benchmarks which attempt to capture broader capabilities and risks e.g. audits (e.g. Buolamwini and Gebru (2018)), adversarial testing (e.g. Niven and Kao (2019)) and red teaming (Ganguli et al., 2022)
- **Do** involve experts and community members in the evaluation of the models – **don't** rely on your intuitions and initial assumptions alone

Useful Tool(kit)s:

- Tools to facilitate test ideation – Ribeiro et al. (2020)
- Taxonomy of LLM evaluations – Chang et al. (2023) – in particular Section 3.2 on evaluating robustness, ethics, bias, and trustworthiness
- Repository of tests for (English) language generation safety – Dinan et al. (2022)
- Test suite to identifying exaggerated safety behaviour – Röttger et al. (2024)
- Taxonomy of tests for safety and trustworthiness in LLMs – Huang et al. (2023)
- Framework for testing reliability of NLP systems – Tan et al. (2021)
- Bias tests across hundreds of identities (in English) – Smith et al. (2022)
- Framework for addressing Sociotechnical (contextualised) Safety – Weidinger et al. (2023)

8 Deployment

In this Section of LLM ETHICS WHITEPAPER we summarise likely harms of LLMs after deployment. We introduce the notion of dual – both negative and positive – use of LLMs. We discuss the impact of different deployment strategies and the limitations of guardrails. We explain the ramifications of using LLMs to replace humans. Finally we discuss best practice when disseminating your ideas.

Herein we provide a summary of our discussion of risks. In their broad overview of the harms that arise from generative AI, Solaiman et al. (2024) present seven over-arching categories of social harms from technical systems, including representational harms; privacy and data protection; and data and content moderation labour. However, in recognition that these cannot be separated from impacts on society, Solaiman et al. (2024) also present categories of "impacts" on society, such as trustworthiness and autonomy, marginalisation and violence, the concentration of authority, and ecosystem and environmental impacts. Weidinger et al. (2021) and Kumar et al. (2022) have also addressed risks of generative AI. While these sets of authors have focused on generative AI, many of the same concerns—such as bias, stereotypes, and representational harms—have been well documented for other NLP technologies (e.g., Anand et al., 2024; Bolukbasi et al., 2016; Davidson et al., 2019; De-Arteaga et al., 2019).

Key Resources
Do's and Don'ts

- **Do** consider integrating watermarking into your generative models – **don't** rely on supervised detection models alone
- **Do** pre-release audits to identify biases and security vulnerabilities (Madnani et al., 2017) – **don't** put the onus on marginalised people to discover harms
- **Do** release LLMs in stages, with an initial release to trusted researchers, followed by a gradual wider release (Solaiman et al., 2019) – **don't** forget the model will change its own environment in terms of both training data and people's expectations
- **Do** continually monitor post-deployment to assess new risks (Anderljung et al., 2023) – **don't** count on brittle guardrails to prevent harm
- **Do** consider how AI might enhance human experience of work, as well as performance – **don't** assume LLMs can effectively replace human participants
- **Do** consider how the public perceive your technology – **don't** contribute to the hype cycle

Useful Tool(kit)s:

- Framework to encourage AI that enhances rather than replaces human performance – Shneiderman (2020)
- Overview of harms and ramifications of generative AI technologies – Solaiman et al. (2024)
- A definition of dual use, and a checklist for consideration in research projects – Kaffee et al. (2023)
- Documentation methodology for risks of LLMs, that could be adapted to document dual use impact – Derczynski et al. (2023)

9 Limitations and Future Directions

Whilst our Do's and Don'ts are applicable regardless of model language, some of our recommended resources are specific to English. Further, all language-specific resources we discuss in LLM ETHICS WHITEPAPER are specific to English. This reflects a tendency for evaluation resources to be produced only for English, but also the first authors' lack of familiarity with non-English language resources. We extend a similar qualifier in our inclusion of ethical resources that reflect a largely Western moral perspective. Similarly, this paper primarily addresses the text modality, and does not cover other modalities like speech and images. As LLM ETHICS WHITEPAPER is intended as a living document, we can integrate further non-English, non-Western and non-text based resources in future. Moreover, the do's and don'ts are presented in terms of languages for which large amounts of resources already exist. Languages for which few resources exist may need additional consideration in terms of data and data subject safety.

This paper and LLM ETHICS WHITEPAPER do not provide full considerations of the topics covered, but rather serve as syntheses with directions for future reading. Moreover, LLM ETHICS WHITEPAPER and this paper are informed by the literature they rely on, and do not claim to cover all topics of relevance for the development of LLM and LLM-based applications. The Do's and Don'ts we have drafted are not intended to be the final rule in LLM design. Further, it is crucial that readers of this and LLM ETHICS WHITEPAPER situate the considerations of harms of their work within the contexts that their tools will be applied in. LLM ETHICS WHITEPAPER and the Do's and Don'ts can be thought of as starting points, which we will revise in response to community feedback and further consideration for the subject matter of each section. We welcome input from practitioners on

how to make this resource most useful. We are hosting LLM ETHICS WHITEPAPER on Github to expedite this process. We will periodically update the version available on Arxiv to facilitate scholarship.

10 Conclusion

In this paper, we have briefly summarised the topics covered in LLM ETHICS WHITEPAPER, and highlighted topics of particular interest. We synthesise arguments that LLMs and LLM-based applications can have large impacts on society, and therefore developers of such systems need to attend to the types of harms they risk, and seek to mitigate such risks. Here, and in LLM ETHICS WHITEPAPER, we seek to address the gap in resources for conducting ethical research with LLMs that falls between professional association guidelines, and AI frameworks with extremely broad scope, and provide researchers and practitioners with a starting point for their inquiry into ethical research with and development of LLM applications.

Acknowledgements

This work was partly funded by the Generative AI Laboratory (GAIL), University of Edinburgh. Alexandra Birch was partly funded by the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number 10039436 (Utter)]. Eddie L. Ungless was supported by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh, School of Informatics.

References

Gavin Abercrombie, Amanda Cercas Curry, Tanvi Dinkar, Verena Rieser, and Zeerak Talat. 2023. Mirages. on anthropomorphism in dialogue systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4776–4790, Singapore. Association for Computational Linguistics.

Abhishek Anand, Negar Mokhberian, Prathyusha Naresh Kumar, Anweasha Saha, Zihao He, Ashwin Rao, Fred Morstatter, and Kristina Lerman. 2024. Don't Blame the Data, Blame the Model: Understanding Noise and Bias When Learning from Subjective Annotations. *Preprint*, arXiv:2403.04085.

Markus Anderljung, Joslyn Barnhart, Jade Leung, Anton Korinek, Cullen O'Keefe, Jess Whittlestone, Shahar Avin, Miles Brundage, Justin Bullock, Duncan Cass-Beggs, et al. 2023. Frontier ai regulation: Managing emerging risks to public safety. *arXiv preprint arXiv:2307.03718*.

Nesrine Bannour, Sahar Ghannay, Aurélie Névéol, and Anne-Laure Ligozat. 2021. Evaluating the carbon footprint of NLP methods: a survey and analysis of existing tools. In *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*, pages 11–21, Virtual. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, Virtual Event Canada. ACM.

Luciana Benotti, Karën Fort, Min-Yen Kan, and Yulia Tsvetkov. 2023. Understanding Ethics in NLP Authoring and Reviewing. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 19–24, Dubrovnik, Croatia. Association for Computational Linguistics.

Charlotte Bird, Eddie Ungless, and Atoosa Kasirzadeh. 2023. Typology of Risks of Generative Text-to-Image Models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, pages 396–410, New York, NY, USA. Association for Computing Machinery.

Steven Bird. 2020. Decolonising Speech and Language Technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The Values Encoded in Machine Learning Research. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 173–184, Seoul Republic of Korea. ACM.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets.

Su Lin Blodgett and Brendan O'Connor. 2017. Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English. *arXiv:1707.00061 [cs]*. ArXiv: 1707.00061.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91. PMLR. ISSN: 2640-3498.

Agostina Calabrese, Michele Bevilacqua, Björn Ross, Rocco Tripodi, and Roberto Navigli. 2021. AAA: Fair Evaluation for Abuse Detection Systems Wanted. In *13th ACM Web Science Conference 2021*, pages 243–252, Virtual Event United Kingdom. ACM.

Tommaso Caselli, Roberto Cibin, Costanza Conforti, Enrique Encinas, and Maurizio Teli. 2021. Guiding Principles for Participatory Design-inspired Natural Language Processing. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 27–35, Online. Association for Computational Linguistics.

Yu-Chu Chang, Xu Wang, Jindong Wang, Yuanyi Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Weirong Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qian Yang, and Xingxu Xie. 2023. A Survey on Evaluation of Large Language Models. *ArXiv*, abs/2307.03109.

Anthony Cuthbertson. 2023. ChatGPT "grandma exploit" helps people pirate software. Publication Title: The Independent.

David Danks. 2022. Digital ethics as translational ethics. In *Applied ethics in a digital world*, pages 1–15. IGI Global.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128, Atlanta GA USA. ACM.

Leon Derczynski, Hannah Rose Kirk, Vidhisha Balachandran, Sachin Kumar, Yulia Tsvetkov, M. R. Leiser, and Saif Mohammad. 2023. Assessing Language Model Deployment with Risk Cards. Publisher: [object Object] Version Number: 1.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2022. Theories of "Gender" in NLP Bias Research. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 2083–2102, New York, NY, USA. Association for Computing Machinery.

Emily Dinan, Gavin Abercrombie, A. Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. 2022. SafetyKit: First Aid for Measuring Safety in Open-domain Conversational Systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4113–4133, Dublin, Ireland. Association for Computational Linguistics.

Jesse Dodge, Taylor Prewitt, Remi Tachet Des Combes, Erika Odmark, Roy Schwartz, Emma Strubell, Alexandra Sasha Luccioni, Noah A. Smith, Nicole DeCario, and Will Buchanan. 2022. Measuring the Carbon Intensity of AI in Cloud Instances. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1877–1894, Seoul Republic of Korea. ACM.

Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hanna Hajishirzi, Noah A. Smith, and Jesse Dodge. 2024. What's In My Big Data? *arXiv preprint*. ArXiv:2310.20707 [cs].

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, pages 1–79.

Deep Ganguli, Liane Lovitt, John Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Benjamin Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zachary Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom B. Brown, Nicholas Joseph, Sam McCandlish, Christopher Olah, Jared Kaplan, and Jack Clark. 2022. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. *ArXiv*, abs/2209.07858.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2020. Datasheets for Datasets. *arXiv:1803.09010 [cs]*.

Seraphina Goldfarb-Tarrant, Eddie Ungless, Esma Balkir, and Su Lin Blodgett. 2023. This prompt is measuring <mask>: evaluating bias evaluation in language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2209–2225, Toronto, Canada. Association for Computational Linguistics.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove

Them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Luke Guerdan, Amanda Coston, Zhiwei Steven Wu, and Kenneth Holstein. 2023. Ground(less) Truth: A Causal Framework for Proxy Labels in Human-Algorithm Decision-Making. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 688–704, Chicago IL USA. ACM.

Lucy Havens, Melissa Terras, Benjamin Bach, and Beatrice Alex. 2020. Situated Data, Situated Systems: A Methodology to Engage with Power Relations in Natural Language Processing Research. *arXiv:2011.05911 [cs]*. ArXiv: 2011.05911.

Peter Henderson, Jie Hu, Joshua Romoff, E. Brunskill, Dan Jurafsky, and Joelle Pineau. 2020. Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning. *ArXiv*.

Rachel Hong, William Agnew, Tadayoshi Kohno, and Jamie Morgenstern. 2024. Who's in and who's out? A case study of multimodal CLIP-filtering in DataComp. *arXiv preprint*. ArXiv:2405.08209 [cs].

Sara Hooker. 2021. Moving beyond "algorithmic bias is a data problem". *Patterns*, 2(4):100241.

Xiaowei Huang, Wenjie Ruan, Wei Huang, Gao Jin, Yizhen Dong, Changshun Wu, Saddek Bensalem, Ronghui Mu, Yi Qi, Xingyu Zhao, Kaiwen Cai, Yanghao Zhang, Sihao Wu, Peipei Xu, Dengyu Wu, André Freitas, and Mustafa A. Mustafa. 2023. A Survey of Safety and Trustworthiness of Large Language Models through the Lens of Verification and Validation. *ArXiv*, abs/2305.11391.

Nanna Inie and Leon Derczynski. 2021. An IDR Framework of Opportunities and Barriers between HCI and NLP. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 101–108, Online. Association for Computational Linguistics.

Lucie-Aimée Kaffee, Arnav Arora, Zeerak Talat, and Isabelle Augenstein. 2023. Thorny Roses: Investigating the Dual Use Dilemma in Natural Language Processing. Publisher: [object Object] Version Number: 3.

Atoosa Kasirzadeh. 2021. Reasons, values, stakeholders: A philosophical framework for explainable artificial intelligence. *arXiv preprint arXiv:2103.00752*.

Atoosa Kasirzadeh. 2024. Plurality of value pluralism and ai value alignment. In *Pluralistic Alignment Workshop at NeurIPS 2024*.

Atoosa Kasirzadeh and Iason Gabriel. 2023. In conversation with artificial intelligence: aligning language models with human values. *Philosophy & Technology*, 36(2):27.

Anna Kawakami, Amanda Coston, Haiyi Zhu, Hoda Heidari, and Kenneth Holstein. 2024. The Situate AI Guidebook: Co-Designing a Toolkit to Support Multi-Stakeholder Early-stage Deliberations Around Public Sector AI Proposals. ArXiv:2402.18774 [cs].

Sachin Kumar, Vidhisha Balachandran, Lucille Njoo, Antonios Anastasopoulos, and Yulia Tsvetkov. 2022. Language Generation Models Can Cause Harm: So What Can We Do About It? An Actionable Survey. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, 296:103473.

Brian Larson. 2017. Gender as a Variable in Natural-Language Processing: Ethical Considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain. Association for Computational Linguistics.

Jochen L. Leidner and Vassilis Plachouras. 2017. Ethical by Design: Ethics Best Practices for Natural Language Processing. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 30–40, Valencia, Spain. Association for Computational Linguistics.

Dave Lewis, Linda Hogan, David Filip, and P. J. Wall. 2020. Global Challenges in the Standardization of Ethics for Trustworthy AI. *Journal of ICT Standardization*.

Calvin Liang. 2021. Reflexivity, positionality, and disclosure in HCI.

Calvin A. Liang, Sean A. Munson, and Julie A. Kientz. 2021. Embracing Four Tensions in Human-Computer Interaction Research with Marginalized People. *ACM Transactions on Computer-Human Interaction*, 28(2):1–47.

Jie Liu et al. 2012. The enterprise risk management and the risk oriented internal audit. *Ibusiness*, 4(03):287.

Zoey Liu, Crystal Richardson, Richard Hatcher, and Emily Prud'hommeaux. 2022. Not always about you: Prioritizing community needs when developing endangered language technology. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3933–3944, Dublin, Ireland. Association for Computational Linguistics.

Nicola Lucchi. 2023. Chatgpt: A case study on copyright challenges for generative artificial intelligence systems. *European Journal of Risk Regulation*, page 1–23.

Alexandra Sasha Luccioni and Alex Hernandez-Garcia. 2023. Counting Carbon: A Survey of Factors Influencing the Emissions of Machine Learning. *arXiv preprint*. ArXiv:2302.08476 [cs].

Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. 2023. Estimating the carbon footprint of bloom, a 176b parameter language model. *Journal of Machine Learning Research*, 24(253):1–15.

Sasha Luccioni, Yacine Jernite, and Emma Strubell. 2024. Power Hungry Processing: Watts Driving the Cost of AI Deployment? In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 85–99, Rio de Janeiro Brazil. ACM.

Nitin Madnani, Anastassia Loukina, Alina von Davier, Jill Burstein, and Aoife Cahill. 2017. Building Better Open-Source Tools to Support Fairness in Automated Scoring. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 41–52, Valencia, Spain. Association for Computational Linguistics.

Keoni Mahelona, Gianna Leoni, Suzanne Duncan, and Miles Thompson. 2023. OpenAI's whisper is another case study in colonisation. *Papa Reo*.

James W Malazita and Korryn Resetar. 2019. Infrastructures of abstraction: how computer science education produces anti-political subjects. *Digital Creativity*, 30(4):300–312.

Moreno Mancosu and Federico Vegetti. 2020. What You Can Scrape and What Is Right to Scrape: A Proposal for a Tool to Collect Public Facebook Data. *Social Media + Society*, 6(3):2056305120940703. Publisher: SAGE Publications Ltd.

Nina Markl. 2022. Mind the data gap(s): Investigating power in speech and language datasets. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 1–12, Dublin, Ireland. Association for Computational Linguistics.

Andrew McNamara, Justin Smith, and Emerson Murphy-Hill. 2018. Does acm's code of ethics change ethical decision making in software development? In *Proceedings of the 2018 26th ACM joint meeting on european software engineering conference and symposium on the foundations of software engineering*, pages 729–733.

Boaz Miller. 2021. Is technology value-neutral? *Science, Technology, & Human Values*, 46(1):53–80.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent Advances in Natural Language Processing via Large Pre-trained Language Models: A Survey. *ACM Comput. Surv.*, 56(2):30:1–30:40.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229. ArXiv: 1810.03993.

Saif Mohammad. 2022. Ethics Sheets for AI Tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8368–8379, Dublin, Ireland. Association for Computational Linguistics.

Lucia Nalbandian. 2022. An eye for an 'I:' a critical assessment of artificial intelligence tools in migration and asylum management. *Comparative Migration Studies*, 10(1):32.

Lisa P. Nathan, Predrag V. Klasnja, and Batya Friedman. 2007. Value scenarios: a technique for envisioning systemic effects of new technologies. In *CHI '07 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '07, pages 2585–2590, New York, NY, USA. Association for Computing Machinery.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

Anaelia Ovalle, Arjun Subramonian, Vagrant Gautam, Gilbert Gee, and Kai-Wei Chang. 2023. Factoring the Matrix of Domination: A Critical Review and Reimagination of Intersectionality in AI Fairness. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 496–511, Montr\'{e}al QC Canada. ACM.

David Patterson, Joseph Gonzalez, Urs Hölzle, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David R. So, Maud Texier, and Jeff Dean. 2022. The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink. *Computer*, 55(7):18–28. Conference Name: Computer.

Giada Pistilli, Alina Leidinger, Yacine Jernite, Atoosa Kasirzadeh, Alexandra Sasha Luccioni, and Margaret Mitchell. 2024. Civics: Building a dataset for examining culturally-informed values in large language models. *arXiv preprint arXiv:2405.13974*.

Inioluwa Deborah Raji, Morgan Klaus Scheuerman, and Razvan Amironesei. 2021. You can't sit with us: Exclusionary pedagogy in ai ethics education. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 515–525.

Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and

Parker Barnes. 2020. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. page 12.

Maribeth Rauh, John Mellor, Jonathan Uesato, Po-Sen Huang, Johannes Welbl, Laura Weidinger, Sumanth Dathathri, Amelia Glaese, Geoffrey Irving, Iason Gabriel, William Isaac, and Lisa Anne Hendricks. 2022. Characteristics of Harmful Text: Towards Rigorous Benchmarking of Language Models. Publisher: [object Object] Version Number: 2.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Anna Rogers, Timothy Baldwin, and Kobi Leins. 2021. 'Just What do You Think You're Doing, Dave?' A Checklist for Responsible Data Use in NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4821–4833, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5377–5400, Mexico City, Mexico. Association for Computational Linguistics.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Talat, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional Tests for Hate Speech Detection Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR.

Boaz Shmueli, Jan Fell, Soumya Ray, and Lun-Wei Ku. 2021. Beyond Fair Pay: Ethical Implications of NLP Crowdsourcing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3758–3769, Online. Association for Computational Linguistics.

Ben Shneiderman. 2020. Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human–Computer Interaction*, 36(6):495–504. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10447318.2020.1741118.

Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. 2022. Participation Is not a Design Fix for Machine Learning. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–6, Arlington VA USA. ACM.

Eric Michael Smith, Melissa Hall Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. "I'm sorry to hear that": finding bias in language models with a holistic descriptor dataset. Technical Report arXiv:2205.09209, arXiv. ArXiv:2205.09209 [cs] version: 1 type: article.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, and Jasmine Wang. 2019. Release Strategies and the Social Impacts of Language Models. Technical Report arXiv:1908.09203, arXiv. ArXiv:1908.09203 [cs] type: article.

Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Canyu Chen, Hal Daumé III, Jesse Dodge, Isabella Duan, Ellie Evans, Felix Friedrich, Avijit Ghosh, Usman Gohar, Sara Hooker, Yacine Jernite, Ria Kalluri, Alberto Lusoli, Alina Leidinger, Michelle Lin, Xiuzhu Lin, Sasha Luccioni, Jennifer Mickel, Margaret Mitchell, Jessica Newman, Anaelia Ovalle, Marie-Therese Png, Shubham Singh, Andrew Strait, Lukas Struppek, and Arjun Subramonian. 2024. Evaluating the Social Impact of Generative AI Systems in Systems and Society. *Preprint*, arXiv:2306.05949.

Ryan Steed, Swetasudha Panda, Ari Kobren, and Michael L. Wick. 2022. Upstream Mitigation Is Not All You Need: Testing the Bias Transfer Hypothesis in Pre-Trained Language Models. In *ACL*.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650. Conference Name: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics Place: Florence, Italy Publisher: Association for Computational Linguistics.

Nishant Subramani, Sasha Luccioni, Jesse Dodge, and Margaret Mitchell. 2023. Detecting Personal Information in Training Corpora: an Analysis. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 208–220, Toronto, Canada. Association for Computational Linguistics.

Zeerak Talat, Smarika Lulz, Joachim Bingel, and Isabelle Augenstein. 2021. Disembodied Machine Learning: On the Illusion of Objectivity in NLP. *arXiv:2101.11974 [cs]*. ArXiv: 2101.11974.

Samson Tan, Shafiq Joty, Kathy Baxter, Araz Taeihagh, Gregory A. Bennett, and Min-Yen Kan. 2021. Reliability Testing for Natural Language Processing Systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4153–4169, Online. Association for Computational Linguistics.

Eddie L. Ungless, Nikolas Vitsakis, Zeerak Talat, James Garforth, Björn Ross, Arno Onken, Atoosa Kasirzadeh, and Alexandra Birch. 2024. Ethics Whitepaper: Whitepaper on Ethical Research into Large Language Models. *arXiv preprint*. ArXiv:2410.19812 [cs].

Levent Uzun. 2023. Are Concerns Related to Artificial Intelligence Development and Use Really Necessary: A Philosophical Discussion. *Digital Society*, 2(3):40.

Ibo Van de Poel. 2020. Embedding values in artificial intelligence (ai) systems. *Minds and machines*, 30(3):385–409.

Maggie Walter, Raymond Lovett, Bobby Maher, Bhiamie Williamson, Jacob Prehn, Gawaian Bodkin-Andrews, and Vanessa Lee. 2021. Indigenous Data Sovereignty in the Era of Big Data and Open Data. *Australian Journal of Social Issues*, 56(2):143–156. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ajs4.141.

Xiaorong Wang, Clara Na, Emma Strubell, Sorelle Friedler, and Sasha Luccioni. 2023. Energy and Carbon Considerations of Fine-Tuning BERT. *arXiv preprint*. ArXiv:2311.10267 [cs].

Mengyi Wei and Zhixuan Zhou. 2022. AI Ethics Issues in Real World: Evidence from AI Incident Database. Publisher: [object Object] Version Number: 2.

Laura Weidinger, John F. J. Mellor, M. Rauh, C. Griffin, J. Uesato, Po-Sen Huang, M. Cheng, Mia Glaese, B. Balle, A. Kasirzadeh, Z. Kenton, S. Brown, W. Hawkins, T. Stepleton, C. Biles, A. Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William S. Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from Language Models. *undefined*.

Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, et al. 2023. Sociotechnical safety evaluation of generative ai systems. *arXiv preprint arXiv:2310.11986*.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 214–229.

David Gray Widder and Dawn Nafus. 2023. Dislocated accountabilities in the "ai supply chain": Modularity and developers' notions of responsibility. *Big Data & Society*, 10(1):20539517231177620.

Langdon Winner. 1980. Do Artifacts Have Politics? *Daedalus*, 109(1):121–136. Publisher: The MIT Press.

Richmond Y. Wong, Michael A. Madaio, and Nick Merrill. 2023. Seeing Like a Toolkit: How Toolkits Envision the Work of AI Ethics. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–27.

Albert Xu, Eshaan Pathak, Eric Wallace, Suchin Gururangan, Maarten Sap, and Dan Klein. 2021. Detoxifying Language Models Risks Marginalizing Minority Voices. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2390–2397, Online. Association for Computational Linguistics.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.

Jiancheng Yang, Hongwei Bran Li, and Donglai Wei. 2023. The impact of chatgpt and llms on medical imaging stakeholders: perspectives and use cases. *Meta-Radiology*, page 100007.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, page 100211.