Provisional Draft of the NeurIPS Code of Ethics

Samy Bengio **Alina Beygelzimer** **Kate Crawford** **Jeanne Fromer** **Iason Gabriel**

Amanda Levendowski **Deborah Raji** **Marc'Aurelio Ranzato**

Abstract

Over the past few decades, research in machine learning and AI has had a tremendous impact in our society. The number of deployed applications has greatly increased, particularly in recent years. As a result, the NeurIPS Conference has received an increased number of submissions (approaching 10,000 in the past two years), with several papers describing research that has foreseeable deployment scenarios. With such great opportunity to impact the life of people comes also a great responsibility to ensure research has an overall positive effect in our society.

Before 2020, the NeurIPS program chairs had the arduous task of assessing papers not only for their scientific merit, but also in terms of their ethical implications. As the number of submissions increased and as it became clear that ethical considerations were also becoming more complex, such *ad hoc* process had to be redesigned in order to properly support our community of submitting authors.

The program chairs of NeurIPS 2020 established an ethics review process, which was chaired by Iason Gabriel. As part of that process, papers flagged by technical reviewers could undergo an additional review process handled by reviewers with expertise at the intersection of AI and Ethics. These ethical reviewers based their assessment on guidelines that Iason Gabriel, in collaboration with the program chairs, drafted for the occasion.

This pilot experiment was overall successful because it surfaced early on several papers that needed additional discussion and provided authors with additional feedback on how to improve their work. Extended and improved versions of such ethics review process were later adopted by the NeurIPS 2021 program chairs as well as by other leading machine learning conferences.

One outstanding issue with such a process was the lack of transparency in the process and lack of guidelines for the authors. Early in 2021, the NeurIPS Board gave Marc'Aurelio Ranzato the mandate to form a committee to draft a code of Ethics to remedy this. The committee, which corresponds to the author list of this document, includes individuals with diverse background and views who have been working or have been involved in Ethics in AI as part of their research or professional service. The first outcome of the committee was the Ethics guidelines, which was published in May 2021.

The committee has worked for over a year to draft a Code of Ethics. This document is their current draft, which has not been approved by the NeurIPS Board as of yet. The Board decided to first engage with the community to gather feedback. We therefore invite reviews and comments on this document. We welcome your encouragement as well as your critical feedback. We will then revise this document accordingly and finalize the draft, hoping that this will become a useful resource for submitting authors, reviewers and presenters.

Preprint. Under review.

1 Preamble

The Code of Ethics aims to guide the NeurIPS community towards higher standards of ethical conduct. It outlines conference expectations about the broad ethical practices that must be adopted by participants and by the wider community, including submitting authors, members of the program committee and speakers. The Code of Ethics complements the NeurIPS Code of Conduct, which focuses on professional conduct and research integrity, by considering also elements of research ethics and the broader societal and environmental impact of research in our field. Notice that conference contributors are also obliged to adhere to additional ethical codes or review requirements arising from other stakeholders like funders or research institutions.

The NeurIPS Code of Ethics is a living document that will be updated regularly to reflect our community understanding and evolving sensitivity towards matters related to ethics. Ethical considerations are often complex and nuanced. Therefore, this document aims at supporting the comprehensive evaluation of ethics in AI research by offering material of reflection and further reading, as opposed to providing a strict set of rules.

In general, what is legally permissible might not be considered ethically acceptable in AI research. There are two guiding principles that underlie the NeurIPS Code of Ethics. First, AI research is worthwhile only to the extent that it can have a positive net outcome for our society. Second, since the assessment of whether a research practice or outcome is ethical might be difficult to establish unequivocally, transparent and truthful communication is of paramount importance.

The rest of the document is organized in two parts. The first part (§2) pertains to research ethics, i.e., ethical considerations that arise while researchers pursue and develop their own research artifact. In particular, section 2.1 deals with how human subjects might have been treated (e.g., were they harmed during the process? Was their wage fair?). This is followed by a discussion on potential environmental damages due to research activities (§2.2). Section 2.3 is dedicated to how data has been used in the development of the research work. This section touches on several aspects, including privacy (§2.3.1), documentation (§2.3.2), dataset deprecation (§2.3.3), application of copyright and fair use (§2.3.4), and evaluation (§2.3.5).

The second part of the Code of Ethics (§3) focuses on the societal impact of the research, that is, on reflections about the likely impact that a piece of research (once it is completed and deployed) can have in our society and environment. For instance, section 3.1 encourages researchers to engage early on with those who are likely going to be affected by the deployment, particularly the most vulnerable ones, as well as with domain experts to better understand the effects of a piece of research. Section 3.2 emphasizes the importance of candid communication about potential safety (§3.2.1) and security (§3.2.2) issues deriving from the research work, potential labor issues (§3.2.3), increase in discrimination against subgroups of individuals (§3.2.4), and issues related to privacy (§3.2.5), deception (§3.2.6), environmental damage (§3.2.7), violation of human rights (§3.2.8), and use of weapons (§3.2.9). The section concludes with a recommendation to communicate mitigation strategies (§3.3) to remedy any potential negative impact, and with a list of pointers (§3.3.1) for further reading and reflection on the topic.

2 Research Ethics

This section considers potential ethical issues related to all research activities and design decisions. This is not about deployment scenarios or the consequences of applications of the research but the ethical challenges that arise from the execution and reporting of the research process itself. For example, experimental design, data collection practices, research communication and more. Though we acknowledge this to be a component of research ethics, we will specifically not address research integrity issues in this section, as these are dealt separately in Code of Conduct.

2.1 Research involving human subjects or participants

Many research projects involve human research participants and human subjects. This includes research assistants, crowdsourced annotators, and those from whom data is passively collected or actively solicited.

2.1.1 Inclusive Stakeholder Participation

Consider the influence of research activities and processes on relevant stakeholders, especially those that are the most vulnerable or who are likely to be negatively impacted by the research.

In the context of applied research in particular, it is important to engage with relevant stakeholders to whatever extent is possible to better understand their viewpoint as it relates to the project, and communicate about the outcome of such discussions when they happen. This could include focus groups or surveys with an impacted community to consult on research projects that impact their community.

2.1.2 Ensure research topics and methods minimise direct harm to living beings

Reflect on potential consequences of research activities throughout the development and application process. Be aware that when multiple stakeholders, in this case people involved in the development of the research artifact, have competing interests, we must minimize harm, and especially prioritize the concerns of those who are the most vulnerable or disadvantaged.

2.1.3 Fair Wages

If you make use of crowdsourcing or contract work for a particular task as part of your research project, please communicate transparently about the wages offered to participants for their labor. If one chooses to categorize people as part of their research process (ie. data creation, task design, etc.), please reflect on the ethical implications of doing so.

2.1.4 There should be an internal review process for research projects involving human subjects

A project involves human subjects when making use of data collected of human subjects or involving the extensive contribution of human participants. For example, any new dataset making use of human information must provide a detailed explanation of why it was necessary to do so, including if there was an assessment by an institutional review board (IRB).

When a project involves human subjects, there must be some internal review process, during which researchers reflect on ethical considerations prior to the design and implementation of the research project.

Please go through an Institutional Review Board (IRB) process, internal ethics review process imposed by institutions and funders, or any equivalent process, or informal review when possible. Authors are encouraged to communicate transparently about the decision to enter this process or not, as well as report the details of the process, status and outcomes in their submission.

2.2 Environment

Ensure research topics, activities and methods minimise direct harm to the environment.

For example, limit computational resources, e.g. energy consumption, when developing research artifacts.

2.3 Data-related concerns

There are challenges inherent to papers presenting new datasets, papers presenting experiments on new or existing datasets, and papers using identity characteristics (e.g., gender, race, ethnicity) as variables in data. The following are a set of issues and expectations related to the use of data in machine learning research that authors should be actively aware of.

2.3.1 Privacy and Consent Considerations

Researchers should be attentive to the privacy protections of data subjects and research participants. Datasets should not contain any personally identifiable information, without any measures taken to protect the identity of the involved parties, unless some form of informed consent from those

individuals is provided to do so. Even when a dataset has been approved by an IRB, subsequent users should conduct their own independent assessment to address potential privacy and consent issues.

2.3.2 Data and model documentation should accompany research artifacts

Researchers should communicate about the details of the dataset or the model as part of their submissions. Data documentation requirements should adhere to templates (for example, Model Cards & Datasheets) including a need to communicate about data provenance (ie. context in which the data was sourced, dataset collection process and conditions) and data curation processing (i.e. what measures were taken to address bias).

2.3.3 Deprecated datasets

Researchers should not use datasets that have been deprecated for technical, legal, or ethical reasons. NeurIPS will maintain a list of datasets that are known to have been deprecated, and papers will not be accepted that use these datasets. Dataset makers should complete a deprecation sheet with NeurIPS when they no longer want a dataset they have made to be used, and NeurIPS will share that information with the wider community. For more details on dataset deprecation, see section A in the Appendix.

2.3.4 Copyright and Fair Use

There are different expectations for copyright and consent depending on the geographical region and context. The use of the data should be consistent with the terms for the data source or as directed through any contractual agreement. This includes the terms of licenses for distribution of datasets, and the preferences of data creators regarding the terms of use for their datasets. These terms should only be violated in situations involving whistleblowing, in which context those circumstances should be explicitly acknowledged. Researchers with questions about copyright law or fair use can choose to consult with an attorney or contact a university-based technology law clinic.

For data that is already publicly available and without any copyright claim, researchers should be considerate of the subjects in the data and their privacy concerns. Copyright concerns related to curated datasets that are already made available publicly should be addressed by the original authors. If an authorized dataset has already been leaked, then researchers are highly discouraged from using such datasets .

Using copyright law to determine which datasets to use can be complicated. The norms of machine learning research are still evolving, and use of a dataset may be permitted under fair use but raise ethical issues. Researchers should strive to have consent to use a dataset and must provide a detailed explanation to reviewers of why consent was not or could not be obtained, including if it is for the purposes of audit or a critical examination of the dataset, particularly of a deprecated dataset (see §2.3.3). The distinction between what is legally permissible and what is ethically acceptable is particularly relevant when researchers intend to use datasets featuring real data of real people, including photographs, videos, voice recordings, and text-based messages, without their consent. Researchers are expected to take a subject-based approach to consent, rather than an approach centering consent from a copyright owner, which requires both (1) awareness of the subject that the data was being recorded or documented, and (2) permission from the subject to use the data in the dataset. A dataset containing data that fail to meet this standard is understood to be using the data without consent. Any paper that chooses to use a dataset with real data of real people without consent must provide a detailed explanation to reviewers of why it was necessary to do so, including if it is for the purposes of audit or a critical examination of the dataset.

Please refer to Section 1.5 and 1.7 of the ACM code of ethics for further information.

2.3.5 Engaging in representative data and evaluation practice

Datasets should aim to be representative of a diverse population when collecting new datasets or making decisions about which datasets to use. In instances when the data collection or curation process cannot be inclusive, authors should communicate this limitation clearly.

Model performance should be reported on disaggregated test sets on the relevant sub-populations when possible.Guidance on how to begin thinking about this can be found in related work on the topic. This should be communicated to contextualize broader model performance.

Any suspected biases or limits to the scope of performance of models should be clearly communicated. This also includes a necessary inspection or warning if the dataset encodes, contains or might exacerbate bias against people of a certain gender, race, sexuality, or who have other protected characteristics.

3 Societal Impact

All contributors are expected to reflect upon the wider social impact of their research, discuss potential domains of application, and identify ways to mitigate risks of harm when these arise. Unless otherwise specified in the yearly Call for Papers, these reflections must be placed in a section dedicated to societal impact. Members of the reviewing committee are expected to assess the value of a research contribution in terms of the insight and perspectives offered in this discussion. The closer the work is to downstream application, the more central these considerations will be to reviewer evaluation.

3.1 Engage with stakeholders and domain experts to identify potential application concerns

Contributors are asked to identify those who are likely to be impacted by the publication or deployment of research – particularly if they are from vulnerable or disadvantaged groups. Researchers are encouraged to proactively engage with these stakeholders in order to learn from their perspectives. When this is not feasible, researchers should engage with relevant literature or documentation that explores the impact of their research on vulnerable populations or marginalized groups. Researchers should then use information acquired in either way to inform their discussion of the risk and to propose mitigations.

3.2 Reflect and transparently communicate the known or anticipated consequences of research

Researchers are expected to communicate the known or anticipated consequences of their research in good faith, to explore the real-world implications of their research, and to evaluate the likelihood that it will lead to beneficial or harmful outcomes. Researchers should be particularly mindful of risk in the following areas:

3.2.1 Safety Considerations

Safety is concerned with protection from harm where this extends to include prospective negative impact on a person's physical, emotional or psychological well-being. Contributors should consider whether there are situations in which their technology could be used to harm, injure or kill other people through its direct application, side effects, or potential misuse.

Please, refer to Section 1.2 of the ACM code of ethics for further information.

3.2.2 Security Considerations

Security refers to the protection from vulnerabilities that can be exploited by malicious actors. Researchers should endeavour to protect data and systems from access or interference by malicious actors. Researchers are also expected to engage in the responsible communication of identified vulnerabilities or weaknesses to adversarial attack.

Researchers should consider whether, for example, there is a risk that applications could open security vulnerabilities or cause serious accidents when deployed in real world environments and recommend ways to mitigate any potential harms.

3.2.3 Labor Issues

Machine learning research intersects with the labor market in a variety of ways. People provide labor to build these systems and future employment will sometimes be impacted through deployment.

Taken together these harms potentially include, but are not limited to, disruptions to livelihood, automation of tasks, and worsening worker conditions.

Researchers should consider whether, for example, the research could result in a threat to certain sections of the job market? Could it be applied to automate certain sections of the economy at scale in ways that could result in unemployment?

3.2.4 Discrimination

Researchers should take steps to prevent their work from advancing, perpetuating or entrenching discrimination through dataset bias, model bias, or through incorporation into harmful applications that lead to discriminatory outcomes.

Researchers should also consider whether, for example, the technology be used to discriminate, exclude, or otherwise negatively impact people, including impacts on the provision of vital services, such as healthcare, education or access to credit.

Please consult the Toronto Declaration for further details.

3.2.5 Privacy Violations & Surveillance

Certain forms of surveillance violate reasonable expectations around privacy and consent causing harm to individuals and communities. This has been a particular problem for marginalized groups.

Researchers should consider whether, for example, the research could be used to collect or analyze bulk surveillance data to predict immigration status or other protected categories, or be used in any kind of criminal profiling? More generally, researchers should seek to minimize the collection of personal data, and in particular, personal information without explicit consent.

3.2.6 Deception & Harassment

Research should not be applicable for the purposes of deception or the harassment of people. This includes research that further the willful misrepresentation of facts.

Researchers should consider, for example, whether their approach could be used to facilitate deceptive interactions that would cause harm such as theft, fraud, or harassment? Could it be used to impersonate public figures and influence political processes, or as a tool that promotes hate speech or abuse?

3.2.7 Environment

Contributors are expected to consider the environmental consequences, if any, of their work. For instance, a description of the computational resources (e.g., energy consumed, number of devices used in a certain amount of time) needed to train a model or validate the research is considered best practice for empirical works. Discussion of other negative impacts that research may have on the environment should also be included.

Researchers should consider, for example, whether their research is going to negatively impact the environment by promoting fossil fuel extraction, deforestation, or by increasing pollution.

3.2.8 Legality & Human Rights

We hold the members of the NeurIPS community to higher standards than what is required by the law. We prohibit circulation of any research work that builds upon or facilitates illegal activity. Moreover, we do not allow work that aims at breaching human rights, such as technology that could be used to deny people rights to privacy, speech, health, liberty, security, legal personhood, or freedom of conscience or religion.

3.2.9 Weapons

Research should not be designed for incorporation into weapons or weapons systems, either as a direct aim or primary outcome. We encourage researchers to consider ways in which their contribution could be used to develop, support, or expand weapons systems and to take measures to forestall this outcome.

For example, is the research likely to be used in the production of lethal autonomous weapons or the targeting of these weapons systems? (UN LAWS)

3.3 Communicate risk mitigation

Whenever potential negative consequences have been identified, we expect authors and presenters to elaborate on what tools, approaches and strategies could be used to mitigate the risk. The goal is to make a convincing argument that once such mitigations have been taken into account, the potential benefits of the research work are likely to outweigh the potential negative consequences. That calculus can only be made if authors have transparently and fully accounted for all the potential risks and mitigations.

3.3.1 Proposed Mitigation Strategies

Available risk mitigation strategies to address potential negative downstream impacts include, but are not limited to:

- Selection of a responsible release and publication strategy to allow for controlled use of the model, when risk for misuse or harm is high. See for instance [1]
- Documentation and audits to declare intended use and allow for external scrutiny of these systems. See for instance [here]
- Privacy protocols, encryption, or anonymization, as measures to mitigate data leakage and security threats after model release. See for instance [here].
- Modifications to prevent misuse or inappropriate use [here]
- Data curation or distribution to mitigate downstream discriminatory impacts.See for instance [here]
- Liaise with civil society organizations and affected communities to anticipate potential under-considered harms and allow for external oversight
- Consult with law school clinics specializing in intellectual property and technology issues, to prevent legal liability. These clinics work with student attorneys to advise clients for free. A list of many relevant clinics as of 2019-2020 is available here (pages 150-53).

Competing Interest Disclosure

The committee was formed in February 2021. At the time Samy Bengio was affiliated to Google Research and is now affiliated to Apple. Alina Beygelzimer is affiliated to Yahoo! Research. Kate Crawford is a Research Professor at the University of Southern California Annenberg, a Senior Principal Researcher at Microsoft Research, and an Honorary Professor at the University of Sydney. Jeanne Fromer is a Professor of Law at New York University. Iason Gabriel is affiliated to both DeepMind and University of California at Berkeley. Amanda Levendowski is an Associate Professor of Law at Georgetown Law. Deborah Raji is a Mozilla fellow and affiliated to University of California at Berkeley. Marc'Aurelio Ranzato is currently affiliated to DeepMind, but prior to June 2021 was affiliated to Facebook AI Research.

This document was drafted under the NeurIPS Board's mandate for the best interests of the NeurIPS community.

Appendix A A Framework for Addressing Deprecated Datasets in the NeurIPS Community

A proposal based on research conducted by Sasha Luccioni, Francis Corry, Hamsini Sridharan, Mike Ananny, Jason Schultz, and Kate Crawford from the Knowing Machines Project

A.1 Summary

Datasets are central to machine learning. The NeurIPS community has been making improvements to data stewardship and documentation practices across the model development life cycle, and now has a Datasets and Benchmarks track. However, there is no centralized way to register datasets, check if they are still valid for paper submission, or flag or remove them if they are found to have technical, legal, or ethical issues associated with their use. For example in 2021, the NeurIPS chairs consulted with members of the community to crowdsource which datasets should no longer be used, but this is a time consuming and incomplete process. After conducting a study of dataset deprecation with a group of research scientist colleagues in industry and academia, we propose that NeurIPS is in a strong position to lead the field by hosting a centralized, sustainable dataset repository, which could be used for archiving datasets, tracking dataset modifications, deletions, or derivations, and facilitating practices of good data stewardship that can be integrated into research and conference publication processes.

A.2 Background context

In recent years, ML scholarship has unearthed many issues with commonly-used datasets, including technical errors, discriminatory and offensive labels, and privacy violations. This has brought attention to the importance of careful data stewardship practices in all steps of the dataset life cycle, ranging from proper documentation at the data creation stage via checklists and data sheets to the study of dataset usage and genealogy, licensing and auditing, as well as data stewardship and maintenance.

However, less attention has been given to dataset deprecation, such as when and how datasets should be removed or modified. Datasets can be deprecated for many reasons, ranging from issues around legality (e.g., datasets with images gathered that breach privacy laws) and ethics (e.g., datasets that perpetuate harmful stereotypes or biases), but also more mundane reasons like the creation of a new derivative dataset. In all deprecation cases, to ensure datasets' integrity and legitimacy, the NeurIPS community needs mechanisms to ensure that information about deprecations are easy to find, well-documented, clearly communicated, and follow a timeline with due process appeals built in.

Without these mechanisms, ad-hoc deprecation practices can and have become the norm, resulting in datasets that continue to circulate after deprecation, and a community that does not know which datasets to use. When deprecated datasets are cited in NeurIPS papers or used for training models, the problems can compound. In some cases, multiple versions and derivatives of datasets live on after deprecation and continue to do harm, despite their creators' intentions to remove them. Researchers continue to use these sites not out of carelessness but because it's extremely difficult to know which datasets have been deprecated and why. The use of deprecated datasets is further sustained by the lack of documentation about when, why, and how a dataset was removed. Poor documentation practices about dataset deprecation, along with their continued circulation and use within ML communities, can perpetuate the harms that deprecations often seek to address.

NeurIPS is powerfully positioned to support dataset creators and the wider ML community with a framework for good dataset stewardship, which will also help researchers and practitioners to avoid technical, legal and ethical issues when using datasets that have been, or should be, deprecated. NeurIPS can address these issues by instituting a Dataset Deprecation Protocol. It would include three elements:

- A centralized repository for managing deprecation information
- A deprecation sheet template for dataset creators to use
- A system of permanent identifiers for datasets (i.e like DOI, digital object identifiers)

Dataset Deprecation Protocol would help researchers and practitioners track the status of datasets and reasons for their deprecation, and over time, would support a culture of good stewardship for datasets over their lifespan. While permanent identifiers for datasets are optional, they can provide stronger infrastructural stability in the long term.

A.3 Examples of deprecated datasets that still circulate

Several datasets have recently been deprecated or edited to create derivatives after work by researchers that revealed underlying problems. They include: Tiny Images, MS-Celeb-1M, ImageNet, Duke MTMC, MegaFace, and HRT Transgender. In our study of the reasons for these datasets to be changed or deleted, we found that despite the removal of these datasets from their original hosting location, some still continue to circulate, are used to train models, and are cited in ML papers. Sometimes this research is published years after the deprecation. For example, MS-Celeb-1M's harms and deprecated status were well-documented in popular press accounts when it was removed in April 2019. Yet, as Peng et al. have noted, the underlying data for MS-Celeb-1M were used hundreds of times in published papers since its 2019 retraction. Today, it continues to circulate on sites like Academic Torrents. Duke MTMC and Tiny Images have also been deprecated but are still widely circulated, used, and cited. Popular repositories like Papers With Code and Exposing.ai (by Adam Harvey and Jules LaPlace) are useful ways to track datasets that have already been investigated and deprecated, but there remains a need for a centralized resource that aggregates the deprecations of datasets that the ML community has used, regardless of the dataset's prominence or role in a public controversy.

To understand which datasets they should use – and why – researchers need to be able to see the status of active, updated, and deprecated datasets in one place. Such a directory would help researchers make informed choices about model training and could, over time, show the NeurIPS community how and why it deprecates datasets and become a valuable tool for reflecting on the field's values and practices.

A.4 Reasons for Dataset Deprecation

A.4.1 Technical Considerations

There are several technical reasons why a dataset is deprecated, which can result in downstream issues in model relevance and performance. For instance, some datasets suffer from contamination – i.e. the presence of the same data in both the training and testing of a given dataset. This is increasingly problematic for datasets scraped from the Internet, which may contain millions of documents or images. In a recent study, Dodge et al. analyzed C4, the "Colossal Clean Crawled Corpus" used for training many language models and found that up to 14.4

A.4.2 Legal considerations

There are important legal considerations when using datasets, including potential violations of laws governing privacy, discrimination, data protection, intellectual property licenses, fair decision-making processes, consumer protection, and use of an individual's image or likeness, among numerous others. For example, in the United States, numerous legal actions have been brought against companies such as Clearview AI and IBM for their facial recognition datasets. Clearview has been sued under California law for commercial appropriation of individual faces within photographs, violations of California's constitutional privacy protections, and for aiding illegal government surveillance efforts. Both Clearview AI and IBM's Diversity in Faces (DiF) dataset have faced legal action under Illinois' Biometric Information Privacy Act (BIPA). Facebook has also settled a case alleging BIPA violations for $550 million in 2020.

The landscape of potential legal issues applicable to datasets is complex and will vary based on content, jurisdiction, and application. Therefore, it will be important for dataset creators and providers to consistently assess them over time, especially during periods of dramatic changes in the law, such as when GDPR was implemented in May 2018, or if, for example, the EU AI Act becomes law. This is another reason why it is important for dataset creators to be aware of these and other legal issues, and it also impacts downstream users, who may incur liability if they use deprecated datasets with legal issues, even unwittingly. This further supports the urgent need for NeurIPS to support a deprecation protocol, as well as an easily accessible public repository so researchers can check which datasets present risks and may even cause legal liability if they use them.

A.4.3 Social considerations

Datasets necessarily represent a worldview, including how the data is collected, labeled and applied, and these particular representations and classifications of the world have social and political values embedded within them. Systems shaped by these datasets can produce different forms of harm, including allocative harms (a system offering or withholding opportunities from certain groups) and representational harms (a system reinforcing the subordination of particular groups by virtue of identity). As numerous audits of datasets show, such harms tend to disproportionately affect marginalized groups along the intersecting axes of race, ethnicity, gender, ability, and positionality in global hierarchies. For example, after public and scholarly criticism, the creators of ImageNet identified a total of 1,593 harmful labels in the dataset, and subsequently removed them. However, because both Tiny Images and unredacted versions of ImageNet continue to circulate and are potentially used to train production-level systems, these problematic labels and logics could be embedded in ways that entrench harms while being hard to track and investigate: for instance, if an image classification model trained on ImageNet is continuing to use the pre-existing labels such as 'alcoholic,' 'closet queen,' or 'rape suspect.' The continued circulation of problematic data collections threatens to reproduce both allocative harms and representational harms.

A.5 A Dataset Deprecation Framework for NeurIPS

NeurIPS could play a significant role in tracking and addressing these issues by instituting a dataset deprecation framework, consisting of a centralized repository for maintaining status updates, a dataset deprecation sheet, and, optionally a free unique identifier for datasets when they are created.

A.5.1 Central Repository of Deprecated Datasets

Having a central place to store, update and disseminate dataset deprecation information is important to ensure that the ML community is aware of which datasets not to use. NeurIPS is a very suitable entity to act as the keeper of a public, centralized repository of deprecation decisions. This could take the form of a database of deprecation sheets and their accompanying documentation. This would give researchers a single site to visit in order to check if a dataset has been deprecated, thus protecting them from a range of technical, legal, ethical, and organizational problems. It would enable a transparent way to submit, access, and disseminate up-to-date information on a dataset's current status, as well as notifying the ML community when a deprecation is necessary and why. This would also support jurisdiction-specific deprecations, such as when a dataset is illegal in one jurisdiction but allowed in others. The repository maintainer could address this by noting jurisdictional differences in the repository and, if applicable, supporting geo-fence techniques to control access to the dataset going forward. Furthermore, a central repository would enable a reporting function, where people could share issues they have discovered with existing datasets, even if they are not the original creators.

Researchers often discover problems with datasets in the process of their work, but currently have limited options for informing others who are using the same data. A centralized repository would be a useful venue to inform the wider community of any potential issues. As a standard bearer in the ML research community, NeurIPS could implement the deprecation publication check for paper submission that would be a powerful way to create norms across the research community. In this way, conferences can play a gatekeeping role that contributes to curbing the circulation and use of deprecated datasets. It is important to acknowledge the labor, in terms of time and effort, that would be required to host an up-to-date repository of deprecated datasets. It requires an ongoing institutional commitment to the maintenance of this infrastructure.

This is another reason why NeurIPS is the ideal entity for maintaining the repository while dataset creators are responsible for completing the Dataset Deprecation Sheet and lodging it with the repository.

A.5.2 Dataset Deprecation Sheet

The Dataset Deprecation Sheet would communicate all the necessary information to the ML community about a dataset deprecation.

1. Reasons for Deprecation: Using publicly accessible explanations, those responsible for a dataset should clearly explain potential impacts of the deprecation. The discussion of risks

could include which risks are being envisaged and to whom, and over what time frame risks were considered.

2. Execution and Mitigation Plan: dataset creators should provide a plan regarding how the dataset deprecation will happen, including how access to the dataset will be restricted or halted; how changes to access will be announced and maintained on a publicly accessible site; and which derivative datasets are impacted and should potentially also be deprecated.
3. Appeal Mechanism: An appeal mechanism should be included that allows challenges to the deprecation, for users who have strong reasons to need to continue use.
4. Timeline: Deprecation announcements should give stakeholders adequate time to understand the deprecation's rationale, evaluate its impact, and launch any appeals.
5. Post-Deprecation Protocol: Recognizing that deprecated datasets will continue to have value (e.g., as research objects, legal evidence, and historical records), deprecation protocols should also articulate methods for sequestering and accessing datasets post-deprecation, including what principles and procedures will be used to grant access to sequestered datasets. This protocol should be regularly re-evaluated in light of technical changes in data sequestration, new best practices for policy implementation, and any insights gained during the appeals process.
6. Publication Check: NeurIPS could include in the paper acceptance process that authors confirm they are not using deprecated datasets, and their work follows the post-deprecation protocols sanctioned by the conference, or risk that their work be rejected. Additionally, those who are presenting new datasets at the conference should note in their paper that they will follow the conference's framework for dataset deprecation, in the event that their datasets require future removal or modifications.

(Examples of hypothetical dataset deprecation sheets can be provided on request)

A.5.3 Permanent dataset identifiers

Additionally, for a strong technical infrastructure, datasets can be supported by a permanent identifier that accompany datasets from their creation to their deprecation (as well as through updates and version changes). An ideal mechanism for assigning and updating identifiers could be based on the DOI (digital object identifier) system, which has existed for decades. The digital object identifier, or DOI, was introduced by the International Organization for Standardization (ISO) in 2000 as a way to standardize and track the creation and evolution of digital objects ranging from academic articles to government publications. DOIs are fixed and bound to metadata – including the object's URL, date of creation, authors, version, etc. – that can be changed. DOIs are governed by the International DOI Foundation (IDF), a non-profit entity that ensures the perpetuity of the format and prevents third parties from imposing licensing on the system.

Websites such as Zenodo and FigShare have been offering free DOI attribution and maintenance for the academic community, covering various types of outputs ranging from articles to figures and datasets. But the practice of using DOIs to track the creation and modification of datasets remains a rare occurrence in machine learning. This makes it not only hard to find datasets, but to determine their status and current version. Also, while websites such as GitHub, which are also often used for disseminating data, support versioning and change tracking, they often lack structured metadata, datasheets, and interoperability with similar systems. Although using either personal websites or code-hosting repositories may allow for more flexibility in description and indexing, they lack any kind of traceability or standardized structure. In fact, both of these platforms can be used in conjunction with DOI. The existence and usage of DOIs as perennial identifiers of digital objects can serve as a basis for building an infrastructure specifically tailored for the ML community.

A.6 Conclusion

Dataset deprecation is an important part of the dataset life cycle, and good practices for deprecation are overdue. This proposal for a Dataset Deprecation Framework for NeurIPS includes a centralized repository, a dataset deprecation sheet, and an optional system of permanent dataset identification. The framework is informed by a recent review of several major datasets that have been redacted or deprecated where we observed that existing deprecations have been subject to poor documentation

practices, leaving the ML research community with ongoing uncertainty and no clear rationale for stopping their use.

Once NeurIPS has established the central repository and deprecation sheets, it can deploy a publication check to ensure that emerging research no longer uses these data sources.

A.7 Possible Implementation

Establish a committee to oversee dataset deprecation. Committee members could come from the Ethics committee and the Dataset & Benchmark Chairs. In subsequent years when the process is more established, we could move towards a cross-institutional body. Set up a website where people can search for deprecated datasets as well as fill a form to make a request for deprecation. Such requests could come from the dataset creators as well as other researchers. Ideally, the website would mimic a portal like Zenodo. At the very least and perhaps initially, the website could be based upon a github repository for versioning purposes. We need to start associating a DOI to datasets for identifiability purposes, which are particularly important nowadays as we encounter lots of derivative datasets and versions of the same dataset. Authors should include the corresponding DOI in the citation of the dataset. This is a change that should go in the Call for Papers and formatting instructions template. See previous Sec. 2 for the fields that the form requesting deprecation could include.

Appendix B List of relevant resources

- NeurIPS 2021 Ethics guidelines
- NeurIPS 2021 Checklist
- ACM Code of Ethics
- ACL ethics FAQ
- ACL ethics review questions
- ICLR Code of Ethics