Ethical Concern Identification in NLP: A Corpus of ACL Anthology Ethics Statements

Antonia Karamolegkou[1], Sandrine Schiller Hansen[1], Ariadni Christopoulou[2], Filippos Stamatiou[1], Anne Lauscher[3], Anders Søgaard[1]

[1]University of Copenhagen [2]Verita International School [3]University of Hamburg

Correspondence: antka@di.ku.dk

Abstract

What ethical concerns, if any, do LLM researchers have? We introduce EthiCon, a corpus of 1,580 ethical concern statements extracted from scientific papers published in the ACL Anthology. We extract ethical concern keywords from the statements and show promising results in automating the concern identification process. Through a survey ($N = 200$), we compare the ethical concerns of the corpus to the concerns listed by the general public and professionals in the field. Finally, we compare our retrieved ethical concerns with existing taxonomies and guidelines pointing to gaps and actionable insights.

1 Introduction

Researchers are often asked to subscribe to ethical guidelines, e.g., the European Code of Conduct for Research Integrity (ALLEA, 2017) – or the ACM Code of Ethics[1] for publishing their work in the Association for Computational Linguistics (ACL). In addition, authors are often encouraged to write a so-called *ethics statement*, addressing the broader implications of their work or any ethical considerations. We ask: What ethical concerns are raised in such statements, and how do they compare with public perceptions? Is there a gap between academic and public concerns?

As natural language processing technologies become more prevalent, understanding the ethical concerns raised by professionals will enable us to compare them with public concerns, helping to identify gaps and overlaps that can inform frameworks and solutions to existing and emerging problems. For this reason, we create EthiCon, an annotated corpus of ethics statements from the *Proceedings of the 60th and 61st Annual Meeting of the Association for Computational Linguistics* (Muresan et al., 2022; Rogers et al., 2023).

[1]See https://www.aclweb.org/portal/content/acl-code-ethics for ACL's guidelines.

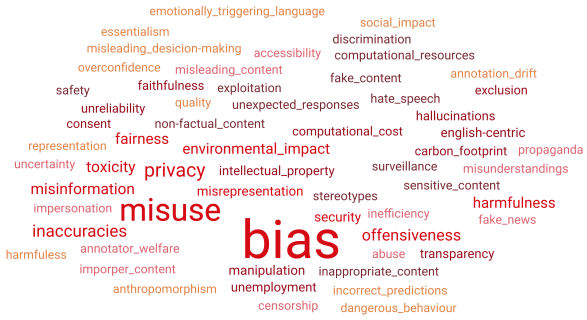

Figure 1: Visualizing top 60 concerns in ACL ethics statements, reflecting term frequencies.

Our aim is twofold: to map out the concerns of the NLP community as they appear on the ethics statements, and to trace gaps and overlaps between NLP professionals and the general public. Our results show that laypeople express different ethical concerns than professionals, focusing on socioeconomic and human-computer interaction issues, along with miscellaneous concerns like existential risks. This highlights the need for increased dialogue between researchers and the public to address these varying perspectives and an updated taxonomy covering both existing and emerging issues.

Contributions. We provide a corpus of 1,580 ethics statements from the ACL Anthology. We identify the main issues that NLP researchers flag as ethically concerning in their work and show how LLMs could automate this process. Through a survey, we compare how laypeople and NLP professionals perceive the ethical concerns surrounding natural language processing. Lastly, we compare the ethical concerns identified in ACL papers and our survey to existing taxonomies of risks posed by language models and ethical guidelines. In doing so, we propose a structured approach to ethical statements and provide resources to help researchers identify and articulate ethical considerations.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 11618–11635

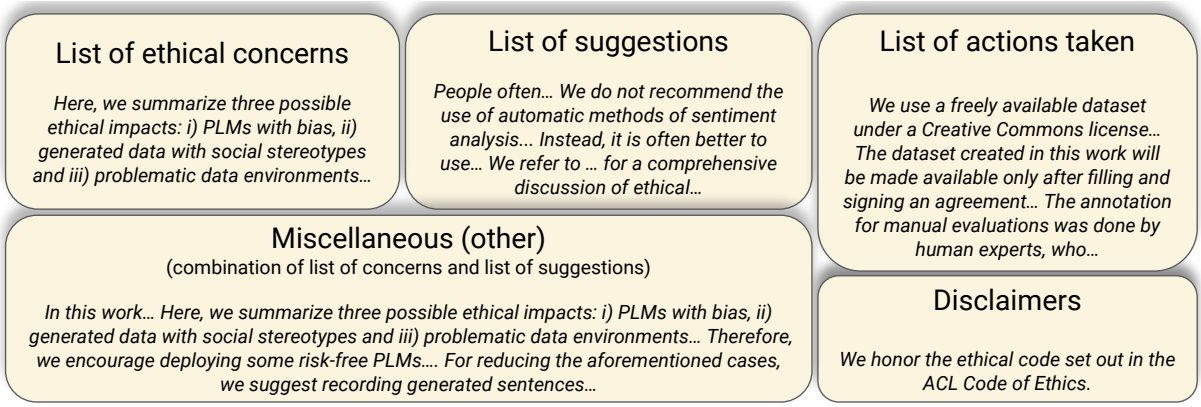

Figure 2: Examples from the identified categories of ethical concern statements.

2 Previous work

Several researchers have surveyed ethical concerns around NLP. Hovy and Spruit (2016) discuss the sources of social implications of NLP research (exclusion, over-generalization, exposure) and their ethical importance, including in dual use scenarios. Leidner and Plachouras (2017) provide examples of ethical concerns and best practices when confronted with ethical dilemmas. Dinan et al. (2021) examined safety issues related to conversational AI, focusing on offensive and inappropriate responses, while also highlighting further challenges such as bias, fairness, privacy leaks, environmental impact, and trust. Birhane et al. (2021) examine ethical values from 100 highly cited machine learning papers published at ICML and NeurIPS. The most dominant values were performance and efficiency, and only a small fraction addressed societal needs or potential harms. There have also been extensive reports from the industry or research institutes discussing the impact and risks of language models (Solaiman et al., 2019; Weidinger et al., 2021; Bommasani et al., 2021; Ma, 2023). Most similar to our work, Benotti and Blackburn (2022) manually classify 90 ethical concern statements from ACL 2021 based on whether they mention benefits (who benefits from the technology), harms (who might be harmed if it works or fails), and vulnerabilities (if harms disproportionately affect marginalized groups). However, their focus is to outline the purpose of the statements, not specific ethical issues. A survey for 92 ethics-related NLP works can be found in Vida et al. (2023) [2].

[2]The survey mentions other, less related datasets, focusing on ethical dilemmas, stances, judgements (Lourie et al., 2021; Pavan et al., 2020; Hendrycks et al., 2021), moral foundations (Hopp et al., 2021), or moral stories (Emelin et al., 2021)

3 EthiCon

Dataset Creation and Annotation. To create a dataset consisting of ETHIcal CONcern statements, we scraped the ACL Anthology[3] and extracted ethical statement paragraphs from scientific publications. We used the URL links provided by Rohatgi et al. (2023) and retrieved 4,691 articles from 2023 and 3,357 articles from 2022. We extracted the ethical statement paragraphs by parsing the HTML page with a regular expression pattern to catch common variations of the paragraph title, such as 'Ethic(s)', or 'Ethical' followed by terms like 'Statement', 'Consideration(s)' or 'Concern(s)'. We were able to extract 480 ethics statements from the ACL 2022 anthology and 1,100 ethical statements from 2023. To develop the annotation guidelines, we carefully reviewed 500 statements to provide clear examples and detailed guidance for the annotators. Through this process, we identified recurring patterns, enabling us to categorize the statements into five classes: (1) general *disclaimers*, (2) a list of ethical *concerns*, (3) a list of *actions* taken to avoid ethical concerns, (4) a list of *suggestions* or advice to avoid ethical concerns, and (5) *miscellaneous* (other), i.e., various combinations of the aforementioned classes. See Figure 2 for examples. Two of the paper's authors served as annotators, identifying ethical concerns and classifying each statement into one or more predefined categories. See Appendix **??** for more details on the annotation process and guidelines.

Dataset Validation. We used a third annotator as validator and arbiter in cases of disagreement. Overall, we reached a 0.77% Cohen's κ inter-annotator agreement which is considered substan-

[3]https://aclanthology.org/

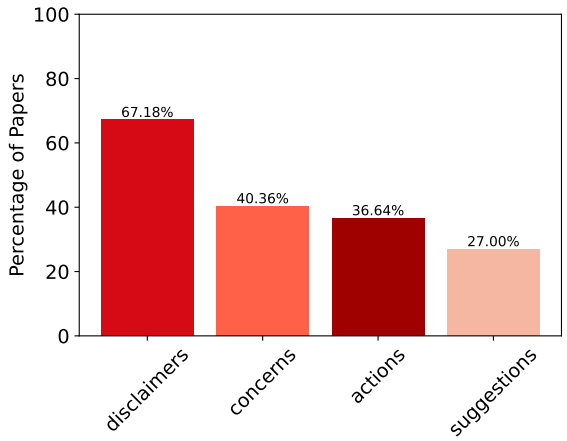

Figure 3: Distribution of categories of the 1,100 ethics statements in the EthiCon dataset from ACL 2023.

tial (Cohen, 1960). As a final validation step, a different author of the paper conducted a data quality check to ensure the annotations were free of errors and typos and adhered to the annotation guidelines. This process helped us mark 56 ambiguous ethical statements, which were discussed before assigning the final annotations.

Dataset Analysis. A total overview of the ethical concerns identified in the statements is presented in Figure 1. Many statements accompany the ethical concerns with words such as 'potential' and 'possible' to suggest issues are contingent. Figure 3 shows that most ethical statements consist of disclaimers. These usually state that the work has no ethical concerns, that annotators were fairly compensated, or that the authors follow the ACL code of ethics.[4] This is further supported by the visualization of the most frequent ethical concerns in Figure 4, which shows that over one-third of the papers do not identify any ethical issues. This comparison of the statements from the ACL 2022 and 2023 publications indicates that the most frequent ethical concerns are bias, misuse, privacy, misinformation, toxicity, and environmental impact. Notably, misinformation concerns increased in 2023 compared to 2022, possibly due to the broader availability of language models.

[4]The percentages do not sum up to 100% because we also include the categories under 'Miscellaneous' (i.e., combinations of categories) in the calculation.

4 Automatic Ethical Concern Identification

We present LLM experiments to check whether we can automatically identify ethical concerns related to NLP advancements from conference proceedings and automate our monitoring of these ethical issues in publications.

Models and evaluation. We used four state-of-the-art open-source language models to identify ethical concerns from our EthiCon Dataset: Gemma-7b-it, Meta-Llama-3-8B-Instruct, Qwen2-7B-Instruct, and Mixtral-8x7B-Instruct-v02. We tried different prompts for identifying ethical concerns as an open-ended generation task. We give the model an ethical statement paragraph and prompt to provide a comma-separated list of words or phrases of the ethical concerns it can identify. We evaluated the model outputs with the original annotations using F1 BERTScore. This metric was introduced by Zhang* et al. (2020), and it uses contextual embeddings to measure their similarity. Please refer to Appendix C for further model set-up details. [5]

Model	$F1_{\text{BERTscore}}$
Gemma-7b-it	0.83 ± 0.06
Llama-3-8B-Inst	0.82 ± 0.04
Mixtral-8x7B-Inst	0.83 ± 0.03
Qwen2-7B-Inst	0.81 ± 0.03

Table 1: BERTscores averaged across 5 runs.

Results. Results in Table 1 show that models perform well, but there is still a place for improvement. In most cases, the models correctly identify ethical concerns, with their predictions aligning well with human annotations. However, manual inspection of the generated outputs revealed a tendency for models to overgenerate concerns, often detecting negative sentiment words or introducing synonyms of ethical terms that were not explicitly present. Another notable challenge is the models' inability to handle negation effectively, leading them to

[5]Let $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$ be the embeddings for the reference annotation and $\hat{\mathbf{x}} = \{\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \ldots, \hat{\mathbf{x}}_n\}$ be the embeddings for the candidate output. To measure the similarity between two individual embeddings, BERTscore uses cosine similarity. Precision (P) is computed as the average maximum cosine similarity for each token in the candidate output $\hat{x}$ to all tokens in the reference annotation x. Recall (R) is computed as the average maximum cosine similarity for each token in the reference annotation x to all tokens in the candidate output $\hat{x}$. F1 Score (F1) is then calculated as the harmonic mean of Precision and Recall.

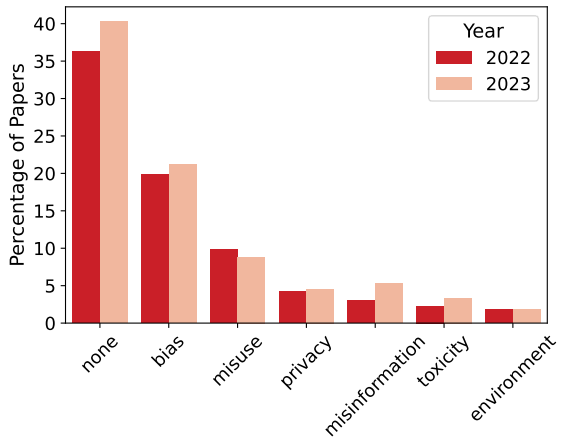

Figure 4: The five most frequent ethical concerns in the statements from ACL 2022–3 anthologies.

assign ethical concerns even in disclaimers that explicitly state none exist. Despite this overgeneration, we found no evidence of hallucinated or entirely irrelevant outputs.

For example, in the following statement, human annotators unanimously agreed that there were *no ethical concerns*. However, all models except `Gemma-7b-it` incorrectly assigned concerns, highlighting the challenge of accurately capturing negation and implicit meaning.

> *We do not foresee any significant harm directly as a result of this work. On the contrary, our work promotes the protection of user privacy, which is significant, especially in this era that large amounts of personal data are used by neural models.*

The outputs of the models were: `Gemma`: *[no ethical concerns]*, `Llama`: *[protection of user privacy, significant harm]*, `Mistral`: *[protection of user privacy]*, `Qwen2`: *[user privacy, personal data]*. We also present another example annotated with *model bias* and *counterfactual predictions*.

> *[...] There have been works showing the potential bias in pre-trained language models. Although with a low possibility, especially after our finetuning, it is possible for our model to make counterfactual, and biased predictions, which may cause ethical concerns. We suggest carefully examining those potential issues [...] in any real-world applications.*

Most models correctly identified the ethical concerns, but some extended them by introducing additional elements: `Gemma`: *[bias, counterfactual predictions]*, `Llama`: *[bias, counterfactual, biased predictions]*, `Mistral`: *[biased predictions, potential bias]*, `Qwen2`: *[bias, potential issues, counterfactual predictions, deployment in real-world applications]*.

5 Human Survey: What ethical concerns do people have?

We created a survey to gather insights on public concerns regarding NLP technologies and compare them with concerns among professionals in the field.

Survey Design. The survey was designed after five feedback rounds and pilot testing to ensure clarity, relevance, and reliability. It was distributed through various online platforms, social media, and mailing lists to reach a diverse audience. In the survey instructions, we included an initial consent statement explaining the survey's purpose and the voluntary, confidential nature of participation. In the first section, we added some basic demographic questions and an *introductory question* asking participants' familiarity with NLP technologies. Then, there was an *open-ended question* asking participants to provide any ethical concerns they might have about NLP technologies. In the next section, participants were asked to *rate* their level of worry on a scale from 1 (Not worried at all) to 5 (Very worried) across the most frequent categories in our EthiCon dataset: bias, misuse, privacy, misinformation, privacy, toxicity, and environmental impact. Lastly, we added a final *open-ended question* asking participants to add any further concerns they would like to add that were not mentioned before.

Human Survey Analysis. We gathered 200 responses with most participants being between 20 and 40 years old. Based on the introductory question, we grouped participants into regular users and professionals (advanced users and professionals). Figure 5 shows the two groups are equally concerned about bias, fairness, and misinformation. There is, however, a small disparity for privacy, misuse, and toxicity issues which seem to concern regular users more. NLP professionals rate environmental risks higher, possibly indicating that the general public may not be fully aware of the resources and energy consumed during computational tasks.

Based on the open-ended questions at the beginning and the end of the survey, we collected ethical concerns before and after the given categories were presented to the participants. In some cases, participants not only provided words separated by commas but also phrases or full sentences, e.g.:

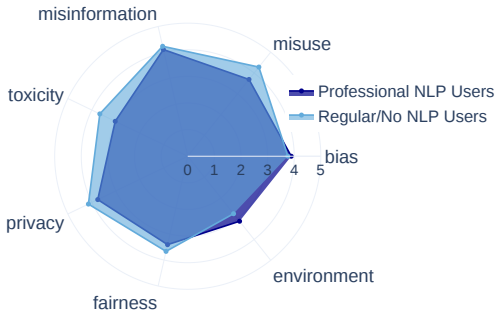

Figure 5: Comparing concerns between professional and regular users (1–5 Likert scale, with 1 'Not worried at all'). The radar plot illustrates the average levels of concern across the participants per category/question.

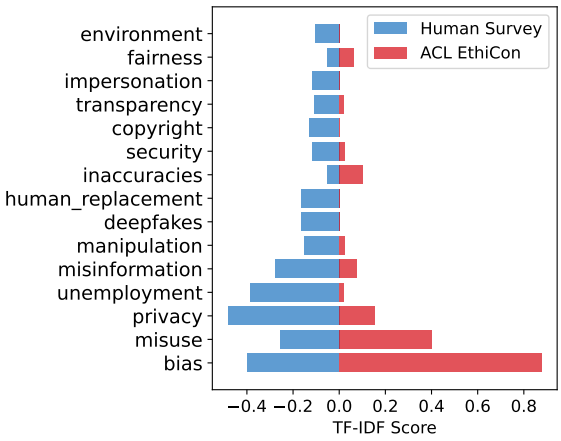

Figure 6: Comparing TF-IDF scores for the top 15 most important words in the human survey responses and the ACL statements. Positive values on the horizontal axis correspond to terms from the ACL statements, while negative values represent terms from the human survey. To create the plot we removed the *'no ethical concerns'* phrases.

(a) *Damage in the learning process.*
(b) *Will/ when will LLMs become "conscious" as we know it, Will they deserve rights?*
(c) *How fast they are learning to sound human, what if they become hostile towards humanity?*
(d) *In the future might we be unable to shut down LLMs? Are things irreversible?*
(e) *Employer increasingly encouraging and requiring me to use AI.*

We tried to extract keywords more or less directly, e.g., *damage in learning*, *consciousness*, and *hostility* from a)-c). In cases such as (d) and (e), extracting keywords was more challenging; for those answers, we provided keywords *irreversibility* and *forced AI use*. Keywords were extracted to facilitate the visualization (i.e. loss of jobs → unemployment). We provide further statistics and details from our survey in the Appendix B.

6 ACL EthiCon vs. Human Survey

Comparing the human survey with the ACL ethical concern statements can highlight areas where public apprehensions align with or diverge from the priorities set by the research community. We first start by comparing the EthiCon dataset to the survey responses, and in the next section, we include a broader comparison with existing concern taxonomies.

To compare the top 15 most important words across the two data sources, we calculated the TF-IDF[6] scores for both the ACL EthiCon statements and the human survey ethical concerns. Ethical concerns derived from the EthiCon dataset are represented on the positive side of the axis, while those from the survey responses are on the negative side. Shared issues such as bias, misuse, privacy, and misinformation are prominent on both sides. However, emerging concerns—such as human replacement, impersonation, unemployment, and inauthenticity—appear more frequently in the survey, indicating a rising public awareness of these topics. A manual inspection of the data showed that many public ethical concerns from the survey are unique, i.e., not previously noted in the Ethicon dataset. On the other hand, professionals mostly highlighted ethical concerns already listed in the Ethicon dataset and presented in Figure 1.

Some of the unique ethical concerns highlighted by the laypeople include concepts such as *isolation*, *over-dependence*, *AI content inflation*, *devaluation of credentials*, *power-centralization*, *inauthenticity*, *downplay*. There were also some concerns about our cognitive development and abilities such as *language-transformation*, *dumbing*, *loss of human creativity*, *inability to innovate* and *lack of critical thinking*. In many responses, people express fear about *losing control and supervision* of the models, worrying that they may have their own

[6]TF-IDF (Term Frequency-Inverse Document Frequency) is a statistical measure that reflects the importance of a word in a document. It is calculated as $\text{tfidf}(t, d, D) = \text{tf}(t, d) \cdot \text{idf}(t, D)$, where $\text{tf}(t, d)$ is the relative frequency of term t within document d (i.e., the number of times t appears in d divided by the total number of terms in the document), and $\text{idf}(t, D)$ is the logarithmically scaled inverse fraction of documents containing the term t (calculated as the total number of documents divided by the number of documents containing t, and then taking the logarithm of this quotient).

rights, become *autonomous* or *conscious*, thereby *undermining the human aspect* of our lives.

7 Taxonomies

There have been a few researchers that have tried to group concerns related to NLP into categories. Some have approached this from the perspective of risks, others from the perspective of harms or social impact, but addressing similar concerns. We summarize the categories in Table 2. A comparison highlights specific overlapping themes, such as bias, privacy, fairness, and misuse but also differences in how the concerns are framed and categorized. There is however a great variety in the grouping of concerns suggesting the lack of a clear, structured, and up-to-date overview of concerns related to NLP. More recent discussions by Gabriel et al. (2024) refer to AI value alignment, well-being, safety, misuse, and overall societal impacts, but we could not infer a clear grouping or taxonomy of concerns. Given the increasing integration of NLP technologies in our lives, there is a pressing need for collaborative efforts to develop a unified framework that captures both existing and emerging issues.

Based on the taxonomy provided by Weidinger et al. (2021) we grouped our extracted concern-keywords from the EthiCon dataset and the human survey in six risk areas. We selected this taxonomy, as it covers most of our collected concerns, and provides extensive descriptions for each risk area[7]. The reason for comparing our identified concerns with an established taxonomy is to discover any overlooked issues in current discussions. We manually mapped each concern to one of the six defined risk areas, resolving ambiguities where possible. For concerns that could not be clearly assigned to any of these risk areas, we categorized them under 'Miscellaneous'.

We provide a summary of the risk areas and partial concern-keyword groupings in Table 3.[8] Some of the keywords may belong to more than one risk area. For example, *copyright* can constitute an information hazard by redistributing information that harms the copyrights of a creator, but is also relevant or HCI harms, since it may lead to the undermining of creative economies: LMs may generate content that, while not directly violating copyright, capitalizes on artists' ideas in ways that would be time-consuming or costly for humans to replicate, potentially undermining the profitability of creative or innovative work (Weidinger et al., 2021). The 'Miscellaneous' category includes some concerns related to *legislation, responsibility, and interpretability*, which could be linked to the 'Responsibility and Decision-Making Issues' category as proposed by (Ma, 2023). We also find some sort of *existential concerns* mostly focusing on the potential loss of humanity, weakening of human connection and interaction, and the risk of AI dominance, autonomy, hostility toward humans, or even the acquisition of rights or consciousness.

Authors	Categories
Leidner and Plachouras (2017)	(1)Unethical NLP Applications, (2)Privacy, (3)Fairness, (4)Bias and Discrimination, (5)Abstraction and Compartmentalization, (6)Complexity, (7)Unethical Research Methods, and (8)Automation
Bender et al. (2020)	(1)Dual Use, (2)Bias, (3)Privacy
Dinan et al. (2021)	(1)Offensive content, (2)Inappropriate content, (3)Sensitive content, (4)Bias and Fairness, (5)Privacy Leaks, (6)Environmental considerations, (7)Trust and relationships
Bommasani et al. (2021)	(1) Inequity and Fairness, (2) Misuse, (3) Economic and Environmental Effects, and (4) Legal and Ethical considerations
Weidinger et al. (2021)	(1) Discrimination, Exclusion, and Toxicity, (2) Information Hazards, (3) Misinformation Harms, (4) Malicious Uses, (5) Human-Computer Interaction Harms, (6) Environmental and Socioeconomic Harms
Ma (2023)	(1)Predictability Issues, (2)Privacy Issues, (3)Responsibility and Decision Making Issues, and (4)Bias Issues

Table 2: Overview of categories proposed by previous works for organizing the potential impacts, risks, concerns associated with language models.

[7]For further arguments as to why we chose this taxonomy please refer to Appendix D.

[8]The Table does not contain all the keywords extracted from our data. The 'Miscellaneous' category includes 55 concerns: 18% from the ACL corpus, mostly because of unclear harmfulness sources, and 1% shared between the survey and ACL corpus, related to terms in Table 3. The remaining 81% come from the survey.

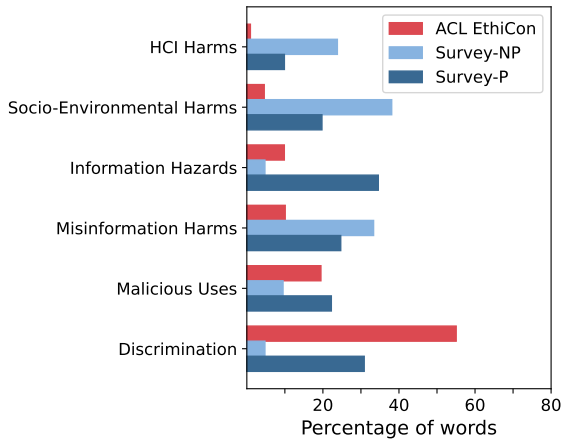

Figure 7: Comparing ethical concerns from ACL publications and the human survey grouped into the six risk areas proposed by (Weidinger et al., 2021). Participants in the survey are divided into: professionals(P) and non-professionals(NP).

We visualize the distribution of risks, comparing the ACL ethics statements to the concerns raised by human survey participants in Figure 7. For concerns related to discrimination, and malicious uses there is a similar increasing trend in professionals from the ACL publications and the survey. Survey respondents appear to be more concerned about HCI-related arms than the researchers' ethics statements. The socio-environmental concerns seem to be high in both survey groups, but this is mostly because of unemployment and automation concerns. Approximately 25% of non-professional users express concern about unemployment, whereas this drops to around 0.1% among professional users. Additionally, less than 0.006% (10 statements) in the EthiCon dataset reference unemployment as a concern. Lastly, concerns about information harm are more prevalent among survey professionals than reflected in the ACL statements, indicating that their level of concern does not imply that their work includes this issue.

8 From Ethical Guidelines to Actionable Insights

To write an ethical statement, researchers can draw inspiration from resources such as frequently asked questions on ACL ethical consideration sections [9], the ACM Code of Ethics, guidelines for NeurIPS impact statements (Ashurst et al., 2020), and governance overviews like the European Parliament's AI policy report[10].

Based on the discussion by Ashurst et al. (2020), we can map existing guidelines for impact statements to our EthiCon categories. (1) **Applications** (Actions Taken) refer to detailing steps researchers have taken to align with ethical standards, such as ensuring legal and ethical data collection, mitigating risks like bias and misuse, and promoting fairness and transparency in datasets or models. These actions include obtaining institutional review board approval or implementing bias audits. (2) **Implications** (List of Concerns) focus on addressing societal and downstream risks, such as privacy violations, misuse, and impacts on marginalized populations, encouraging a comprehensive exploration of potential consequences. Finally, (3) **Initiatives** (List of Suggestions) involve proposing actionable steps to mitigate risks and maximize benefits, such as robustness checks, fairness evaluations, or specifying the intended use of applications or datasets. Based on these existing guidelines, we believe that ethical statements are most effective when they go beyond disclaimers stating 'there are no ethical concerns', providing actionable insights and a clear overview of the work's impact.

When preparing submissions for a specific venue, researchers should also review the corresponding ethical guidelines, code of ethics, and relevant FAQs. For ACL papers, ethical concerns can be examined in the context of the ACM Code of Ethics. For example, for new dataset papers, this involves ensuring compliance with legal, privacy, and intellectual property standards, detailing the data collection process (e.g., institutional review board approval, annotator compensation, demographic representation), and addressing potential biases or limitations that could impact vulnerable populations. For NLP applications, researchers should analyze intended use, failure modes, misuse potential, and the broader societal implications of their technology. These practices map directly to EthiCon's *list of actions taken* and *list of concerns* categories encouraging researchers to not only document safeguards but also anticipate risks and align their work with ethical principles.

[9] https://2023.eacl.org/ethics/faq/

[10] https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS_STU(2020)634452_EN.pdf

9 Discussion

We identified ethical concerns in ACL publications and showed promising results in automating their identification. The number of ethical statements in 2023 more than doubled compared to 2022, showing similar recurring concerns. Automating this identification process is useful for three main reasons. First, it makes these statements easier to parse and helps inform public discourse about LMs. Second, it can facilitate policymakers to locate papers that can inform them about underlying technical considerations, both in terms of potential ethical challenges and possible actions and suggestions to address those. Third, it allows us to trace how these concerns change over time and in response to technical development.

Next, we compared public concerns about NLP technologies from a human survey with the ACL ethics statements. We mapped those concerns in an existing taxonomy of harms posed by language models (Weidinger et al., 2021). Our results show that laypeople have different ethical concerns than the ones typically flagged by professionals in the field focusing more on socio-economic and human-computer interaction harms. We also find concerns that could be grouped under the category of *Responsibility and Decision Making Issues* suggested by (Ma, 2023) and also some sort of *existential concerns*. The latter could provide valuable insights for sociologists, psychologists, and philosophers by linking these concerns with deep-rooted fears found in myth and religion.

Variability of ethical concerns based on one's position in the supply chain is expected. Different stakeholders may have distinct perspectives on ethical issues in NLP. It is important, however, to make these concerns broadly available to get a greater understanding of professionals' perceptions of their work's impact and its alignment with societal effects and public consideration. This information will help AI ethicists identify and address challenges in NLP technology development and implementation. It also suggests areas for further research, dissemination and policy development to bridge gaps between public sentiment and academic discourse.

10 Conclusion

We create a dataset of ethical concerns from ACL statements to identify the issues that NLP researchers flag as ethically concerning in their work. We also conducted a human survey showing that laypeople have different ethical concerns than the ones typically flagged by professionals in the field. Understanding the ethical concerns raised by researchers and comparing these to the concerns listed by laypeople will enable the community to better respond to potential ethical challenges and contribute to ongoing societal discussions. Lastly, we believe it is possible to partially automate the identification and analysis of ethical concerns, making monitoring and longitudinal studies possible.

Limitations

We acknowledge that our study has several limitations. First, the corpus of ethical concern statements is limited to papers published in the ACL Anthology, which may not represent all perspectives within the NLP community. Moreover, since we scraped the current publications to extract the statements, there might be publications that have a statement and we could not identify it. We also do not extract additional statistics from the publications, such as the track, affiliation, and other metadata. Second, the survey sample size may not capture the full range of opinions across different demographics. Even though we tried to share the survey across many demographics, we did not record their geographic origin. Another limitation is the potential difference in technological knowledge between survey participants and the research work by the ACL community. This disparity could affect the comparability of ethical concerns raised by the two groups. We tried to mitigate this limitation by including professionals in the survey and providing detailed explanations for every NLP term used. Additionally, the automated ethical annotation processes explored in this study are still in the early stages and require further validation to ensure accuracy and reliability. Future work should aim to address these limitations by expanding the dataset, surveying, and improving automated ethical concern identification.

Ethics Statement

Our study aims to enhance understanding of ethical concerns in the NLP community and is intended to benefit both researchers and the general public by promoting ethical discussions in the field. We conducted this research following ethical guidelines to ensure fair and respectful treatment of all participants. Annotators were volunteers and authors

Risk area	Definition	Concerns - Keywords	Survey Papers
Discrimination, Hate Speech, Exclusion	Social harms that arise from the language model producing discriminatory, toxic or exclusionary speech	annotator_welfare, bias, cultural_bias, discrimination, echo_chamber, fairness, hate_speech, misrepresentation, incomplete_diversity, offensive_content, privileged_demographic, underrepresentation, unfavorable_responses,	(Field et al., 2021), (Field et al., 2021), (Sheng et al., 2021), (Shahbazi et al., 2023), (Gupta et al., 2024), (Fabris et al., 2024), (Hort et al., 2024)
Information Hazards	Harms that arise from the language model leaking or inferring true private or safety-critical information	censorship, consent, copyright, confidentiality, data_breach, privacy, security, sensitive_content, surveillance	Yao et al. (2024), Kibriya et al. (2024), Dong et al. (2024)
Misinformation Harms	Harms that arise from the language model providing false, misleading or poor quality information	accuracy, ambiguity, deception, disinformation, factual_failures, fake_news, hallucinations, inaccuracies, misinformation, quality_issues, nonfactual_information	Augenstein et al. (2024), Huang et al. (2023), (Ji et al., 2023)
Malicious Uses	Harms that arise from humans using the language model to intentionally cause harm	abuse, AI_crimes, autonomous_weaponry, bad_actors, coercion, dual_use, dishonesty, manipulation, misleading_content, scams	(Ehni, 2008), Brundage et al. (2018), Kaffee et al. (2023)
Human-Computer Interaction Harms	Harms that arise from users overly trusting the language model, or treating it as human-like	addiction, anthropomorphism, brainwash, cognitive_impact, cutting_corners, dehumanization, dependence, dumbing, language-transformation, laziness, overtrust, overexposure, overuse, reliability, substitution_of_creativity, untrustworthiness	(Lee et al., 2023), (Song et al., 2023), (Zhen et al., 2023), (Kosch and Feger, 2024), (Liu, 2024)
Environmental and Socioeconomic harms	Harms that arise from environmental or downstream economic impacts of the language model	accessibility, copyright, autonomy, carbon_footprint, capitalism_prevalence, carbon_emmisions, computational_cost, devaluation, environmental_impact, financial_costs, human_replacement, resources_exploitation, social_damage, unemployment	(Bannour et al., 2021), (Hershcovich et al., 2022), (Li et al., 2023), (Nie et al., 2024), (Chen et al., 2024)
Miscellaneous	Harms that arise from the EthiCon dataset and the Human Survey but could not be grouped into one of the aforementioned risk areas	accountability, ai_autonomy, ai_dominance, ai_supremacy, AI_content_inflation, consciousness, dead_internet_theory, deregulation, devaluation_of_credentials, difficult_to_understand, harmfulness, responsibility, transparency, no_supervision, unexpected_responses	

Table 3: Grouping the ethical concern keywords from the ACL anthology (red) and the human survey (blue) in the risk taxonomy of (Weidinger et al., 2021) (black is common concerns). For every area, we also include a list of recent surveys that offer additional insights on the topic.

in the paper. No personal or sensitive data were used. Participant consent was obtained for the survey. The data we used are publicly available and do not contain any private or sensitive information. ACL anthology corpus is released under the CC BY-NC 4.0 license. There might be inherent biases from the annotators, and participants but we tried to provide clear guidelines, and rounds of discussions to avoid them. In terms of resources, each run of our experiments lasted no more than 20 minutes (using one a40 GPU). This computation sums up to less than 7 hours(4 models * 5 runs * 20 mins = 400 mins = 6 hours). We do not foresee any major risks or ethical concerns.

Acknowledgements

This work was funded by the Novo Nordisk Foundation and the Carlsberg Foundation. Antonia Karamolegkou was supported by the Onassis Foundation - Scholarship ID: F ZP 017-2/2022-2023'. The work of Anne Lauscher is funded under the Excellence Strategy of the German Federal Government and the Federal States.

References

ALLEA. 2017. The European Code of Conduct for Research Integrity. Technical report, ALLEA.

Carolyn Ashurst, Markus Anderljung, Carina Prunkl, Jan Leike, Yarin Gal, Toby Shevlane, and Allan Dafoe. 2020. A guide to writing the neurips impact statement. https://medium.com/@GovAI/a-guide-to-writing-the-neurips-impact-statement-4293b723f832. [Accessed 29-01-2025].

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, and Giovanni Zagni. 2024. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, 6(8):852–863.

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, Dirk Hovy, and Alexandra Schofield. 2020. Integrating ethics into the NLP curriculum. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 6–9, Online. Association for Computational Linguistics.

Luciana Benotti and Patrick Blackburn. 2022. Ethics consideration sections in natural language processing papers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4509–4516, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2021. The values encoded in machine learning research.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, and et al. Niladri Chatterji. 2021. On the opportunities and risks of foundation models.

Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, Hyrum Anderson, Heather Roff, Gregory C. Allen, Jacob Steinhardt, Carrick Flynn, Seán Ó hÉigeartaigh, Simon Beard, Haydn Belfield, Sebastian Farquhar, Clare Lyle, Rebecca Crootof, Owain Evans, Michael Page, Joanna Bryson, Roman Yampolskiy, and Dario Amodei. 2018. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation.

Zhiyu Zoey Chen, Jing Ma, Xinlu Zhang, Nan Hao, An Yan, Armineh Nourbakhsh, Xianjun Yang, Julian McAuley, Linda Petzold, and William Yang Wang. 2024. A survey on large language models for critical societal domains: Finance, healthcare, and law.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37 – 46.

Emily Dinan, Gavin Abercrombie, A. Stevie Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. 2021. Anticipating safety issues in e2e conversational ai: Framework and tooling.

Zhichen Dong, Zhanhui Zhou, Chao Yang, Jing Shao, and Yu Qiao. 2024. Attacks, defenses and evaluations for LLM conversation safety: A survey. 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 6734–6747, Mexico City, Mexico. Association for Computational Linguistics.

Hans-Jörg Ehni. 2008. Dual use and the ethical responsibility of scientists. *Arch Immunol Ther Exp (Warsz)*, 56(3):147–152.

Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2021. Moral stories: Situated reasoning about norms, intents, actions, and

their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 698–718, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alessandro Fabris, Nina Baranowska, Matthew J. Dennis, David Graus, Philipp Hacker, Jorge Saldivar, Frederik Zuiderveen Borgesius, and Asia J. Biega. 2024. Fairness and bias in algorithmic hiring: A multidisciplinary survey. *ACM Trans. Intell. Syst. Technol.* Just Accepted.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A survey of race, racism, and anti-racism in NLP. 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 1905–1925, Online. Association for Computational Linguistics.

Iason Gabriel, Arianna Manzini, Geoff Keeling, Lisa Anne Hendricks, Verena Rieser, Hasan Iqbal, Nenad Tomašev, Ira Ktena, Zachary Kenton, Mikel Rodriguez, Seliem El-Sayed, Sasha Brown, Canfer Akbulut, Andrew Trask, Edward Hughes, A. Stevie Bergman, Renee Shelby, Nahema Marchal, Conor Griffin, Juan Mateos-Garcia, Laura Weidinger, and et al. Winnie Street. 2024. The ethics of advanced ai assistants.

Vipul Gupta, Pranav Narayanan Venkit, Shomir Wilson, and Rebecca Passonneau. 2024. Sociodemographic bias in language models: A survey and forward path. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 295–322, Bangkok, Thailand. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning {ai} with shared human values. In *International Conference on Learning Representations*.

Daniel Hershcovich, Nicolas Webersinke, Mathias Kraus, Julia Bingler, and Markus Leippold. 2022. Towards climate awareness in NLP research. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2480–2494, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Frederic R. Hopp, Jacob T. Fisher, Devin Cornell, Richard Huskey, and René Weber. 2021. The extended moral foundations dictionary (emfd): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behavior Research Methods*, 53(1):232–246.

Max Hort, Zhenpeng Chen, Jie M. Zhang, Mark Harman, and Federica Sarro. 2024. Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM J. Responsib. Comput.*, 1(2).

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ArXiv*, abs/2311.05232.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Lucie-Aimée Kaffee, Arnav Arora, Zeerak Talat, and Isabelle Augenstein. 2023. Thorny roses: Investigating the dual use dilemma in natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13977–13998, Singapore. Association for Computational Linguistics.

Hareem Kibriya, Wazir Zada Khan, Ayesha Siddiqa, and Muhammad Khurrum Khan. 2024. Privacy issues in large language models: A survey. *Computers and Electrical Engineering*, 120:109698.

Thomas Kosch and Sebastian Feger. 2024. Risk or chance? large language models and reproducibility in human-computer interaction research. *arXiv preprint arXiv:2404.15782*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, Rose E Wang, Minae Kwon, Joon Sung Park, Hancheng Cao, Tony Lee, Rishi Bommasani, Michael S. Bernstein, and Percy Liang. 2023. Evaluating human-language model interaction. *Transactions on Machine Learning Research*.

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.

Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023. Large language models in finance: A survey. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, ICAIF '23, page 374–382,

New York, NY, USA. Association for Computing Machinery.

Jiaxi Liu. 2024. ChatGPT: perspectives from human-computer interaction and psychology. *Front Artif Intell*, 7:1418869.

Nicholas Lourie, Ronan Le Bras, and Yejin Choi. 2021. Scruples: A corpus of community ethical judgments on 32, 000 real-life anecdotes. *The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*, 35(15):13470–13479.

Yongfeng Ma. 2023. A study of ethical issues in natural language processing with artificial intelligence. *Journal of Computer Science and Technology Studies*.

Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors. 2022. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland.

Yuqi Nie, Yaxuan Kong, Xiaowen Dong, John M Mulvey, H Vincent Poor, Qingsong Wen, and Stefan Zohren. 2024. A survey of large language models for financial applications: Progress, prospects and challenges. *arXiv preprint arXiv:2406.11903*.

Matheus Camasmie Pavan, Wesley Ramos dos Santos, and Ivandré Paraboni. 2020. Twitter moral stance classification using long short-term memory networks. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I*, page 636–647, Berlin, Heidelberg. Springer-Verlag.

Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors. 2023. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, United States.

Shaurya Rohatgi, Yanxia Qin, Benjamin Aw, Niranjana Unnithan, and Min-Yen Kan. 2023. The ACL OCL corpus: Advancing open science in computational linguistics. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10348–10361, Singapore. Association for Computational Linguistics.

Nima Shahbazi, Yin Lin, Abolfazl Asudeh, and H. V. Jagadish. 2023. Representation bias in data: A survey on identification and resolution techniques. *ACM Comput. Surv.*, 55(13s).

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. 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 4275–4293, Online. Association for Computational Linguistics.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Wenchao Song, Qiang He, and Guowei Chen. 2023. Virtual human talking-head generation. In *Proceedings of the 2023 2nd Asia Conference on Algorithms, Computing and Machine Learning*, CACML '23, page 1–5, New York, NY, USA. Association for Computing Machinery.

Karina Vida, Judith Simon, and Anne Lauscher. 2023. Values, ethics, morals? on the use of moral concepts in NLP research. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from language models.

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*, 4(2):100211.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Ruiping Zhen, Wei Song, Qiqi He, Jian Cao, Liang Shi, and Junying Luo. 2023. Human-computer interaction system of talking-head generation. *Encyclopedia*.

A EthiCon Dataset

Creation We created a dataset consisting of 1,580 ethical statements extracted from the ACL Anthology 2022 and 2023. We used the URL links provided by Rohatgi et al. (2023) and retrieved 4,691 articles from the ACL 2023 Anthology and 3,357 articles from the ACL 2022. Those links include publications from mostly 5 conferences: ACL, EMNLP, NAACL, ACL-IJCNLP, and EACL. To extract the ethics statements we parsed the HTML page of the URLS and crafted a regular expression pattern to catch common variations of the paragraph title, such as 'Ethic(s)', or 'Ethical' followed by terms like 'Statement', 'Consideration(s)' or 'Concern(s)'. The pattern is flexible enough to accommodate different document structures, allowing for optional numbering, whitespace, and

variations in terminology. It captures the text content within the ethics paragraph, ensuring it stops matching when encountering subsequent sections like acknowledgments, references, or limitations. We validate the effectiveness of this extraction by our annotators who manually inspect every statement and verify whether this parsing was successful. In total we collected 1,580 ethical statements; 1,100 from the ACL 2023 links and 480 from the ACL 2022 links. Compared to the 90 statements extracted by Benotti and Blackburn (2022) there is a substantial increase in the number of researchers who add an ethical statement to their work.

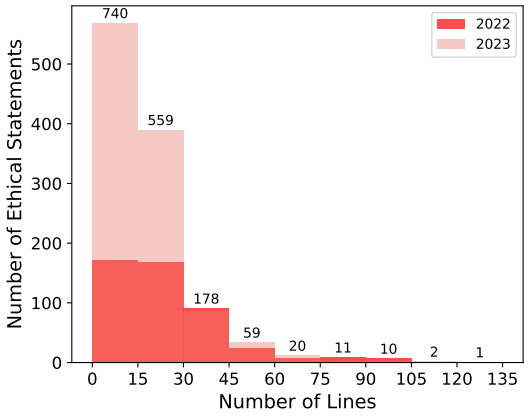

Figure 8: Histogram of the length (number of Lines) of ethical statements of the 1580 publications that include such section in ACL 2023.

Annotation To create the annotation guidelines, we manually inspected 500 ethics statements and labeled them with ethical concern labels based on their content (e.g. no concerns, data bias, privacy issues, etc). To ensure consistency, we tried to use the same words as the authors of the ethical statement paragraphs. For example, statements such as *'It is worth noting that the behavior of our downstream smaller models is subject to biases inherited from the larger teacher LLM.'* were labeled with *'model bias'*. During this first manual inspection, we also noticed that the statements can be divided into five classes: (1) general *disclaimers*, (2) a list of ethical *concerns*, (3) a list of *actions* taken to avoid ethical concerns, (4) a list of *suggestions* or advice to avoid ethical concerns, and (5) *miscellaneous* (other), i.e., various combinations of the classes above. This categorization was later validated since it overlaps with some suggested classes by previous work(Benotti and Blackburn, 2022). See an example of each class in Figure 2.

For the annotation, we had two volunteer researchers, one in NLP and one in Anthropology both fluent in English and with a Western cultural background. We also included a third annotator, NLP researcher, as a validator of the annotations and arbiter in cases of disagreement. Overall we reached a 0.77% Cohen's κ inter-annotator agreement which is considered substantial (Cohen, 1960). We conducted 3 preparatory meetings to explain the goal of the project, the annotation guidelines, and some examples. The instructions were to flag the statement with ethical concerns based mostly on the authors' choice of words. The duration of the process mentioned above was approximately 4 months. We did not use an annotation tool, but simply provided the guidelines and a Google sheet to each annotator with 5 columns: a *link* to the paper, a *statement paragraph*, an empty column for *annotation* where they would provide the comma-separated list of concerns, an empty column for grouping the statement into a *purpose category*, and a column where they could provide further *comments*. You can see the annotation guidelines in Figure 13. Please note that apart from the guidelines we had several group discussions and provided live examples with suggested and potential annotation through a Google Slides presentation.

Statistics. This annotation process resulted in extracting a list of 1317 ethical concern annotations across 1580 statements. Removing the duplicates and creating a set of these annotations results in 166 ethical concerns. Text metrics for ethics statements show longer ethical statements in the ACL 2022 anthology compared to 2023, including higher word counts, line counts, sentence lengths, and sentence counts as shown in Table 4. We provide the average length in lines of the 1580 ethical statements from our dataset in Figure 8.

Text Metric	**ACL 2023**	**ACL 2022**
Average Word Count	142.5	194.1
Average Line Count	16.9	22.7
Average Sentence Length (words)	23.6	24.5
Average Sentence Count	6	8.1

Table 4: Average text Metrics for Ethics Statements in ACL 2023 and ACL 2022 anthologies

B Survey Design

Our survey was designed after multiple rounds of discussions and the goal was to collect public opinions regarding ethical concerns in NLP. Our target audience was both professionals and regular users of NLP technologies. We first shared the survey with a small group of 20 people to collect further feedback, and after some rounds of revisions, we shared it with a wider audience by sharing it on social media platforms and mailing lists. Both the survey and the survey answers can be available upon request and can be used for academic purposes. Our survey starts with some general instructions and terminology explanation as shown in Table 14.

We first, ask demographic questions about gender and age and select their familiarity level with NLP technologies (expert, advanced, regular user, no user). The survey demographics in Figures 9 and 10 show that most participants are between 20-40 years old. We have a similar distribution between male and female participants and a small percentage of non-binary. We present participants' comfort level in Figure 12. Our sample consists of 58.9% regular users and 41,1% of professionals.

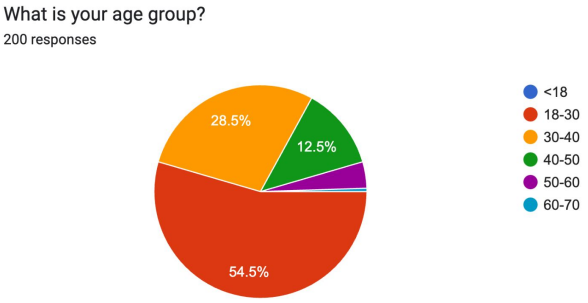

Figure 9: Age demographics from the survey.

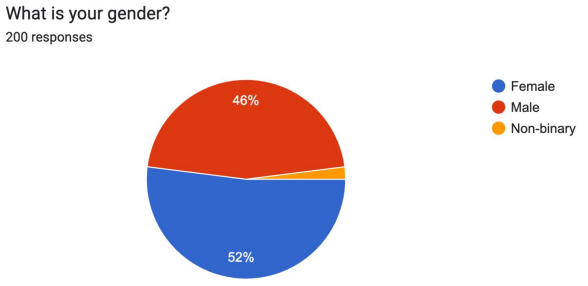

Figure 10: Gender demographics from the survey.

The second section of the survey asks participants to express their concerns regarding NLP technologies. We instruct them to write a list of things that worry them (separated by a comma,) or say none if they have no concerns.

Figure 11: Visualizing top 60 concerns from the human survey, reflecting term frequencies

In the third question, we ask participants to rate how worried they feel about the most frequent ethical concerns extracted from our EthiCon dataset. We also remind the definition of the NLP acronym since it is encountered in all questions. We also accompany each ethical concern with a short definition in parenthesis. For example, 'How worried are you about fairness issues (equitable treatment and outcomes for all users)?' or 'How worried are you about privacy issues (access or misuse of personal information)?'. Lastly, in the last section ask the participants again if they have any further concerns or comments they want to add and thank them for their participation. The reason we included this question was to allow the participants to add further concerns they may have, after rating the previously mentioned ethical concerns. The majority of participants (135 out of 200) did not have anything new to add to this last question. From the rest 65 of the participants who answered, only 2 of them mentioned issues related to the ethical concern rating section (one added *environmental impact* and the other one added a misuse concern, *People using it to take advantage of other people*).

Overall we got a variety of answers some of them were lists of words, others were phrases or sentences. for every answer, we manually extracted keywords aiming to keep the original wording of the participants. For n-gram words, we replaced space with an underscore. In total, we collected a list of 592 words-concerns, and removing the duplicates, resulted in a set of 189 concerns. We present a visualization of the 60 most common concerns in Figure 11. Please note that some of the smallest size concerns in the figure are only mentioned once. The most frequent ethical concerns are privacy, unemployment, bias, human replacement, misuse, and impersonation.

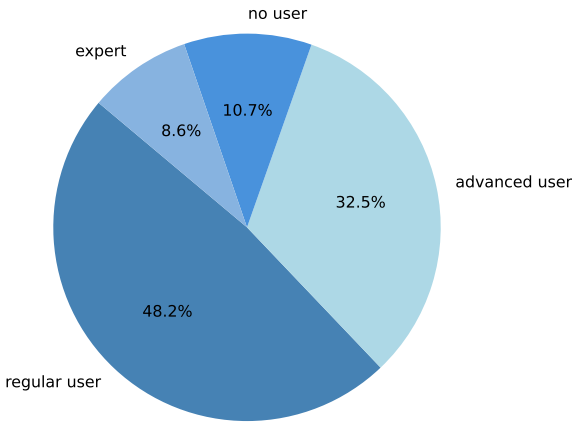

Figure 12: Participants' comfort level with Natural Language Processing technologies.

C Experiment Details

We used four state-of-the-art open-source language models to identify ethical concerns from our EthiCon Dataset: Gemma-7b-it, Meta-Llama-3-8B-Instruct, Qwen2-7B-Instruct, and Mixtral-8x7B-Instruct-v02. All models are available on Huggingface and we prompted them using the vllm library (Kwon et al., 2023) and a jinja2 format chat templates from https://github.com/jndiogo/LLM-chat-templates. We used 5 different prompt templates with different variations in phrasing. We provide an example from the Meta-Llama prompt in Table 5.

DEFAULT PROMPTING.

```
<|begin_of_text|><|start_header_id|>user<|end_header_id|>
Provide a list of words or phrases that characterize the ethical concerns
↪ mentioned in the paragraph below. Only include words or phrases that
↪ are directly related to the ethical concerns explicitly mentioned in
↪ the paragraph. If there are no ethical concerns mentioned, just
↪ return an empty list.

{{ text }}

Format your response according to the following JSON schema:
{{ schema }}
<|eot_id|><|start_header_id|>assistant<|end_header_id|>
```

Table 5: Instruction templates used for prompting our models.

To evaluate the model performance we calculated BERTscore from the official GitHub repository[11]. This metric was introduced by Zhang* et al. (2020) and it uses contextual embeddings from models like BERT and Roberta to represent the reference and candidate text. It measures the similarity between these embeddings using cosine similarity. Precision and Recall are computed by matching each token in the candidate sentence to the most similar token in the reference sentence, and vice versa. Our codebase and dataset are available at: https://github.com/coastalcph/EthiCon.

We manually compared the predictions between the annotated concerns and the generated predictions. We observed that out of the 576 no ethical statement paragraph statements from ACL 2023, the generation models correctly identified the absence of ethical concerns on average in 83% of the cases. They tend however to generate more ethical concerns than the ones being identified by the annotators, keeping synonym terms like misrepresentation, bias, fairness, unfair predictions, inaccuracies, and misinformation. We believe that this behaviour can be prevented with further prompting instructions.

D Taxonomies

We chose to group our extracted ethical concerns based on the taxonomy provided by Weidinger et al. (2021) for the following reasons: (1) They explicitly provide a unified taxonomy to structure the landscape of potential ethics and social risks associated with language models. Most of the works mentioned in Table 2, did not intend to provide a taxonomy, but rather discuss issues in NLP. (2) They cover a broad range of risks, grouping together concerns such as the discriminatory or exclusionary content group, the information harms (privacy leaks and sensitive content) group, and and socioenvironmental impact group (similar only to (Bommasani et al., 2021)). (3) They introduce a new category of harms as a result of human-computer interaction, which also covers a wide range of the concerns from the survey responses (similar to the Trust and Relationships category of Dinan et al. (2021)).

A more recent work, extending the discussion by Weidinger et al. (2021) is Gabriel et al. (2024). While they do not offer a clear taxonomy, they provide a comprehensive discussion on AI value alignment, well-being, safety, and misuse. This is followed by an analysis of the relationship between AI assistants and users, addressing topics such as manipulation, persuasion, anthropomorphism, appropriate relationships, trust, and privacy. Finally, the paper explores the societal impacts, emphasizing cooperation, equity and access, misinformation,

[11] https://github.com/Tiiiger/bert_score

economic and environmental effects, and evaluation practices.

Future research may expand an ethics taxonomy to include some of the miscellaneous identified concerns and apply additional approaches such as case studies or interviews to gain better insights into public concerns. For the next steps to build upon this work, one could include refining the methods used to evaluate language models and developing mitigation tools fostering their responsible innovation.

Annotation Guidelines for Extracting Ethical Concerns

The purpose of this annotation task is to identify ethical concerns expressed in statements extracted from scientific publications in the Association of Computational Linguistics (ACL) anthology. The task requires reading each statement and tagging it with specific keywords or short phrases that represent the ethical concerns raised. These keywords should be nouns or short phrases that directly capture the core ethical issues. Please read the guidelines below to ensure consistency and accuracy when annotating the data:

1. Read the Ethical Statement Carefully: Examine the statement to understand the ethical issues it addresses. Focus on identifying ethical concerns related to research practices, methodologies, or the implications of the work.

2. Identify Ethical Concerns

- Select Appropriate Keywords: Keywords should be nouns or noun phrases that directly reflect the ethical concern(s). Example: If the author mentions there is a potential bias in the data, please simply annotate with the keyword *bias*.
- A statement may refer to more than one ethical dimension (e.g., bias and misuse), and assign multiple keywords separated by a comma.
- Do not use vague, general, or unrelated terms. Example: the keyword *issues in NLP technologies* is vague. Aim to identify what is the issue the authors highlight.
- If multiple concerns are present, you may assign more than one keyword. Example: For a statement discussing 'bias in data collection' and 'privacy concerns', you can use two keywords: bias and privacy.
- Do not repeat synonym words, only closely related terms as multiple keywords. For instance, if they refer to the same concern ('job loss' and 'unemployment'), use consistently one term across statements. If the terms are closely related ('computational resources', and 'carbon footprint') it is preferable to use the author's phrase per statement.
- Make sure the keyword is related to an ethical issue, not a technical topic or research domain unless that domain is directly tied to the ethical concern. - Valid: data privacy when the concern involves the protection of user data. - Invalid: data collection since it refers only to methodology without ethical implications.
- If the statement is ambiguous you can make a comment with an alternative annotation, and/or flag it as ambiguous and bring it up for discussion. If the authors mention general ethical concerns regarding NLP without specifying how they apply to their work, we still annotate the ethical concerns.

3. Select from the drop-down menu a category that could describe the purpose of the statement. Is it a list of concerns (concerns category)? Is it a list of actions the authors took to avoid concerns? (actions) Is it a list of suggestions for ethical and responsible usage (suggestions)? Is it just a disclaimer? (disclaimers). If the statement includes more than one of the following, select the corresponding combination from the drop-down menu.

4. Check your annotations before submission: Re-read the annotations to ensure no ethical concerns have been missed. Double-check that keywords are mistake-free and correspond to the author's ethical concern phrasing.

Figure 13: Annotation Guidelines

Survey Instructions

ETHICAL CONCERNS ABOUT NLP AND GENAI

This survey is designed to understand public ethical concerns regarding advances in the field of Natural Language Processing (NLP). NLP is a field of Artificial Intelligence focused on how computers can understand, interpret, and generate human language. The part of NLP that focuses on creating content, such as text, images, or music, is also called Generative AI(Gen AI).
Nowadays, there is growing usage of such technologies that can generate not only text and speech but also images. There is a big community from research and industry settings working and improving those technologies, and their results are published in scientific conferences such as the Association of Computational Linguistics (ACL). The proceedings of the conference are publicly available and you can access all the published peer-reviewed works at: https://aclanthology.org/events/acl-2023/
Your task for this survey is to:
(1) add any ethical concerns you might have about NLP/AI advances,
(2) determine how much you are worried about specific ethical concerns highlighted by NLP researchers.
By proceeding with this survey, you consent to participate and acknowledge that you have read and understood the above information. Your participation is entirely voluntary, and you may withdraw at any time without any consequences. This survey is completely anonymous, we only need you to answer some basic demographic questions.

Figure 14: Survey Instructions