Handling and Presenting Harmful Text

Leon Derczynski
IT University of Copenhagen
Denmark
ld@itu.dk

Hannah Rose Kirk
University of Oxford /
The Alan Turing Institute
United Kingdom
hannah.kirk@oii.ox.ac.uk

Abeba Birhane
Mozilla Foundation /
University College Dublin
Ireland
abeba@mozillafoundation.org

Bertie Vidgen
The Alan Turing Institute
United Kingdom
bvidgen@turing.ac.uk

Abstract

Textual data can pose a risk of serious harm. These harms can be categorised along three axes: (1) the harm type (e.g. misinformation, hate speech or racial stereotypes) (2) whether it is *elicited* as a feature of the research design from directly studying harmful content (e.g. training a hate speech classifier or auditing unfiltered large-scale datasets) versus *spuriously* invoked from working on unrelated problems (e.g. language generation or part of speech tagging) but with datasets that nonetheless contain harmful content, and (3) who it affects, from the humans (mis)represented in the data to those handling or labelling the data to readers and reviewers of publications produced from the data. It is an unsolved problem in NLP as to how textual harms should be handled, presented, and discussed; but, stopping work on content which poses a risk of harm is untenable. Accordingly, we provide practical advice and introduce HARMCHECK, a resource for reflecting on research into textual harms. We hope our work encourages ethical, responsible, and respectful research in the NLP community.

1 Introduction

Textual harms can arise through a multitude of complex channels. The first degree of complexity comes from *what* type of harm is contained in the content itself. Misinformation can spread a culture of distrust, contaminate the information landscape and polarise groups (Mihailidis and Viotty, 2017; Au et al., 2021). Hate speech and abusive language can pollute online communities, inflict long-lasting trauma on its victims, and desensitise bystanders (Waldron, 2012; Vidgen et al., 2019). Negative social stereotypes and misrepresentations of individuals or groups can perpetuate traditional power imbalances in society, lead to the unjust allocation of opportunities or resources and promote discrimination towards the 'undersampled majority' (Buolamwini, 2017; Blodgett et al., 2020).

The second axis of textual harms concerns *why* they appear during the research process. Some NLP research actively studies harmful phenomena in language, such as hate speech, extremism, misinformation, prejudice, or toxicity. In these areas, the risk comes from *elicited* harm, where exposure is a direct consequence of research design. However, work in seemingly unrelated NLP domains (e.g. NLG, part-of-speech tagging, or semantic search) may still encounter *spurious* harms in datasets, especially if these are large-scale and scraped from internet sources (Luccioni and Viviano, 2021; Dodge et al., 2021; Kreutzer et al., 2022).

Finally, harm can be categorised as to *who* it affects and *when*. Broadly, we focus on three broad groups of people: (1) *data subjects*, i.e., the humans and groups represented in the data who may suffer primary trauma from negative stereotypical representations, damaging associations or toxic language; (2) *data handlers and researchers*, i.e., those who curate, collect, annotate, or audit the data and those who code, analyse or write-up the results may face a risk of vicarious trauma from exposure to harmful material (Vidgen et al., 2019; Newton, 2020; Pyevich et al., 2003); and (3) *readers and reviewers*, i.e., those who consume publications written about the harmful content may be exposed to verbatim examples or screenshots.

Each of these axes presents concerning ethical and methodological challenges which need to be addressed for the field to advance in a responsible and equitable manner. If left unaddressed, a lack of adequate safeguarding could lead to harm perpetuating through the research process; academics

from less well-represented groups leaving or not entering the community; reduced readership of academic outputs; and reduced awareness of existing harmful content. Simply avoiding all research that poses a risk of harm is an untenable solution because at the same time, researchers have a duty to investigate textual harms contained in our datasets and appearing in the wider online landscape.

As a community of researchers and data practitioners, we need to communicate clearly about linguistic elements contained in data as a basic function of research. Often, this involves giving verbatim examples; synthetic examples which closely reflect the original content; or detailed descriptions of data. We need to study and discuss problematic data without inadvertently entrenching problematic ideas or propagating harms to data subjects, without distressing readers, with minimal emotional toll to the researcher, and without the risk of being misquoted or misconstrued as aligning with the views represented in the harmful content.

A new framework is needed to outline ethical obligations on researchers, and to find a common ground for continuing research on textual harms in a safe, responsible, and respectful manner. In this paper, we address the lacuna in existing NLP research practise by describing the harms and risks contained in text, who they affect, and what can be done about it. To ease the adoption of these recommendations, we present HARMCHECK, a checklist for transparent, responsible, and reflexive reporting of textual harms.

Any paper that discusses harmful content should have a clear content warning in the introduction or abstract at least a page before any examples are shown. The content warning should contain a brief description of the harms and distance the authors from any harmful examples. For this paper:

Content Warning: *This document discusses examples of harmful content (hate, abuse, misinformation and negative stereotypes). The authors do not support the use of harmful language, nor any of the harmful representations quoted below.*

2 Harms and Risks in NLP Data

Data is fundamental to nearly all areas of NLP but in recent years, more energy has been directed to *quantity* over *quality*. The advent of large-scale, pre-trained models has intensified the search for more data. To deal with the demands of deep learning, data curators and researchers have turned to enormous internet-scraped datasets such as Common Crawl Corpus or WebText. As these unstructured corpora become larger, the risk of them containing harmful content increases, and the larger the dataset, the more difficult it is for humans to explore what is in the dataset and audit for quality or toxicity (Hanna and Park, 2020; Luccioni and Viviano, 2021; Kreutzer et al., 2022).

The harms posed in a dataset itself can be echoed by models trained on it. The transfer of dataset harms to model harms is particularly pertinent with pre-trained, large-scale models because of two reasons. First, as Bommasani et al. (2021) argues, large language models have emergent capabilities which are difficult to fully understand. Second, these models are high-performing and widely-accessible on repositories such as HuggingFace so have been adopted by researchers and practitioners for a variety of downstream tasks. However, the original training data is rarely available and even if it were, an enormous amount of resources is required to train a model from scratch on alternative or augmented data. These factors make it more likely large-scale pre-trained models will be applied 'out-of-the-box' (Kirk et al., 2021b), that users of them do not audit the data before applying it, nor understand the risks contained in the pre-trained checkpoints. As such, *spurious* harm can be encountered both in the dataset (when audited) and in the model's behaviour and outputs.

The proclivity of large-scale models to inherit negative stereotypes and toxicity from the training data emboldens the need for audits and improved data *quality*. The movement towards data-centric AI prioritises data acquisition and diversity over model complexity (Paullada et al., 2021). A body of evidence shows how data optimisation can lead to substantial performance gains (Xu et al., 2021) and that dataset quality is critical for safer and more robust AI models (Sambasivan et al., 2021). While this move is welcome, it may mean NLP researchers increasingly spend more time working with textual data, qualitatively inspecting datasets, auditing their contents and reviewing labels.

We recognise a distinction between *elicited* and *spurious* harms. Some activities elicit harms: compiling a dataset of hate speech, for example, or auditing a dataset for stereotype propagation. In this case, an actor uses their agency in an attempt to discover and thus elicit harmful text. While on the one hand these *elicited* harms pose a greater

risk due to the increased density of harmful content, on the other hand, researchers, annotators and reviewers are aware of the risks a priori and can prepare (§3.1). However, the potential for encountering harms exists even when one is not looking for them, in part due to wider use of large-scale datasets and pre-trained models. This latent risk, which we term *spurious* harm, is concerning as people who interact with the data may not be aware of or prepared for it.

In the following sections, we first offer a general definition and discussion of harmful content, then describe *who* is at risk of harm.

2.1 What is Harmful Content?

By 'harmful content' we mean content that negatively impacts the emotional, psychological or physical well-being and safety of an individual, group or society of humans. What constitutes harmful is deeply predicated on historical and contemporaneous context, as well as who it comes from and is directed at. Harmful content is thus an open class, that is, we cannot enumerate all possible sources and types. Weidinger et al. (2021) summarised the NLP 'risk landscape', with a taxonomy of six risk areas: fairness and toxicity, privacy, false information, malicious use of NLP tools, interactions between humans and AI agents, and wider societal impacts on the environment and the economy. We focus on harm contained in NLP data such as disinformation (Derczynski et al., 2015), propaganda (Da San Martino et al., 2020), incendiary and manipulative messages, descriptions of harmful acts, hate speech (Vidgen and Derczynski, 2020), threats of violence, abuse, slurs, sexist, racist and otherwise marginalising and negative stereotypes (Birhane and Prabhu, 2021).

In addition to the variety of textual harms complicating our definition, the harms suffered by individuals from such content can vary; being both short- and long-term in effect, and affecting individuals both directly and indirectly. For instance, online hate can create severe mental health problems for victims (a *long-term* and *internalised* form of harm) but also cause a second-order harm on those handling or moderating the content (Pyevich et al., 2003; Dubberley et al., 2015; Spangenberg, 2022; Newton, 2020). Quantifying the degree of *internalised* harm to an individual is difficult because experiences of harm are intimately related to that individual's identity and lived experience. The same piece of content could affect individuals idiosyncratically: for example, a self-identifying woman studying a dataset of online misogyny may be more affected by its content than a equivalent man, especially if she has been personally targeted by similar attacks in her past. Furthermore, not every individual can observe the same harms due to societal positionality (§2.2).

Harm to society interacts with individual harms. In one direction, societal-level harms can deepen individual-level harms. Representational harms, for example, emerge from sexist, racist, ableist, and otherwise unjust historical, cultural and norms (Blodgett et al., 2020; Ahmed, 2007). These representational harms can lead directly or indirectly to allocational harms, where under-served groups face inequitable opportunity and access to resources, reflecting back a deep-rooted culture of injustice and discrimination onto individuals. In another direction, individual-level harms can amass into societal-level harms. Electoral disinformation can lead to individuals attending the wrong location to vote, disrupting the democratic process, while climate change or vaccine misinformation targeted at individuals can enforce a negative externality on wider society and its members.

2.2 Who Decides What is Harmful?

The designation of content as harmful has social and political (and, as we argue, methodological) implications. This is perhaps best exemplified through content moderation on social media, where the labelling of content as 'misinformation' or 'hateful' is routinely contested. In a study of perceptions of hate speech, Costello et al. (2019) show that men and political conservatives find hateful material less disturbing than women or liberals. Critical data scholars (Benjamin, 2019; D'ignazio and Klein, 2020; Birhane, 2021) contend that those at the receiving end of harm and injustice hold the epistemic privilege to define harm from their lived experience – while those occupying the most privileged position in society are poorly equipped to recognise it, a phenomena D'ignazio and Klein (2020) have termed as the *privilege hazard*. For example, given the problematic history of the term, it shouldn't be up to white folks to decide if a use of n*gga is offensive or not. The experiences of individuals and communities at the margins of society who often disproportionately face abuse, hate speech and marginalisation must control and shape

understanding of harm (Weidinger et al., 2021).

2.3 Who is at Risk of Harm from Text Data?

We identify three groups that can be harmed from text data:

Those represented in the data The humans represented in the dataset are at risk of harm both from (a) what it does contain and (b) what it omits. The first of these can be considered harms from 'hypervisibility' (Noble, 2013), where groups may be the subjects of false claims, bigotry, negative stereotypes and/or derogatory terms. The harm begins with such textual (mis)representations, that is individuals and groups are already harmed by the text before it becomes data, and they become data subjects. When the relevant text is subsumed into a dataset, that harm becomes "frozen in time" and perpetuated as far as that dataset is spread. The second risk of harm comes from 'erasure', where the lived experiences of entire groups and communities are omitted from the data and thus rendered invisible to NLP systems (Jo and Gebru, 2020). These two forms of harm – the inclusions and the omissions – can interact in pernicious ways, i.e., when certain groups are represented rarely and these representations are harmful portrayals.

Presenting harmful content in research publications without the necessary precautions and safeguards risks propagating the harm to data subjects. This is the case for misinformation where spreading known-harmful ideas and false claims without making the problems with them unavoidably evident can lead to ambiguity. Exposure to false headlines increases the chance of their false claims being accepted and normalised, even when the reader knows they are false (Pennycook et al., 2018). Research that presents negative stereotypes without cautions contextualisation and qualification also risks further entrenching the associations in the dataset, deepening the harm to the data subjects (Barlas et al., 2021). In particular, when researchers are from a different background to those that are subject to harms, there is a greater risk of treating the content as de-humanised data that can be studied from a relative distance in an abstract manner – rather than something that has direct implications for the subjects' representation, welfare, and safety. This dehumanisation-by-datafication increases the risk of a disrespectful or harmful representation of the data subjects (Leurs and Shepherd, 2017).

Those working with the data People who are exposed to harmful text at any stage during the research process are at risk of vicarious trauma. At the earliest stage, *dataset curators and creators*, i.e., the people who search for or collect dataset entries, may come into contact with harms, for example, using keyword searches on the Twitter API to find online hate or scraping a political sub-reddit and finding racist posts. After the data is collected, *data handlers and processors*, i.e., the people who write code or clean the data, may come into contact with harms. For example, with a dataset containing a high-proportion of abuse, simply using commands to view the data like `df.head(3)` can inadvertently expose the coder. Once the data has been processed, harms arise during analysis. For example, in unsupervised learning, topic labels are assigned by reading the most representative documents, or in supervised learning, entries are given labels and models may be interrogated with qualitative error analyses. *Data labellers* or *annotators* are at particular risk of harm, especially in situations where harmful language is the phenomena of study. From repeatedly viewing harmful content for extended periods of time, annotators of harmful content may face similar psychological risks and emotional toll to content moderators, such as post-traumatic stress disorder, secondary trauma, and burnout (Steiger et al., 2021).

During the write-up stage of research, there is a direct welfare risk to the authors, who discuss, summarise or directly quote examples of harmful content. There is also an indirect reputational risk, that their examples may be misconstrued as representations of author beliefs, through careless reading, ambiguous presentation, or being taken out of context. While misconstrual can occur maliciously, it can also happen accidentally, for example through unfortunate crops created in photographs of presentations, screenshots of papers, or even resting on a screen while re-arranging windows.

Those consuming research about the data People reviewing and reading papers or attending talks produced about or from the data, may be distressed by exposure to harmful examples. With poor justification for and presentation of harmful examples, reviewers can object to inclusion of such content and give lower reviews or even ask for desk rejects. Consider this paraphrased review com-

ment:[1]

> Ethical issue: Even though the authors added a trigger warning in the paper, it was still uncomfortable for me to see examples along the lines of *"I want to murder Muslims"* in this manuscript. Researchers should confine themselves to discussion of their novel methods; it's not relevant to include so many distracting and useless quotes.

On the one hand, this comment summarises the welfare risk to readers (as well as the harm and distress this hateful content might cause the Muslim community themselves) from presenting verbatim examples of harmful content. On the other hand, these examples of harmful statements may need to be pointed out to demonstrate the severity of problems that authors are seeking to solve.

3 Guidelines for Handling and Mitigating Textual Harms

In other areas of academic research, there are well-established practices for reducing the risk and severity of harm. People working in a chemistry lab are protected by safety protocols for working with materials that present hazards, such as poisonous gases or radioactive substances. Similarly, there should be protocols for reducing the risk of harm to those working with hazardous materials in language. This section outlines ways of reducing the risk of harm from NLP data. Researchers are encouraged to use these guidelines as they are often best placed to take steps to protect themselves, those represented in the data, and those who consume outputs about the data. We present our guidance in chronological order of the contact points which can arise during research and practice.

3.1 Mitigating Textual Harms From the Offset

Data never emerge in a social, cultural, historical and contextual vacuum – they embed and perpetuate deeply held social norms, historical injustices, and uneven power dynamics. NLP datasets, including the harmful content they contain, cannot be neatly disentangled from these factors. The harms in text, therefore, cannot be solved through technical fixes from individual researchers. Instead, they require acknowledging the systemic roots of harm, challenging unjust systems, envisioning alternative world views and eventually working towards making such visions a reality. However, although tackling harmful content in datasets requires broader systemic change, actions that challenge structural change are far from useless. There also exist various ways in which NLP practice might assuage harms and eventually contribute towards culture change, even if only in an incremental way.

The first opportunity for harm mitigation is during data curation and selection, in order to avoid harms to the data subjects becoming frozen in time in a dataset. Indeed, some unlabelled corpora used to train large language models are filtered to remove the most obvious forms of toxic language. Pre-filtering, while arguably *a priori* desirable, must be cautiously approached because it can itself censor and erase marginalised experiences (Dodge et al., 2021). For example, by crudely removing language that could be considered harmful (e.g. by removing *any* use of potentially reclaimed terms, such as n*gga), the language of entire communities can also be excluded. When harmful language is being studied, data curators typically have no choice but to include harmful content in their datasets. For example, it is difficult to study hate speech without a dataset that contains some instances of hate. However, despite this practical necessity, they still have an obligation to mitigate the propagation and entrenchment of harms through their work. Thus, in our guidelines for handling, presenting and publishing research, we refer primarily to handling and presenting *elicited* harms because this is where the risk of harm is greatest and most direct.

3.2 Handling Textual Harms

The process of studying harmful text – whether as a machine learning engineer, social scientist, data labeller, auditor or otherwise – creates a clear risk of harm from repeated exposure (Einwiller and Kim, 2020). Several practical steps can be taken during the research process to mitigate this harm.

Brief It is important that researchers understand the goal and social mission of research which involves toxic content, as well as the likely risks that will be encountered. Without having reviewed the dataset by hand, lead researchers should outline the likely harms that will be presented given prior experience. Research teams should avoid engaging in projects without any researcher who has some

[1] We have added a watermark to warn of the harmful statement quoted in this excerpt.

prior experience or without extensively reviewing prior research and critically examining the upcoming task. Ultimately, researchers should have a realistic understanding of the likely risks they are facing *before* starting work.

Check in There should be a direct channel of communication between all involved parties - senior researchers, research assistants and annotators. Feedback from the research team should be explicitly and frequently elicited during research. Feedback mechanisms should be available that are both anonymous and individual, giving opportunities for people with different preferences to provide meaningful updates on their experience of the work. Ensuring adequate feedback opportunities is the responsibility of the senior researchers on a team. Regular feedback can also aid the research process by creating multiple touchpoints between all parts of the research team. Note using crowd-sourced workers may limit the effectiveness of communication about the annotation process, a concern which researchers should consider when designing their annotation process and building annotation teams.

Limit The risk of harm faced by researchers can be minimised by reducing their exposure to content. For some researchers, such as annotators, this exposure is unavoidable – but can be minimised by using more efficient techniques for working with data. For instance, active learning and continuous learning minimise the total amount of data that is needed for a given project. In some fields, such as computer vision, techniques have been developed to enable researchers to carry out their work whilst minimising the risk of harm, such as greyscaling images (Das et al., 2020). Similar approaches for text could be considered, such as masking harmful words, although this is likely to constrain research and may not be a worthwhile tradeoff. For people involved in other parts of the research process, such as machine learning engineers, engagement with data can be substantially minimised by more effective data processing. For instance, harmful text can be replaced with dummy data whilst establishing coding pipelines – and the real data only merge back in once models need to be trained.

Support Mental health and psychological support services should be in place for all who come into contact with harmful text, and made accessible with as few barriers as possible. This is particularly important in contexts where there is social stigma associated with seeking help or where those working with data are concerned about how they may be perceived. This can help both address negative experiences when they occur, and build resilience within teams (Steiger et al., 2021). These endeavours should translate into practical tools and processes for providing support through different interventions. They should be varied, and fit the needs of the person at risk of harm. In some cases and when possible, research teams may need to consider paying for support. In-person counselling services should also be considered, when possible, perhaps through the host institution. At a minimum, there should be a space for people working with harmful text to talk about that text with other humans, even just anecdotes or venting (Marwick et al., 2016).

De-Brief At the end of the research process, senior researchers should explain to their team the impact of the work and comment any unique or unanticipated issues that were encountered. This process should be as 'horizontal' as possible, enabling all researchers to express their views and experiences in an open dialogue. The de-brief is a useful opportunity for researchers to identify lessons learnt, refine processes and take steps to mitigate the risk of harm in the future.

3.3 Presenting Textual Harms for Publication

When publicising research about harmful phenomena, authors need to take steps (1) to protect and respect those represented in the dataset, (2) to warn of harm and limit exposure to those reading the research, and (3) to distance their own opinions from the harmful views or examples being discussed. These aims can be achieved using a selection of five techniques: preview – distance – disclaim – replace – respect. These steps are inspired by journalistic practice (Politifact, 2014; The Annenberg Public Policy Center, 2012), where it is important to be precise when establishing and positioning narratives. While the best way of reducing harm is to not give examples of harmful content at all, precise exposition and argumentation of a method or motivating problem in research sometimes requires these examples. This section details measures that can be taken to reduce the negative impacts of those examples.

Preview Readers need to know what to expect. The relevant section of a paper, video, audio, or code should indicate the kind of harmful content

that is coming up. Authors should **preview** or signpost the upcoming content in a consistent fashion. For example, give a warning in a visually distinct style. Avoid placing harmful content on the first page or above the fold, so that the audience gets a chance to decide whether they want to see it. Some might like to give a 'content warning' ahead of potentially troubling content but trigger warnings can risk reinforcing harm (Bridgland and Takarangi, 2021). When an example is required, format examples all the same way, so there is a consistent theme in presentation and examples are thus more readily identifiable as distinct from core content.

- No harmful content on page 1 / above the fold
- Warning about the content at least a page ahead
- Place a warning sign over the content

Distance In the case of harmful content, it is important to clearly **distance** the research *on the data* from the viewpoints and material contained *in the data*. It must be absolutely clear to even a casual observer that an example of inciteful, biased, false or hateful content is not from the authors themselves. This can be achieved visually, by, for example, including a bold highlight by each problematic example or including a watermark that overlaps the example in the paper (see Figure 1). This would also make taking misrepresentative screenshots of a paper or presentation difficult, further protecting the authors. Harmful content examples can also be presented less strongly, perhaps with reduced contrast / opacity by using for example a grey font (Karunakaran and Ramakrishan, 2019) or blurred images (Das et al., 2020). An alternative is to replace some terms with placeholders, e.g. *"that [IDENTITY] is a [SLUR]"* or *"I hate [IDENTITY]"*, to convey syntax and some semantics but avoiding actual hate towards a specific target group (Röttger et al., 2021; Kirk et al., 2021a).

- Visual distancing through watermark overlay
- Format harmful examples consistently
- State that examples are examples
- State that harmful content is harmful
- Use minimal examples: crop, recolor, blur, truncate

Disclaim Clearly identify the content's origin and thereby **disclaim** it as an example. For example, political ads should be labelled as "political

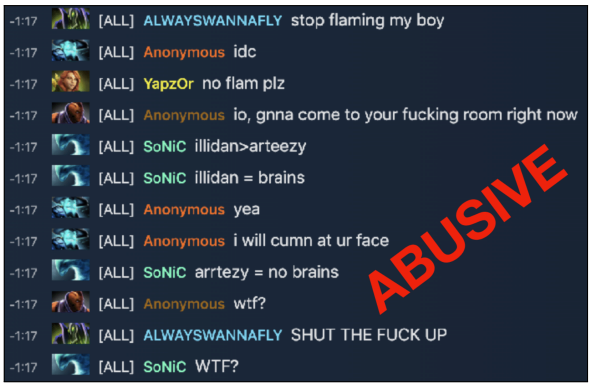

Figure 1: Example of chat history from an online game.

content" and a distinct visual style should be used. This can intersect with the content in order to make it clear that the example is not part of the body of a paper or a scientific graphic. If possible, the provenance of the content should be identified, provided this is in-line with privacy regulations and ethical concerns; it sometimes may be best to say platform of origin and date e.g. "From a Twitter user, November 2020". It may make sense to present a disclaimer that the harmful text does not represent authors' views. This should be close to the examples so that people who are focused on only a sub-part of a document (e.g. skim readers) are made aware of the distancing without reading everything that came beforehand.

- Place disclaimer overlapping the example
- Consider giving attribution to harmful content, stating its source, if this discredits it

Replace It's important to not let the sentiments expressed, or conclusions reached, in harmful content stand uncorrected – the harmful narrative should be **replaced**, where possible, with accurate messaging. For instance, examples of false claims should be explicitly flagged and accompanied by the relevant true claim. For abusive language, this could be a statement or re-statement of the nature of abuse in the example.

- Include a correction close to the content
- Wrap up sections discussing examples with a corrected claim
- Re-state the type of harm close to the content

Respect Present people targeted in harmful text with **respect**. Harmful content is disproportionately targeted on identities who are already marginalised, vulnerable or under-represented, including groups defined by their race, gender,

(dis)ability, sexuality, religion and more (Amnesty, 2018; Abid et al., 2021). This is particularly pertinent in relation to online hate, but is true for other harmful content (e.g. personal attacks and misinformation).Giving examples of harmful content that targets these groups (who we have referred to as "data subjects"), even with appropriate safeguards, risks repeating and propagating those harms. As such, it is critical that the groups represented in such content are treated with respect. Researchers should adopt active and continual reflexive practices, such as striving to adopt the perspective of the data subject and developing awareness of the social and historical roots of groups/concepts that are subject to harm. This is yet another reason why diversity in research teams, with multiple perspectives and positionalities represented, is crucial to raise awareness of, and mitigate against, the risk that harm is reproduced through research. The dignity and personal privacy of data subjects can be protected by removing identifying information (such as Twitter handles) or blurring any identifying images. With hate speech, when possible, the vowels in slurs, profanities or offensive terms should be starred out. This may not always be feasible, such as when showing screenshots of conversations or quotes which contain emoji, where the original content may need to be shown.

- Consider how examples reflect on the people harmful text is about
- Blur or star-out non-reclaimed uses of slurs
- Remove PII
- Blurring images of people / faces in multimodal work

3.4 Preparing for Releasing Textual Harm Research

Research related to harmful content comes with its own risks post-publication. Researchers working in these areas have faced attacks, both online and offline (Marwick et al., 2016). Harassers have used many attack vectors (Vogels, 2021), and researchers have been subjected to online abuse, death threats, deepfake revenge porn, doxxing (finding and publishing personal information) and even swatting (in the USA, having an armed unit storm the researcher's house with guns) (Mortensen, 2018; Greyson et al., 2018). For example, one paper published at NeurIPS focusing on gender bias (Kirk et al., 2021b) prompted a wave of misogynistic attacks against the lead author on Twitter. Another paper published at ACL 2021 studying online misogyny (Zeinert et al., 2021) prompted: large amounts of online abuse and doxxing directed at the authors by name; frivolous freedom of information requests explicitly for the purpose of wasting time; complaints made to the authors' external funding organisations; public attacks from politicians against the authors and their institution; and pejorative opinion articles in the national press against the research. Researchers publishing research about hate speech, misinformation, bias or other forms of textual harms should be aware of and prepared for these kinds of interactions, even though this is far from the norm for most areas of academic research. Fortunately, concrete steps can be taken to reduce the risks to those who handle and present harmful content in the course of their NLP research.

Brief Give your organisation – and its press, communications and legal department – advance warning that you are publicising the research, and that it may bring some harassment. They should have procedures for handling this and protecting their members, though many can be unprepared (Ketchum, 2021). If there are no procedures in place, then guidance and policy templates are publicly available for researchers to initiate a dialogue with their organisation.[2]

Protect identity Would-be harassers use online search to find details about their targets or to identify routes to attack them (e.g. by sending them abusive social media messages). Consider editing, hiding, or removing online information about you that you would not want malicious parties to use (Glaser, 2020).

Get support There is great value in having someone to discuss harassment with. They do not have to be a collaborator or even someone in the field. Let them know that the work might spark a backlash before it happens. It is OK just to vent (Steiger et al., 2021).

Curate outreach Talking to the press is rarely compulsory: not every media request has to be answered. Some discussions are likely to result in negative coverage. It is often worthwhile looking to see what the journalist and outlet in question has

[2]E.g. Data & Society's sheet on 'Online Harassment Information for Universities'

published before, so that you know that research is treated appropriately.

Our comments here are not intended to constrain academic and civic discussion about research – and it is certainly the case that some criticism of research outputs in risky areas will be legitimate. Indeed, proper documentation of research outputs (such as datasets, models and annotation frameworks) increases research transparency. Thus, the steps that we propose will not only mitigate the risk of harm but will also improve academic scrutiny and debate.

4 HARMCHECK: A checklist for handing and presenting harmful text

In recent years, there has been a growing movement towards the responsible and transparent documentation of research artefacts (Bender and Friedman, 2018; Mitchell et al., 2019; Rogers et al., 2021). Some conferences now require that authors fill in a responsible NLP checklist to accompany their submission.[3] In a similar vein, we wish to encourage a standardised and transparent discussion for the risk of harms contained in a research output. We thus present HARMCHECK, a simple checklist drawing on our above advice, which works as a standalone piece of documentation or could be appended onto existing documentation standard and filled in by people specifically researching textual harms. We encourage reflexivity and transparency and each section is intended to be filled in as a statement (such as in a data statement (Bender and Friedman, 2018) or model card (Mitchell et al., 2019)), with some sections being more relevant for different harms types (e.g. toxicity, hate speech or misinformation). To guide researchers, we provide a list of starter questions for each section.

4.1 Proposed Checklist

1. **Risk of Harm Protocol**: Summarise the steps taken during the research progress to identify and mitigate harm to at-risk groups.
    - What are the specific risks of harm and to who? Have you explained how the well-being of any researchers, annotators or data processors was protected during the study period?
2. **Preview**: Summarise any warning of harmful content and presentation of examples.
    - Is there a content warning at least a page before any harmful text instances are presented? Is the content warning clearly visible? Do section, table or figure specific content warnings describe the nature of the harm? Are harmful examples visually distinct and consistent?
3. **Distance**: Summarise distancing statements and views of authors.
    - Is it clear that harmful text is not part of the material's body? Is there visual distinction of harmful examples with a watermark or text color? Are harmful examples given reduced prominence relative to the containing document? Is only the shortest or most relevant part of the harmful text included?
4. **Disclaim**: Summarise documentation of sources and origins of harmful content.
    - Is the origin of the harmful text clearly identified? Are the claims of harmful text explicitly disclaimed?
5. **Replace**: Summarise any corrections, displacements or counter-claims to harmful content.
    - Are rebuttals placed near to harmful text? Are toxic statements, false claims or stereotypes rebutted?
6. **Respect**: Summarise any steps taken to protect the dignity and personal privacy of data subjects.
    - Has personally-identifying information, images or text been removed? Have harmful words, slurs and profanities been starred out?

5 Conclusion

Harms are present in text whether one is looking for them or not, and they can have strong negative impacts on members of many different groups. Some professional areas have established protocols for dealing with inherent harms – we describe these harms and provide these practices in the context of natural language processing research. NLP datasets encode information about the state of world from linguistic traces, predominately from online sources. But the statistical associations in language data are themselves reflective of much larger problematic societal structures and historical injustices. We do not suggest that the NLP community can alone bear the weight of responsibility

[3] https://github.com/acl-org/responsibleNLPresearch

for countering the deep-rooted historical, cultural and societal issues in-grained in language data – this remains an unsolved problem which requires systemic change from multidisciplinary perspectives. However, although systemic changes can't happen overnight, we, as NLP researchers, can still envision the kind of world we want our datasets, research, models, and tools to portray. As part of research, we need to investigate, audit and be transparent about harmful language, but where there is no guidance for doing this safely, we are at risk of not only harming members of our field – with everything from PTSD to publication difficulty to external aggression – but also those outside it.

Thus, a secondary unresolved problem in NLP is how the production, sharing, and consumption of research itself can be handled and presented in a way which limits secondary harm to wider general society and its members. The advice and points of reflection in this paper identify this problem and provide practical solutions. These recommendations are the start of a larger conversation about risks and harms in and from our field, and we hope that delineating and describing them opens an broad dialogue in the NLP community towards creating responsible, just and ethical research.

Acknowledgments

Leon Derczynski was partially supported by the Independent Danish Research Fund under project Verif-AI. Hannah Rose Kirk was supported by the UK Economic and Social Research Council grant ES/P000649/1. Bertie Vidgen was supported by Towards Turing 2.0 under the EPSRC Grant EP/W037211/1 and The Alan Turing Institute.

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Large language models associate Muslims with violence. *Nature Machine Intelligence*, 3(6):461–463.

Sara Ahmed. 2007. A phenomenology of whiteness. *Feminist theory*, 8(2):149–168.

International Amnesty. 2018. TROLL PATROL FINDINGS: Using Crowdsourcing, Data Science & Machine Learning to Measure Violence and Abuse against Women on Twitter. (Accessed on 26/04/2022).

Cheuk Hang Au, Kevin KW Ho, and Dickson KW Chiu. 2021. The role of online misinformation and fake news in ideological polarization: barriers, catalysts, and implications. *Information Systems Frontiers*, pages 1–24.

Pinar Barlas, Kyriakos Kyriakou, Olivia Guest, Styliani Kleanthous, and Jahna Otterbacher. 2021. To" see" is to stereotype: Image tagging algorithms, gender recognition, and the accuracy-fairness trade-off. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3):1–31.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Ruha Benjamin. 2019. Race After Technology: Abolitionist Tools for the New Jim Code. *Social forces*.

Abeba Birhane. 2021. Algorithmic injustice: A relational ethics approach. *Patterns*, 2(2):100205.

Abeba Birhane and Vinay Uday Prabhu. 2021. Large image datasets: A pyrrhic win for computer vision? In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1536–1546. IEEE.

Su Lin Blodgett, Solon Barocas, Hal Daumé, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. *arXiv preprint arXiv:2005.14050*.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Victoria ME Bridgland and Melanie KT Takarangi. 2021. Danger! Negative memories ahead: the effect of warnings on reactions to and recall of negative memories. *Memory*, 29(3):319–329.

Joy Buolamwini. 2017. Limited Vision: The Undersampled Majority.

Matthew Costello, James Hawdon, Colin Bernatzky, and Kelly Mendes. 2019. Social group identity and perceptions of online hate. *Sociological inquiry*, 89(3):427–452.

Giovanni Da San Martino, Alberto Barrón-Cedeno, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the fourteenth workshop on semantic evaluation*, pages 1377–1414.

Anubrata Das, Brandon Dang, and Matthew Lease. 2020. Fast, accurate, and healthier: Interactive blurring helps moderators reduce exposure to harmful content. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 33–42.

Leon Derczynski, Kalina Bontcheva, Michal Lukasik, Thierry Declerck, Arno Scharl, Georgi Georgiev, Petya Osenova, Toms Pariente Lobo, Anna Kolliakou, Robert Stewart, et al. 2015. Pheme: Computing veracity—the fourth challenge of big social data. In *Proceedings of the Extended Semantic Web Conference EU Project Networking session (ESCW-PN)*.

Catherine D'ignazio and Lauren F Klein. 2020. *Data feminism*. MIT press.

Jesse Dodge, Maarten Sap, Ana Marasovic, William Agnew, Gabriel Ilharco, Dirk Groeneveld, and Matt Gardner. 2021. Documenting the English Colossal Clean Crawled Corpus. *arXiv preprint arXiv:2104.08758*.

Sam Dubberley, Elizabeth Griffin, and Haluk Mert Bal. 2015. Making secondary trauma a primary issue: A study of eyewitness media and vicarious trauma on the digital frontline. *Eyewitness Media Hub*.

Sabine A Einwiller and Sora Kim. 2020. How online content providers moderate user-generated content to prevent harmful online communication: An analysis of policies and their implementation. *Policy & Internet*, 12(2):184–206.

April Glaser. 2020. 13 security tips for journalists covering hate online. *The Journalist's Resource*. Https://journalistsresource.org/media/13-security-tips-journalists-hate-online/.

Devon Greyson, Nicole Cooke, Amelia Gibson, and Heidi Julien. 2018. Online targeting of researchers/academics: Ethical obligations and best practices. *Proceedings of the Association for Information Science and Technology*, 55(1):684–687.

Alex Hanna and Tina M Park. 2020. Against scale: Provocations and resistances to scale thinking. *arXiv preprint arXiv:2010.08850*.

Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 306–316.

Sowmya Karunakaran and Rashmi Ramakrishan. 2019. Testing stylistic interventions to reduce emotional impact of content moderation workers. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 50–58.

Alex Ketchum. 2021. Report on the State of Resources Provided to Support Scholars Against Harassment, Trolling, and Doxxing While Doing Public Media Work.

Hannah Rose Kirk, Bertram Vidgen, Paul Röttger, Tristan Thrush, and Scott A Hale. 2021a. Hatemoji: A test suite and adversarially-generated dataset for benchmarking and detecting emoji-based hate. *arXiv preprint arXiv:2108.05921*.

Hannah Rose Kirk, Filippo Volpin, Haider Iqbal, Elias Benussi, Frederic Dreyer, Aleksandar Shtedritski, Yuki Asano, et al. 2021b. Bias Out-of-the-Box: An Empirical Analysis of Intersectional Occupational Biases in Popular Generative Language Models. *Advances in Neural Information Processing Systems*, 34.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, et al. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Koen Leurs and Tamara Shepherd. 2017. Datafication & Discrimination. *The Datafied Society*, 211.

Alexandra Sasha Luccioni and Joseph D Viviano. 2021. What's in the Box? A Preliminary Analysis of Undesirable Content in the Common Crawl Corpus. In *Proc. ACL*.

A. Marwick, L. Blackwell, and K. Lo. 2016. Best Practices for Conducting Risky Research and Protecting Yourself from Online Harassment (Data & Society Guide). Technical report, New York: Data & Society Research Institute.

Paul Mihailidis and Samantha Viotty. 2017. Spreadable spectacle in digital culture: Civic expression, fake news, and the role of media literacies in "post-fact" society. *American behavioral scientist*, 61(4):441–454.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229.

Torill Elvira Mortensen. 2018. Anger, fear, and games: The long event of# GamerGate. *Games and Culture*, 13(8):787–806.

Casey Newton. 2020. Facebook will pay $52 million in settlement with moderators who developed PTSD on the job. The Verge.

Safiya Umoja Noble. 2013. Google search: Hypervisibility as a means of rendering black women and girls invisible. *InVisible Culture*, (19).

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336.

Gordon Pennycook, Tyrone D Cannon, and David G Rand. 2018. Prior exposure increases perceived accuracy of fake news. *Journal of experimental psychology: general*, 147(12):1865.

Politifact. 2014. Fact checking: A studio workshop. Technical report, American Press Institute.

Caroline M Pyevich, Elana Newman, and Eric Daleiden. 2003. The relationship among cognitive schemas, job-related traumatic exposure, and posttraumatic stress disorder in journalists. *Journal of Traumatic Stress: Official Publication of the International Society for Traumatic Stress Studies*, 16(4):325–328.

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

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

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Jochen Spangenberg. 2022. How war videos on social media can trigger secondary trauma. https://www.dw.com/en/how-war-videos-on-social-media-can-trigger-secondary-trauma/a-61049292. Deutsche Welle.

Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J Riedl, and Matthew Lease. 2021. The psychological well-being of content moderators: the emotional labor of commercial moderation and avenues for improving support. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–14.

The Annenberg Public Policy Center. 2012. A Guide to Effective Fact Checking On-air and Online. Technical report, The University of Pennsylvania.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLoS One*, 15(12):e0243300.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*. Association for Computational Linguistics.

Emily A Vogels. 2021. The state of online harassment. *Pew Research Center*, 13.

Jeremy Waldron. 2012. The harm in hate speech. In *The Harm in Hate Speech*. Harvard University Press.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from Language Models. *arXiv preprint arXiv:2112.04359*.

Liang Xu, Jiacheng Liu, Xiang Pan, Xiaojing Lu, and Xiaofeng Hou. 2021. DataCLUE: A Benchmark Suite for Data-centric NLP. *arXiv preprint arXiv:2111.08647*.

Philine Zeinert, Nanna Inie, and Leon Derczynski. 2021. Annotating online misogyny. 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 3181–3197.