'Just What do You Think You're Doing, Dave?'[*]
A Checklist for Responsible Data Use in NLP

Anna Rogers
Center for Social Data Science
University of Copenhagen
Denmark
arogers@sodas.ku.dk

Timothy Baldwin
CIS
The University of Melbourne
Australia
tb@ldwin.net

Kobi Leins
King's College / London
via Australia
kobi.leins@kcl.ac.uk

Abstract

A key part of the NLP ethics movement is responsible use of data, but exactly what that means or how it can be best achieved remain unclear. This position paper discusses the core legal and ethical principles for collection and sharing of textual data, and the tensions between them. We propose a potential checklist for responsible data (re-)use that could both standardise the peer review of conference submissions, as well as enable a more in-depth view of published research across the community. Our proposal aims to contribute to the development of a consistent standard for data (re-)use, embraced across NLP conferences.

1 Introduction

In NLP, as in other research areas that are heavily data-driven, it is impossible to overstate the significance of the creation, collection, storage, and use of data (Paullada et al., 2020). Machine learning models are fundamentally a way of storing (and ideally generalising over) training data in the form of latent representations. The vast majority of NLP tasks involve supervised training of some sort. Ideally, the training data would come from the real-world, be of high quality, diverse, available in large quantities, obtained with revocable consent for specific use cases, and in compliance with licensing and other legal obligations, as well as broader ethical considerations.

Unfortunately, many of these considerations are often overlooked, to the potential detriment of both quality and integrity of the resulting datasets and models. A key issue is that the legal frameworks vary from jurisdiction to jurisdiction, as do the workplace cultures and norms around research institutions.We may not even have established definitions for commonly used terms like "fairness"

(Mulligan et al., 2019; Xiang and Raji, 2019) and "bias" (Blodgett et al., 2020). The result is differing expectations and standards, which makes it harder for fellow researchers to understand the terms of (re-)use of new resources and models, and also complicates peer review.

Experts in biomedical data research have concluded that "truly informed consent ... requires (1) comprehensive disclosure of informational risks to participants, (2) independent governance entities, and (3) data sharing policies that offer guidance for physicians and researchers" (Mittelstadt, 2019). The AI community has already started to work on the latter (Mantelero, 2018; Duan et al., 2019; Sielemann et al., 2020; Asilomar, 2017). Unfortunately, the codes of conduct and promulgation of ethics by industry, government, and academia remain highly contextual, and have "substantive divergence in relation to how these principles are interpreted, why they are deemed important, what issue, domain or actors they pertain to, and how they should be implemented" (Jobin et al., 2019). Further, a more cynical, albeit well-founded view, is that much of the flurry of activity around codes of conduct and ethics has been to avoid regulation, and is heavily driven by industry (Metcalf et al., 2019).

In NLP, the Association for Computational Linguistics (ACL) has adopted the ACM code of ethics (Gotterbarn et al., 2018), conducted a series of workshops (Hovy et al., 2017; Alfano et al., 2018), and, most recently, implemented ethics reviews and submission guidelines.[1] But we are far from a set of rules that would account for all the diverse types of NLP data and applications. Initial discussions after the first round of ethics review in NAACL 2021 highlighted many disagreements.

We submit that the community needs to work

---

[*] A line from the Stanley Kubrick film *2001: A Space Odyssey*.

[1]The NAACL guidelines (2021.naacl.org/ethics/faq) were subsequently adapted by ACL-IJCNLP and EMNLP 2021.

towards a common standard for what constitutes responsible use of data, what review mechanisms should be in place to ensure it, and what should happen when that standard is not upheld. Ideally, the discussion would advance collectively and with representation across different sub-populations of the field, as well as other stakeholders. Hopefully this process would result in more comprehensive disclosure, independent governance structures, and clearer policies that reflect a common purpose by the NLP community.

To this end, we analyse the core legal and ethical principles of data collection, and distill best-practice recommendations into a (non-exhaustive) checklist for potential adoption/adaptation by NLP conferences. Above all, we hope that this paper will advance discussion towards a single standard of responsible data (re-)use in NLP.

2 Key Principles of Data Collection

We do not yet have unified standards for responsible data use, but there are several guiding principles upon which we widely agree. This section briefly discusses such principles and the interactions between them.

2.1 Copyright and Terms of Use

In general, any data that is not in the public domain is considered to be the property of the individual or entities that produced it, and explicit permission is needed to collect/distribute it. It is incumbent upon researchers to ensure that they do not violate the law in undertaking their work. But even this law is not straightforward, because different jurisdictions have different copyright requirements. In 2021, e.g, US-based researchers cannot publish resources based on Hemingway's works, as copyright expires 70 years after the creator's death, but Japan-based researchers can: for them it is 50 years.

In addition to protection by copyright law, data is frequently protected by specific terms of services ("ToS") of the hosting platform, such as Yelp or Twitter. While copyright applies automatically, ToS have to be presented to users before they engage with the content (whether as a data generator or collector), and have to be explicitly agreed to. That said, they are often long, impenetrable, and many users (even researchers) give consent without reading them (Obar and Oeldorf-Hirsch, 2020). But once ToS are accepted, any violations may have potential

legal consequences. Most platforms have strict limitations on who can use what data for what purposes, and any use not in compliance with the ToS is probibited – with the complication that most such services operate internationally, with possibly different versions of ToS in different locations, and with different applicable international, national and regional laws. ToS may also change between the time of data collection and resource publication, and at this point there is no easy answer to the question of which version should apply.[2]

Different countries may also limit specific types of data processing. In France, for example, the automated analysis of legal decisions including personally identifiable data of judges or court clerks is now a crime, potentially carrying a prison sentence (Tashea, 2019).

2.2 Privacy

While privacy may seem like a fundamental right, it is not necessarily clearly defined in national legislation, complicating its interpretation and implementation by researchers. In the US, e.g., current legal provisions are based on a mix of constitutional amendments, federal and state laws, and Supreme Court decisions that are "patchy and in critical ways outdated" (Cate and Cate, 2012). For example, Facebook recently lost its appeal (Facebook vs Davis, 2020) in a case concerning surreptitious tracking of users logged out of Facebook, on the basis of two separate laws: the federal Wiretap Act and California state privacy laws. Most recently in Australia, Facebook lost another case where it was found that they can be held liable for defamation for anything that they post.[3] The landscape is a moving feast for lawyers, let alone for computer scientists and practitioners trying to comply with best practice.

The best-known legislation, setting the current standard for best practice globally regarding data, is the EU General Data Protection Regulation[4]

---

[2]Generally, law does not apply retroactively (Kryvoi and Matos, 2021). Given the bargaining power disparity between the platform and their users, whether the platforms are free to unilaterally change their terms with retroactive effect is a highly complex question concerning many fields of law, including unfair competition and consumer protection. At present, a legally accurate assessment for a particular case would have to start from determining the applicable local law (e.g. the law of Netherlands), and how it treats the particular type of contract (e.g. as service contract).

[3]`eresources.hcourt.gov.au/downloadPdf/2021/HCA/27`

[4]`eur-lex.europa.eu/legal-content/EN/`

Personal data shall be:

(a) processed lawfully, fairly and in a transparent manner in relation to the data subject (**'lawfulness, fairness and transparency'**);

(b) collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes; further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes (**'purpose limitation'**);

(c) adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed (**'data minimisation'**);

(d) accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay (**'accuracy'**);

(e) kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed [...] (**'storage limitation'**);

(f) processed in a manner that ensures appropriate security of the personal data, including protection against unauthorised or unlawful processing [...] (**'integrity and confidentiality'**).

Figure 1: GDPR principles relating to processing of personal data (Article 5)

(GDPR). GDPR governs data processing in all EU member states (and the UK, since it has adopted a copy of GDPR (National Archives, 2019)), as well as processing of personal data of any subjects of EU/UK, *irrespective of where in the world it takes place*. This makes GDPR applicable to most large-scale NLP resources based on web-crawled and social media data, since they are likely to contain at least some samples of data of EU/UK subjects.[5]

It is worth noting that the GDPR regime rests on the right to data protection spelt out in Article 8 of the EU Charter of Fundamental Rights,[6] and so

TXT/HTML/?uri=CELEX:32016R0679

[5]The degree to which extraterritorial laws (including GDPR) can be enforced is debated (Greze, 2019), but: (a) their violation does already pose a threat of reputational damage (Azzi, 2018); and (b) legislation in this sphere is quickly developing, and so it is possible that new mechanisms to enforce such laws e.g. through international cooperation may emerge. This might be the reason why, as of January 2020, an estimated 25% of Fortune 500 US-based retailers simply geo-blocked EU users (Bryan Cave Leighton Paisner, 2020).

[6]The Charter of Fundamental Rights of the European Union OJ C364/01 (2000) became legally binding as part

extends well beyond data privacy to human rights. We cannot do this even bigger issue justice within the scope of this paper, but some of the underlying principles are represented in the issues we discuss, and should frame thinking in governance of NLP.

Figure 1 shows the main GDPR principles that protect privacy. In a nutshell, people need to know and consent to how and why their personal data is being used. Researchers need to obtain consent for specific uses of data, not use it for anything else, not keep it for longer than necessary for the stated purpose, and prevent access to the data by anyone who has not obtained the same consent.

These limitations apply to anything that counts as "personal data", defined as "any information relating to an identified or identifiable natural person ('data subject')". People can be identified both directly (via name, ID numbers, online identifiers, IP addresses, or location data), or indirectly through "one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person".

The GDPR defines pseudo-anonymisation as "processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately" so as to prevent (re-)identification. Anonymisation techniques are an active research area, but at present "even coarse datasets provide little anonymity" (de Montjoye et al., 2013). There is also a fundamental tradeoff between data utility and the privacy it affords: any anonymisation technique necessarily removes information from data, which may significantly decrease its utility (Brickell and Shmatikov, 2008; Zang and Bolot, 2011).

Some argue that de-identified data needs to be better defined and regulated, as much data that is currently defined as de-identified is easily reidentifiable (Culnane and Leins, 2019). In NLP in particular, we have the fundamental problem that longer excerpts of text are almost impossible to anonymise (even in the limited sense of "masking entity references" (Mozes and Kleinberg, 2021)). De-identification may happen in two ways:

- *Direct identification:* the authorship of the text or message may be stripped from the

of the constitutional law of the EU when the Treaty of Lisbon entered into effect on 1 December 2009.

resource, but publicly available elsewhere (through the search function on a social media platform, or generic web search). Combined with metadata (such as that the utterance in question was used on Yelp), even relatively small amounts of texts may potentially be uniquely identifying (Shrestha et al., 2017; Kestemont et al., 2019).

- *Indirect identification* through authorship attribution: it may also be possible to identify the author of non-public texts based on other texts of theirs that *are* public. For example, anonymous social media users can be identified based on stylometry and typo patterns (Narayanan et al., 2012).

In some NLP applications such as search or voice assistants, we can mitigate potential re-anonymisation through differential privacy (Dwork et al., 2006), where the individual data points are deliberately superimposed with noise or include only relatively frequent data points using cardinality estimation (Harmouch and Naumann, 2017). But large language models can and do memorise the training texts (Carlini et al., 2020), and training them with privacy-preserving techniques without sacrificing too much quality is an active research area (Basu et al., 2021). If a language model reproduces a memorised excerpt, and the original author is easy to identify via web search, the author could be exposed to unwelcome attention (similarly to how an out-of-context quote can inflict serious reputational damage).

2.3 Transparency

Transparency "in relation to data subject" is the first GDPR principle cited in Figure 1: the people whose data is being used need to know exactly what was collected, and what it will be used for. It is transparency that prevents the collected data from simply being reused for other purposes to which the data subjects would not have consented.

Article 12 stipulates that the data subjects need to be informed through "a concise, transparent, intelligible and easily accessible form, using clear and plain language". In particular, according to Articles 12–14, the data subject must be aware of the identity and contact details of the controller, the purposes of the processing and its legal basis, the recipients of the data, and the period of storage. They must also be made aware of their rights, including the right to request rectification or erasure

of the data, request its copy, impose restrictions on its processing, or even withdraw consent.

Importantly for NLP, if the application involves any "automated decision-making, including profiling", the data subjects have to be provided with "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject" (Article 12).

2.4 Reproducibility

The reproducibility principle stems from general scientific methodology. It requires that researchers focus on the more robust observations, and guards against falsification and fabrication (as such results would not be reproducible). Most experiments and studies are performed once, under unique conditions, which means that it is difficult to guarantee that the results are valid and trustworthy. In NLP, to reproduce an experiment one would need both the implementation of a given system, and the data on which it would run.

The increased focus on *code availability* in NLP research (Wieling et al., 2018; Raff, 2019; Dodge et al., 2019; Crane, 2018, inter alia) leads to increased expectations for *data availability*.[7] Two recent taxonomies of reproducibility in NLP (Cohen et al., 2018; Tatman et al., 2018) both consider data availability a necessary precondition. This is completely fair when the data is e.g. synthetic or collected with informed consent for this use, but in many other cases the principle of reproducibility is inherently in tension with the other principles, as will be discussed in Section 3.

2.5 Do No Harm

The final principle comes from work on ethical AI and the social impact of NLP. While the NLP community ideally aims to improve the world, the real world is far more complex than the binary categories of 'good' and 'bad' (Green, 2019). Context and culture play a role, with no one-size-fits-all solution. Inclusive, representative groups on projects and reviews are part of the solution, but not enough on their own.

When we think of "do no harm", the first association may be harm inflicted by deployed systems or direct experiments on users (Kramer

---

[7] In addition to lower-level issues such as fixed splits of the data, the choice of evaluation metric(s), and potentially even pre-processing strategies.

et al., 2014; Flick, 2016), but there are also best-practice recommendations for data collection:

- *Document the population from whom the data comes.* This is necessary both for understanding the linguistic data, if the study goals are descriptive, and for ensuring a match with the target user demographics, if the data or models derived from it are meant for real-world use (Rogers, 2021). Much of NLP data is "convenience samples" of naturally-occurring data, which reflects a world riddled with inequalities and social biases. For instance, GPT-2 training data was scraped from links shared on Reddit, reflecting the worldview, language, and interests of predominantly young white men (Bender et al., 2021). Creation of 'perfect' and representative samples is not necessarily possible, but documenting the lacunae and omissions is (Bender and Friedman, 2018).

- *Consider the potential for exposure.* Social media storms are a force to be reckoned with, which can equally serve as an accountability instrument for public figures and companies (Rost et al., 2016; Neu et al., 2019), or a means of identity-based harrassment (Ortiz, 2020; Waisbord, 2020). Taking data from a moment in time and then 'baking it in' presents many problems, potentially leading to inaccurate or harmful representations of the individual (e.g. if they later retract a comment, or appeal a legal case and have the decision reversed).

- *Consider the potential for misuse.* The responsible thing to do before commencing a project is to think through what a bad actor could do with its results. *ACL conferences are increasingly requiring all submissions to consider possible misuse[8] and how the possible harms should be mitigated. In some cases it may not be safe to release even the annotation guidelines, much less the data (Rogers et al., 2019).

3 Tension between Data Collection Principles

While all of the above principles are important, there is a tug-of-war between them, as well as tensions with other factors in the research environment. The reproducibility principle aims to maximise data sharing in the interest of open science, while the others all limit it from ethical and legal perspectives. The transparency principle means that the data processors have to disclose what they are doing with the data, but that complicates the protection of trade secrets. The privacy principle is sometimes fundamentally at odds with public interest (crime, harassment, and accountability for public figures and organisations). Reproducibility and preservation of records of public interest may be in conflict with data subject privacy, the right to be forgotten, client confidentiality, and legitimate business objectives of industry research.

ToS of individual platforms may further cross-cut these principles: e.g. Twitter's requirement that only tweet IDs may be distributed maximises user control over their data, but it makes anonymisation impossible, and also sacrifices reproducibility (due to data attrition). Add to that the differences in national legislation, cultural norms, and priorities of individual researchers, and it is evident that there are no clear solutions that simply follow from all the above principles.

The result is that different research communities have come up with different norms and combinations of these principles, depending on their agenda, the power differential of data controllers and data subjects, and simply the goodwill of researchers willing to volunteer time to engage in unheralded administrative and advocacy work to change the norms.

As an example, consider the tension between privacy and open science in medical research (Minssen et al., 2020). Fundamentally, patient data has to be kept private, and most clinical studies do not publish it. This, however, prevents joint analysis of data from independent clinical trials, which could yield critical insights and literally save lives of other patients. One solution is adaptive clinical trial platforms (Angus et al., 2019), which enable cross-institutional coordination and data sharing in ongoing trials. But currently such initiatives rely largely on volunteer service work by researchers (Hickey and Chen, 2021). The result is patchy: for example, there is a platform for consolidating pancreatic cancer research across institutions,[9] but not for lung cancer.[10]

NLP is not an exception: as an interdisciplinary field, it has a mixture of players with different

---

[8]2021.naacl.org/ethics/faq

[9]www.pancan.org
[10]www.lung.org/research/clinical-trials

priorities and incentive structures, who argue for different combinations of the above principles, in part predicted by the domain they work in (e.g. clinical NLP with a strong focus on privacy and transparency vs. machine learning for NLP with a strong focus on reproducibility). More recently, as part of the rush to develop and distribute pre-trained language models, there has arguably been an over-focus on data volume and too little focus on any of the data principles, including the use of corpora of dubious legality and quality such as BookCorpus (Bandy and Vincent, 2021).

4 Responsible Data Use Checklist

The conflicting principles of data collection, coupled with conflicting priorities of different researchers, make for a torn, confused community. The movement towards standardising ethical and legal expectations has already started: NeurIPS 2020 pioneered obligatory broader impact statements inspired by Hecht et al. (2018), and NAACL, ACL-IJCNLP and EMNLP 2021 adopted a similar ethics policy, with a separate committee ethically reviewing papers flagged by regular reviewers. In parallel with all that, the *ACL conferences now use an adapted version of the reproducibility checklist (Dodge and Smith, 2020).

In this paper, we are hoping to refine the conversation regarding responsible data use, collection and distribution, as one thread of a richer dialog about NLP ethics. To this end, we propose to re-structure the host of issues that the paper authors are asked to consider into three areas:

- *experiment reproducibility*: provision of code, distribution of trained models/model outputs, specification of hyper-parameters, standardisation of resource usage, etc.
- *broader impacts of the NLP task/system*: could the research be (mis)used[11] to amplify existing social inequalities, support/undermine democratic processes, be weaponised, be retooled for propaganda purposes, increase the likelihood of armed conflicts? Are there potential conflicts of interest given the funding sources? Are there significant carbon costs? For more discussion

of these (and many other) issues see Ashurst et al. (2020), Hecht et al. (2018), and the NAACL 2021 Ethics FAQ
- *responsible data use/reuse*: whether the collection methodology is sufficiently described, complies with applicable regulations, and conditions for safe use/reuse are specified.

This paper focuses on the third area. Inspired by Dodge and Smith (2020), we make a first attempt at a checklist for assessing responsible data (re-)use, shown in Figure 2. Similarly to the reproducibility checklist, it is voluntarily filled in by the authors of an NLP study. Ideally it would accompany the papers as an appendix section *both pre- and post-publication*. This way it would not only facilitate peer review, but also help the regular readers of the paper to determine if *they* can use the proposed model or data, given the rules of their institution. If multiple datasets are presented or used, each one would have a separate checklist.

Our proposal draws on the GDPR framework, NAACL ethics guidelines, and work on documenting the populations represented in NLP resources (Bender and Friedman, 2018). While by no means a complete solution, we believe that it is a useful starting point.

The main section of the checklist focuses on the **safe use** of data, with many questions overlapping with those expected in institutional ethics review (see Section 5). It also has a section dedicated to **safe reuse**, which aims to nudge the resource authors towards specifying limitations and safe use cases. At the same time, the model developers are nudged towards specifying whether their use is consistent with any such limitations.

A key provision of safe resue is **ensuring consistent terms of data sharing**. In many cases researchers are allowed access to data that would not be granted to commercial entities, under the provision that this is done solely for research purposes, or for 'social good'. For example, based on Article 89 of GDPR and their local legislation, EU-based researchers may be exempted from observing some data subject rights "for scientific or historical research purposes or statistical purposes" and research performed "in public interest". If researchers gain access to data under such provisions, and then publish the data (or models derived from that data) under licenses allowing commercial use, this creates a loophole.

---

[11] A frequent counter-argument to consideration of broader impacts of NLP research is that anything could be misused, even "neutral" tools like parsers. While that is true, the ease, likelihood, and probable consequences of misuse do matter. NLP is in dire need of research into the cost–benefit of the use of its systems (Rogers, 2021).

<div style="border:1px solid">

For papers using a previously-published resource:

1. ☐ The authors explain their choice of data, given the available resources and their known limitations (e.g. representativeness issues, biases, annotation artifacts) and any data protection issues (e.g. inclusion of sensitive health data). *See Section ___.*

2. ☐ The authors discuss whether their use of a previously-published resource is compatible with its original purpose and license, and any known limitations (e.g. if the target user group is represented in the sample). *See Section ___.*

For papers contributing a new resource:

1. ☐ The authors have the **legal basis** for processing the data and, if it is made public, for distributing it. *(Check one)*

    1.1. ☐ The data is in public domain, and licensed for research purposes;
    1.2. ☐ The data is used with consent of its creators or copyright holders;
    1.3. ☐ If the data is used without consent, the paper makes the case to justify its legal basis (e.g. research performed in the public interest under GDPR). *See Section ___.*

2. ☐ The paper describes in detail the **full data collection protocol**, including collection, annotation, pre-processing, and filtering procedures. In the case that the dataset involves work by human subjects (e.g. data creation or annotation), the paper describes efforts to ensure fair compensation. *See Section ___.*

3. ☐ **Safe use** of data is ensured. *(Check all that apply)*

    3.1. ☐ The data does not include any protected information (e.g. sexual orientation or political views under GDPR), or a specified exception applies. *See Section ___.*
    3.2. ☐ The paper is accompanied by a data statement describing the basic demographic and geographic characteristics of the population that is the source of the language data, and the population that it is intended to represent. *See ___.*
    3.3. ☐ If applicable: the paper describes whether any characteristics of the human subjects were self-reported (preferably) or inferred (in what way), justifying the methodology and choice of description categories. *See Section ___.*
    3.4. ☐ The paper discusses the harms that may ensue from the limitations of the data collection methodology, especially concerning marginalized/vulnerable populations, and specifies the scope within which the data can be used safely. *See Section ___.*
    3.5. ☐ If any personal data is used: the paper specifies the standards applied for its storage and processing, and any anonymization efforts. *See Section ___.*
    3.6. ☐ If the individual speakers remain identifiable via search: the paper discusses possible harms from misuse of this data, and their mitigation. *See Section ___.*

4. ☐ If any data or models are made public: **safe reuse** is ensured. *(Check all that apply)*

    4.1. ☐ The data and/or pretrained models are released under a specified license that is compatible with the conditions under which access to data was granted (in particular, derivatives of data accessed for research purposes should *not* be deployed in the real world as anything other than a research prototype, especially commercially). *See ___.*
    4.2. ☐ The paper specifies the efforts to limit the potential use to circumstances in which the data/models could be used safely (such as an accompanying data/model statements). *See Section ___.*

5. ☐ The data collection protocol was **approved by the ethics review board** at the authors' institution, or such review is not applicable for specified reasons. *See Section ___.*

</div>

Figure 2: Responsible Data Use Checklist

While we argue that the community should aim for a common standard, our checklist of course does not fully specify it: what data is considered sensitive, what payment fair, what legal grounds acceptable for processing data without data subject consent? What it does achieve is forcing the authors to explicitly consider all these issues and present their motivation for the choices they made. If they went through a thorough review by a local board, they would already have done this work, and so workload should not increase.[12]

We readily acknowledge that the development of a shared norm for what is and is not responsible data use is already happening. That process is the reason why from time to time the choices made in a given paper provoke heated discussion, which

---

[12]The added bonus is that once the authors of a resource document their decisions this way, it will be easier for others to motivate their use of that resource, providing an indirect incentive for the dataset creators.

has impact on the planning of further projects by the community. We admit that even with the best checklists the result will likely never be perfect, as it is a reflection of our ever-changing field. The difference is that a standardised practice of showing the legal and ethical reasoning behind different projects in a structured way should accelerate the evolution of the shared norm, and decrease misunderstandings. Also, in the case of difficult legal disputes such as disagreements in ToS interpretation, a professional organisation such as ACL could consult legal professionals, and build a repository of common issues for the community to consult.

5 Can't We Just Defer to IRB?

Our proposed checklist has a separate checkbox for whether the project was reviewed by an ethics review board at the researchers' home institution. In the US it would be an Institutional Review Board (IRB).[13] Ideally, such a board would pre-check everything discussed in this paper, *before* the project is carried out. We absolutely encourage NLP researchers to use the legal and ethical guidance their institutions provide.

However, the primary goal of ethics review boards is *planning* of research projects. We argue that the ethical and legal thinking behind a project should be systematically made available for the community *post-publication*, so that future researchers can make better decisions about what they reuse, more easily document the limitations they inherit, and learn from precedents when developing the ethical and legal motivation for new projects. Reading papers in search of resources and models we can use, given our local rules and legislation, would be much easier if there were a standard way to share such information. In our proposal it is a checklist included in the appendix, with blank fields for specifying the sections of the paper where the given information can be found.

In the context of peer review, we also argue that a professional organisation such as ACL has a role in establishing a set of consistent *minimal* expectations across the NLP community (while the authors are also expected to comply with their national legislation and institutional rules). This is

necessary for the following reasons:

(1) There is a lot of variation in the researcher perception of when ethics review is needed (Shmueli et al., 2021). Santy et al. (2021) showed that less than 0.8% of NLP studies published since 2006 have sought IRB approval, and that was mostly for data collection or annotation. Approvals are rarely sought for data scraping or re-purposing, or for systems. Clearly, more than 1% of NLP papers come from US institutions and have something to do with data collection.

(2) There are major differences in both national legislation and institutional practices. In the worst-case scenario, without a minimal set of expectations, an institution could perform research disregarding privacy or human rights, and ACL would have no basis to block its publication.

(3) Some current review boards do not have the legal and ethical expertise to review legal compliance, and their evaluation may thus be less rigorous. The target scope of their scrutiny also varies, but the average reader or reviewer of the paper may not necessarily know what the review by committee type X entails. E.g. in the debate after NAACL 2021 business meeting several researchers expressed the belief that the ACL ethics review is superfulous because US IRBs cover all the target issues. In fact they are *discouraged* from considering long-term impact of research (Fleischman et al., 2011).

(4) The ethics review process is bound to produce cases where the authors and reviewers rely on different sets of rules, some more stringent than others[14]. The authors should generally be able to argue for their "home" rules, but only as long as they do not go below the minimal standard shared by the community – else we are in the situation (2) above.

6 Challenges

The above checklist is by no means exhaustive in terms of legal and ethical considerations in NLP projects, and does not solve all the problems. In particular, the following topics require a lot more legal thinking in the community.

Web crawled data NLP models are often pre-trained on large volumes of web crawled text: from

---

[13]The names and processes of such boards differ across countries and institutions: Comité de Protection des Personnes (France), Human Research Ethics Committee (Australia), Institutional Ethics Committee (India), Research Ethics Committee (UK).

[14]twitter.com/zehavoc/status/1430793222150840322

4828

6B words of news data in word2vec (Mikolov et al., 2013) to 42B words of Common Crawl data in GloVe (Pennington et al., 2014), and now 300B tokens in GPT3 (Brown et al., 2020), and counting (Fedus et al., 2021). These studies, like any other, have to have a legal basis for data collection, storage, and processing.

GDPR recognises "public interest" as a legal basis for processing of personal data (in this case, data attributable to individuals through public records of authorship). But most current large-scale models are trained by industry labs and made available for commercial use. The only possible basis for such training is the business' "legitimate interest", but then it needs to outweigh the interests of the data subjects, who may suffer significant harms (Bender et al., 2021) Furthermore, those involved in data storage and processing still have to comply with the privacy principles, including the protection of personal data. There is ongoing work in this direction in the Big Science Workshop.[15]

All the above only concerns privacy legislation, which copyright law is orthogonal to. Copyright would be directly violated by the act of copying the data for training, and also potentially by the fact of storing the protected text in trained models. With static word embeddings, it could be argued that the model itself does not "store" any text, but the current generation of language models clearly do reconstruct excerpts of training data. Since this is a blackbox process, it is hard to tell how much "creativity" there is in any particular sample, but the worst-case scenario is that such models could make their users potentially liable for plagiarism.

Legacy data Many datasets that are currently in use have been released with little consideration for copyright, speaker demographics, or proper handling of personal information (Bandy and Vincent, 2021). Since such resources have been published in prestigious venues, some NLP researchers might feel that there is precedent to use them and build new resources in a similar way.

The ongoing use of such data raises many questions. It is not realistic to prohibit its use overnight, but, if a responsible data-use standard is established, it should give rise to a new generation of resources that do not have such problems. In the short term, such a standard should also stimulate efforts to address documentation debt (Bandy and Vincent, 2021; Dodge et al.,

2021), as well as analysis and cleaning of the older resources (similarly to the problem of removal of pornographic images that the computer vision community is tackling[16]). Note that our checklist includes a section for previously-published resources, which should nudge authors towards more carefully documented data (since it would be easier to motivate).

Consent The key provision of use for data protected by both privacy and copyright legislation is that any processing can only happen with consent of the data subject/copyright owner. Fiesler and Proferes (2018) report that regular Twitter users are likely to not even be aware that their tweets can be used by researchers, and the willingness for their data to be analysed depends on the kind of study. GDPR offers some guidance on how data subjects need to be informed of what data was collected and what for, but there are other questions for which we do not have clear answers. In particular, should the participants be made aware of the many ways in which they can be identified in anonymised data (see Section 2.2)? What constitutes "public" data, and does the fact that someone willingly made some text public on their website or social media page constitute tacit permission to use it for any kind of research (or training commercial models)? What about the frequent situation of data that is originally released legitimately, but subsequently withdrawn? Is it permissible to use leaked data? If research is done without consent in the public interest, how can we assess the claims of its benefits vs. possible harms?

We hope that this paper stimulates further discussion of these and other issues.

7 Conclusion

As NLP technologies become more mainstream and commercialised, it is increasingly important that the NLP community leads best practice in legal and ethical standards. This paper is a step in that direction. We propose a tentative checklist for responsible data (re-)use, intended to serve as a complement to the existing reproducibility checklist. We hope that the community will continue working towards a common standard for data collection and sharing, complemented with transparent review mechanisms to ensure that the standard is upheld.

---

[15]bigscience.huggingface.co

[16]See www.losinglena.com, for example.

8 Acknowledgements

References

Mark Alfano, Dirk Hovy, Margaret Mitchell, and Michael Strube, editors. 2018. *Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing*. Association for Computational Linguistics, New Orleans, Louisiana, USA.

Derek C. Angus, Brian M. Alexander, Scott Berry, Meredith Buxton, Roger Lewis, Melissa Paoloni, Steven A. R. Webb, Steven Arnold, Anna Barker, Donald A. Berry, Marc J. M. Bonten, Mary Brophy, Christopher Butler, Timothy F. Cloughesy, Lennie P. G. Derde, Laura J. Esserman, Ryan Ferguson, Louis Fiore, Sarah C. Gaffey, J. Michael Gaziano, Kathy Giusti, Herman Goossens, Stephane Heritier, Bradley Hyman, Michael Krams, Kay Larholt, Lisa M. LaVange, Philip Lavori, Andrew W. Lo, Alex John London, Victoria Manax, Colin McArthur, Genevieve O&, apos, Neill, Giovanni Parmigiani, Jane Perlmutter, Elizabeth A. Petzold, Craig Ritchie, Kathryn M. Rowan, Christopher W. Seymour, Nathan I. Shapiro, Diane M. Simeone, Bradley Smith, Bradley Spellberg, Ariel Dora Stern, Lorenzo Trippa, Mark Trusheim, Kert Viele, Patrick Y. Wen, and Janet Woodcock. 2019. Adaptive platform trials: Definition, design, conduct and reporting considerations. *Nature Reviews Drug Discovery*, 18(10):797–808.

Carolyn Ashurst, Markus Anderljung, Karina Prunkl, Jan Leike, Yarin Gal, Toby Shelvane, and Allan Dafoe. 2020. A Guide to Writing the NeurIPS Impact Statement.

Asilomar. 2017. Asilomar AI Principles.

Adèle Azzi. 2018. The Challenges Faced by the Extraterritorial Scope of the General Data Protection Regulation. *JIPITEC*, 9(2).

Jack Bandy and Nicholas Vincent. 2021. Addressing "Documentation Debt" in Machine Learning Research: A Retrospective Datasheet for BookCorpus. *arXiv:2105.05241 [cs]*.

Priyam Basu, Tiasa Singha Roy, Rakshit Naidu, Zumrut Muftuoglu, Sahib Singh, and Fatemehsadat Mireshghallah. 2021. Benchmarking Differential Privacy and Federated Learning for BERT Models. *arXiv:2106.13973 [cs]*.

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.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *FAccT*.

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

Justin Brickell and Vitaly Shmatikov. 2008. The cost of privacy: Destruction of data-mining utility in anonymized data publishing. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 70–78, New York, NY, USA. Association for Computing Machinery.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]*.

Bryan Cave Leighton Paisner. 2020. What percentage of United States retailers have decided to block European visitors from reaching their websites? Technical report, Bryan Cave Leighton Paisner.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. 2020. Extracting Training Data from Large Language Models. *arXiv:2012.07805 [cs]*.

Fred H. Cate and Beth E. Cate. 2012. The Supreme Court and information privacy. *International Data Privacy Law*, 2(4):255–267.

K. Bretonnel Cohen, Jingbo Xia, Pierre Zweigenbaum, Tiffany Callahan, Orin Hargraves, Foster Goss, Nancy Ide, Aurelie Neveol, Cyril Grouin, and Lawrence Hunter. 2018. Three Dimensions of Reproducibility in Natural Language Processing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Matt Crane. 2018. Questionable Answers in Question Answering Research: Reproducibility and Variability of Published Results. *Transactions of the Association for Computational Linguistics*, 6:241–252.

Chris Culnane and Kobi Leins. 2019. Misconceptions in privacy protection and regulation. *Law in Context*.

Yves-Alexandre de Montjoye, César A. Hidalgo, Michel Verleysen, and Vincent D. Blondel. 2013. Unique in the Crowd: The privacy bounds of human mobility. *Scientific Reports*, 3(1):1376.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show Your Work: Improved Reporting of Experimental Results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194, Hong Kong, China. Association for Computational Linguistics.

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

Jesse Dodge and Noah A. Smith. 2020. Guest Post: Reproducibility at EMNLP 2020.

Yanqing Duan, John S. Edwards, and Yogesh K Dwivedi. 2019. Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda. *International Journal of Information Management*, 48:63–71.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In *Theory of Cryptography*, Lecture Notes in Computer Science, pages 265–284, Berlin, Heidelberg. Springer.

Facebook vs Davis. 2020. Facebook, Inc., Petitioner v. Perrin Aikens Davis, et al.

William Fedus, Barret Zoph, and Noam Shazeer. 2021. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *arXiv:2101.03961 [cs]*.

Casey Fiesler and Nicholas Proferes. 2018. "Participant" Perceptions of Twitter Research Ethics. *Social Media + Society*, 4(1):2056305118763366.

Alan Fleischman, Carol Levine, Lisa Eckenwiler, Christine Grady, Dale E. Hammerschmidt, and Jeremy Sugarman. 2011. Dealing With the Long-Term Social Implications of Research. *The American journal of bioethics : AJOB*, 11(5):5–9.

Catherine Flick. 2016. Informed consent and the Facebook emotional manipulation study. *Research Ethics*, 12(1):14–28.

Don Gotterbarn, Bo Brinkman, Catherine Flick, Michael S. Kirkpatrick, Keith Miller, Kate Varansky, and Marty J. Wolf. 2018. ACM Code of Ethics and Professional Conduct.

Ben Green. 2019. "Good" isn't good enough. In *AI for Social Good Workshop at NeurIPS 2019*.

Benjamin Greze. 2019. The extra-territorial enforcement of the GDPR: A genuine issue and the quest for alternatives. *International Data Privacy Law*, 9(2):109–128.

Hazar Harmouch and Felix Naumann. 2017. Cardinality estimation: An experimental survey. *Proceedings of the VLDB Endowment*, 11(4):499–512.

Brent Hecht, Lauren Wilcox, Jeffrey P. Bigham, Johannes Schöning, Ehsan Hoque, Jason Ernst, Yonatan Bisk, Luigi de Russis, Lana Yarosh, Bushra Anjum, Danish Contractor, and Wu Cathy. 2018. It's Time to Do Something: Mitigating the Negative Impacts of Computing Through a Change to the Peer Review Process. *ACM Future of Computing Blog*.

Matt Hickey and Daphne Chen. 2021. How to Fix the Incentives in Cancer Research (Ep. 449). Freakonomics.

Dirk Hovy, Shannon Spruit, Margaret Mitchell, Emily M. Bender, Michael Strube, and Hanna Wallach, editors. 2017. *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Association for Computational Linguistics, Valencia, Spain.

Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399.

Mike Kestemont, Efstathios Stamatatos, Enrique Manjavacas, Walter Daelemans, Martin Potthast, and Benno Stein. 2019. Overview of the cross-domain authorship attribution task at PAN 2019. In *CLEF (Working Notes)*.

Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790.

Yarik Kryvoi and Shaun Matos. 2021. Non-Retroactivity as a General Principle of Law. *Utrecht Law Review*, 17(1):46–58.

Alessandro Mantelero. 2018. AI and Big Data: A blueprint for a human rights, social and ethical impact assessment. *Computer Law & Security Review*, 34(4):754–772.

Jacob Metcalf, Emanuel Moss, and danah boyd. 2019. Owning Ethics: Corporate Logics, Silicon Valley, and the Institutionalization of Ethics. *Social Research: An International Quarterly*, 86(2):449–476.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of International Conference on Learning Representations (ICLR)*.

Timo Minssen, Neethu Rajam, and Marcel Bogers. 2020. Clinical trial data transparency and GDPR compliance: Implications for data sharing and open innovation. *Science and Public Policy*, 47(5):616–626.

Brent Mittelstadt. 2019. The Ethics of Biomedical 'Big Data' Analytics. *Philosophy & Technology*, 32(1):17–21.

Maximilian Mozes and Bennett Kleinberg. 2021. No Intruder, no Validity: Evaluation Criteria for Privacy-Preserving Text Anonymization. *arXiv:2103.09263 [cs]*.

Deirdre K. Mulligan, Joshua A. Kroll, Nitin Kohli, and Richmond Y. Wong. 2019. This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):119:1–119:36.

A. Narayanan, H. Paskov, N. Z. Gong, J. Bethencourt, E. Stefanov, E. C. R. Shin, and D. Song. 2012. On the Feasibility of Internet-Scale Author Identification. In *2012 IEEE Symposium on Security and Privacy*, pages 300–314.

The National Archives. 2019. The Data Protection, Privacy and Electronic Communications (Amendments etc) (EU Exit) Regulations 2019. Queen's Printer of Acts of Parliament.

Dean Neu, Greg Saxton, Abu Rahaman, and Jeffery Everett. 2019. Twitter and social accountability: Reactions to the Panama Papers. *Critical Perspectives on Accounting*, 61:38–53.

Jonathan A. Obar and Anne Oeldorf-Hirsch. 2020. The biggest lie on the Internet: Ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society*, 23(1):128–147.

Official Journal of the European Communities. 2000. Charter of Fundamental Rights of the European Union. Official Journal of the European Communities.

Stephanie M. Ortiz. 2020. Trolling as a Collective Form of Harassment: An Inductive Study of How Online Users Understand Trolling. *Social Media + Society*, 6(2):2056305120928512.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2020. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *arXiv:2012.05345 [cs]*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Edward Raff. 2019. A Step Toward Quantifying Independently Reproducible Machine Learning Research. In *NeurIPS*.

Anna Rogers. 2021. Changing the World by Changing the Data. 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 2182–2194, Online. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2019. Calls to Action on Social Media: Potential for Censorship and Social Impact. In *EMNLP-IJCNLP 2019 Second Workshop on Natural Language Processing for Internet Freedom*.

Katja Rost, Lea Stahel, and Bruno S. Frey. 2016. Digital Social Norm Enforcement: Online Firestorms in Social Media. *PLOS ONE*, 11(6):e0155923.

Sebastin Santy, Anku Rani, and Monojit Choudhury. 2021. Use of Formal Ethical Reviews in NLP Literature: Historical Trends and Current Practices. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4704–4710, Online. Association for Computational Linguistics.

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

Prasha Shrestha, Sebastian Sierra, Fabio González, Manuel Montes, Paolo Rosso, and Thamar Solorio. 2017. Convolutional neural networks for authorship attribution of short texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 669–674, Valencia, Spain. Association for Computational Linguistics.

Katharina Sielemann, Alenka Hafner, and Boas Pucker. 2020. The reuse of public datasets in the life sciences: Potential risks and rewards. *PeerJ*, 8(e9954).

Jason Tashea. 2019. France bans publishing of judicial analytics and prompts criminal penalty. *ABA Journal*.

Rachael Tatman, Jake VanderPlas, and Sohier Dane. 2018. A Practical Taxonomy of Reproducibility for Machine Learning Research. In *Reproducibility in Machine Learning Workshop (Co-Located with ICML)*.

Silvio Waisbord. 2020. Mob Censorship: Online Harassment of US Journalists in Times of Digital Hate and Populism. *Digital Journalism*, 8(8):1030–1046.

Martijn Wieling, Josine Rawee, and Gertjan van Noord. 2018. Reproducibility in Computational Linguistics: Are We Willing to Share? *Computational Linguistics*, 44(4):641–649.

Alice Xiang and Inioluwa Deborah Raji. 2019. On the Legal Compatibility of Fairness Definitions. In *Workshop on Human-Centric Machine Learning at the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*.

Hui Zang and Jean Bolot. 2011. Anonymization of location data does not work: A large-scale measurement study. In *Proceedings of the 17th Annual International Conference on Mobile Computing and Networking*, MobiCom '11, pages 145–156, New York, NY, USA. Association for Computing Machinery.