An Open Letter from the Mastodon Community

Administrators, Scholars and Users

January 20, 2020

We are writing to raise grave concerns regarding the ethics and methodology of "Mastodon Content Warnings: Inappropriate Contents in a Microblogging Platform", by Matteo Zignani et al. of the University of Milan. The issues with this paper are sufficiently severe that the paper's dataset has been removed from Harvard's Dataverse repository. This open letter will explain the background of this removal and urge further action on the part of the paper's authors, the University of Milan, and the Association for the Advancement of Artificial Intelligence (AAAI), who have published the paper in their conference proceedings. As we detail below, the data analysed in this paper was not collected ethically, failing to take even simple steps to anonymize the data released with the paper, and fundamental errors of methodology make its results irrelevant.

Mastodon is a decentralized, community-operated microblogging platform created in early 2016 by Eugen Rochko and is based on open protocols that allow people to communicate across different servers. Anyone who wishes to create a Mastodon server, or *instance*, can do so by downloading and installing the Mastodon software. Users who register accounts at an instance can then share social-media posts with other users on that instance as well as with other instances. The interconnection of different servers is known as *federation*.

1 Violation of Terms of Service

The authors state that they are aware that the Terms of Service and privacy policies vary from one Mastodon instance to the next: "Even though the distributed nature of Mastodon allows each instance adopts a specific terms of use and service, many instance [sic] are used to adopt the standard terms of service and privacy policy provided by the Mastodon developers." It is evident, however, that they did not respect the Terms of Service and privacy policies of all the instances scraped into their dataset. For example, the `scholar.social` instance documentation plainly states as follows:

> Researchers who wish to study Scholar Social or our users by collecting data using the API or through any other means that does not involve an "opt-in" from individual users are required to submit their protocol, analysis plan and all relevant ethics approval documentation from their institutional review board or departmental internal review documentation by email to: `scholar.social`

> at protonmail dot-com. You may also be required to submit code you plan to execute so that it can be tested to ensure it does not degrade the quality of service to other users.
>
> This does not apply to, e.g. surveys circulated on Scholar Social that individual users can consent to participating in. This does apply to database scraping software, or any means of recording user activity where our users might be surprised that they were included afterward, because they were not given a chance to consent.
>
> Many of our users came to Scholar Social in order to avoid being included in unethical social media research, and so we place a higher value on conducting research on human subjects with informed consent than most other social media, in order to maintain high ethical standards.
>
> At the admin's discretion, you may be asked to submit your research protocol, and the institution that is providing ethical review, even in the case that it is an opt-in survey. Failure to disclose your research plan may result in closure of your account, or having your network banned from our service.

Note that **10487 posts** from the scholar.social domain were captured in the dataset formerly available under doi:10.7910/DVN/R1HKVS, as measured by

```
cat timeline_*.jsonl | jq .uri | grep scholar.social | wc -l
```

2 Failure to de-identify data

The authors state that "Since the Mastodon user may be unaware of their data being public and reusable for research purposes we disposed of the information about the users and we fully anonymized them by hashing the Mastodon user identifier." Unfortunately, this is not the case. The authors appear to have neglected the uri field in the data for each toot. This field contains both the domain name of the instance as well as the user's name and number. This information identifies the post and the Mastodon user who made it, in all of the data that Zignani et al. scraped, aggregated, and made available to the world. The URI as-is can be used by anyone in the world to directly view the original post through their browser (if the post's visibility is "Unlisted" or "Public", as described below). The URI also contains a timestamp on when the post has been published (in milliseconds after 1970, UNIX time) and thus provides the means to identify when certain users have been online over a long term. Consequently, the authors' dataset consists of posts that have been scraped and reproduced not just without users' consent, but also without anonymization.

This contradicts both the authors' claim as well as requirements outlined by the GDPR.

We note that the uri field is visible in the authors' Figure 2, "An example of [a] toot in JSON format". It is difficult to imagine how this part of the data could have been neglected.

Moreover, the authors make no mention of de-anonymization, the practice of discovering the authorship of supposedly anonymous material. Simply put, content itself can reveal its creator's identity even after metadata has been removed. Distributing a dataset without regard for this possibility is highly irresponsible.

3 Fundamental mistakes of data analysis

The topic nominally being addressed by Zignani et al. is the matter of "inappropriate" content on social media. Posts on most Mastodon instances are known as "toots", a jocular reference to the sound that a large elephant-like mammal would make. The Mastodon software allows the user to hide the text of their toot behind a subject line, or "Content Warning". When such a toot is viewed, only the subject line is visible, until a "show more" button is clicked, whereupon the text is revealed. Zignani et al. clearly state that they use Mastodon's Content Warning feature to determine which toots are "inappropriate":

> By clicking on the "CW" button, a user can enter a short summary of what the "body" of her post contains, namely a spoiler-text, and the full content of her toot. Automatically, the system marks this toot as "sensitive" and only shows the spoiler-text in all the timelines. We exploit this latter feature to build our released dataset. This way the toots are labelled by the users, and we assume that they are aware of the policy of the instance and aware of what is appropriate or not for their community.

But this reveals a fundamental misunderstanding of the Mastodon community and how the software is used in practice. Even a brief time spent actually using the software makes clear that Content Warnings have many uses and do not always indicate posts that are "inappropriate" by any standard. A user can click the "CW" button to label a toot as containing discussion of politics, illness, injury, or bigotry (sexism, racism, homophobia, transphobia, and so forth). All of these topics are "appropriate", but a user may at their own discretion decide to provide advance warning for the benefit of those readers who wish to mentally prepare themselves for reading about emotionally damaging subject matter. Such CWs are acts of courtesy, not signals of "inappropriate" content. Users often apply CWs to toots about food and cooking, topics that are safe for children to read but may cause distress among readers with eating disorders. CWs can also hide spoilers about movies, books and television shows, and they can be part of the presentation of a joke: the "Content Warning" text contains the setup, and clicking to open the toot then reveals the punchline. By no stretch of the imagination is hiding the punchline of a joke an example of content that strays outside of community norms or that "may hurt people's feelings".

The authors' own figures cast intense doubt upon their identification of Content Warnings with "inappropriate" material or "bad content". We note the prominent placement of "cauliflower" and "cheesesteak" in the word cloud of "inappropriate" material on `mastodon.social`, and the comparable prominence of "patient", "medication" and "healthcare" in the `octodon.social` word cloud. It is not often that we wonder if the authors of an academic paper have looked at their own plots, but this is one of those times.

We can only describe the authors' discussion of Content Warnings as a total failure of comprehension, and we question the value of any research based on such a faulty grasp of how the Mastodon community operates. The authors claim, "The usage of this dataset empowers researchers to develop new applications as well as to evaluate different machine learning algorithms and methods on different tasks", but we see little chance of empowerment when the foundations are so flawed. As that old computer-science motto has it, *garbage in, garbage out.*

4 Mistaken classification of post privacy

The authors state that "Each toot has a privacy option, and users can choose whether the toot post is public or private," going on to describe "public messages", "private messages", and "local timeline" messages. The authors have unfortunately conflated two things: the way the Mastodon web interface aggregates posts, and the post privacy selections available to Mastodon users through that same interface. A cursory inspection of the Mastodon interface or the onboarding guide made available to new users makes clear the privacy settings, of which there are *four*:

- *Direct message*: These posts are visible only to the sender, the mentioned users, and the administrators of the users' respective instances.

- *Followers only*: These posts are only available to users' followers, and cannot be boosted (i.e., "retweeted", except by the originating user). Notably, *these posts cannot be accessed by navigating a browser to the post's URL.* This amounts to an explicit refusal of consent to scrape or copy the post.

- *Unlisted*: These posts may be boosted by any user that views them, and may be viewed by anyone with access to the post URL. However, they are not included in the aggregated timelines provided by the Mastodon user interface. This post privacy setting is specifically intended to allow circulation of a post without consenting to release of a post outside of Mastodon itself, as review of public Mastodon development discussions on Github would have revealed.

- *Public*: These posts are similar to unlisted posts, but also allow distribution to other instances with which the user's home instance federates. Importantly, *users can opt out of allowing search engines to index their public posts.*

The Mastodon user interface provides a *local timeline*, which is a live, reverse-chronological list of the public posts on the user's home instance. In addition, it provides a *federated timeline*, a live, reverse-chronological list of the public posts from a wider realm. Typically, the federated timeline of an instance contains all the public posts of all the users followed by any user on that instance. This allows rapid dissemination of posts among instances. Posts on local and federated timelines may or may not be open to indexing by search engines.

From a research standpoint, the authors' ignorance of the four visibility types and conflation of those types with post aggregation in the interface has not only led to a breach of ethics, but also a surprising lack of rigor. As indicated above, simply using the Mastodon interface would have dispelled half of the authors' ignorance, and a rigorous methodology consistent with peer-reviewed research work would have included research into the reasons for these visibility types.

Many Mastodon users, including the authors of this letter, are scientists with extensive experience writing and reviewing scholarly articles. Given these serious issues, we would not have accepted a paper in this condition for publication.

5 GDPR compliance issues

The authors state that their dataset is stored in Europe and thus protected by the GDPR, with which they erroneously claim they have complied. However, Harvard Dataverse had made this dataset available through their own infrastructure. It is unclear if this complies with GDPR requirements. That notwithstanding, the ethical breaches and lack of consent as detailed above similarly preclude release of the data by Harvard Dataverse. We express our gratitude for the prompt deaccessioning of the dataset from the Harvard Dataverse repository.

6 Relicensing and redistribution of copyrighted material

By posting to the Harvard Dataverse, the authors have released the scraped dataset under the Creative Commons CC0 license, a choice of license that is tantamount to putting material in the public domain. This blatant relicensing constitutes, at the very least, a serious misuse of CC0 and, in our view, a breach of the basic ethical principles underlying the notion of copyright. Referring again to the `scholar.social` Terms of Service:

> User-submitted content, including profile information, avatars, uploaded media, posted content, replies and any other information submitted to Scholar Social *remain the property of the user who submits it* (provided they owned it to begin with). Scholar Social members and staff do not monetise or sell material posted here. *Users retain complete creative and legal control of their own submitted material.* [emphasis added]

Many users came to Mastodon in order to preserve their control over the content they create, rather than signing it over to one gigantic corporation or another. Attempting a wholesale relicensing of Mastodon user content is a grievous appropriation. Quoting the Creative Commons organization themselves, "You should only apply CC0 to your own work, unless you have the necessary rights to apply CC0 to another person's work."

7 Neglect of risk presented to users

A text search of the paper for any of the terms "lgbt", "minor", or "margin" finds no results. This is concerning, especially in a paper that purports to be written with the goal of addressing "inappropriate" content, which is necessarily a culture-bound concept. A significant number of Mastodon users are members of ethnic, gender, sexual, and/or many other minority groups, and as members of marginalized groups, they face a myriad of risks and harms, including harassment, persecution, and physical harm. (This also applies to users who are minors.) In other words, a great many of these users are significantly more vulnerable than users who are members of majority groups. The authors could have learned of this from any number of public blog posts written by such users, as well as through a simple aggregate count over the data they themselves scraped without consent.

As such, ethical concerns regarding this paper and dataset are heightened. The authors' invasion of such users' privacy is itself a significant harm, but the unexamined hazards created by their publication of the dataset are severe. This neglect is egregious. Even the most cursory research into risks faced by users of color and LGBTQIA+ users around the world reveals these risks to be quite substantial. These users are the people whose very existence has been deemed "inappropriate" throughout many social spheres in history, and this prejudice continues up to the present moment. The published paper as well as the authors' methodology should have included due consideration and discussion of these issues.

Beyond the matters this letter has already detailed, we have also identified other passages where the paper betrays a lack of familiarity, or even casual acquaintance, with Mastodon and the existing writings about it. For example, the authors' description of the switter.at instance reveals ignorance of, or indifference to, a significant chapter in Mastodon history that has already been documented by technology journalists. These shortcomings contribute to the paper's general atmosphere of carelessness, though in our estimation they are less severe than its ethical lapses and its fundamental misapprehension of data.

8 Lack of Acknowledgment of Funding Sources

The authors fail to disclose their sources of funding. Not only is this unusual for an academic paper, to say the least, it prevents the reader from investigating whether the work was done in compliance with the funding agency's standards for responsible research. The AAAI claims that submissions to these conference proceedings were "carefully reviewed by 66 senior program committee members, 117 program committee members, and 328 additional reviewers." We find it puzzling that this process accepted a paper on anthropological research which omits an acknowledgment of its funding sources.

9 Remedies

We ask that the authors take the following actions:

- Apologize publicly in AAAI Proceedings for this violation of vulnerable users' consent
- Retract their paper from publication
- Delete all copies of the dataset they control
- Disclose their funding sources

We ask that the AAAI take the following actions:

- Print this letter in its Proceedings
- Issue a retraction of the "Mastodon Content Warnings" paper until the ethical failures have been addressed
- Should the paper be republished in any form, ensure personally identifiable information (such as verbatim posts, usernames, profile pictures, etc.) within screenshots in the paper or footnotes is redacted
- Investigate the review process for its conference proceedings, with particular attention to the question of whether the committee members and other reviewers have the relevant ethics experience to review social-media dataset papers

We ask that the University of Milan take the following actions:

- Investigate the basis on which this research was initiated, and whether any junior researchers involved were adequately supervised
- Issue a statement of retraction through the University public-relations system and any other means by which the original paper may have been advertised
- Delete all copies of the dataset they control

We ask that Harvard University take the following actions:

- Fully commit to the deaccessioning of the "Mastodon Content Warnings" dataset by removing it and all backups of it from Dataverse and any other Harvard-owned infrastructure

10 Signatories of this letter

We are writing as users of Mastodon and other intercompatible software platforms that participate in the federated social network (the *fediverse*) of which it is the most visible component. The opinions we express herein are not official position statements of our employers or funding agencies.

1. Noëlle Anthony, owner of academic research firm Anthony Expert Services LLC; administrator of `elekk.xyz` (`@noelle@elekk.xyz`)

2. Marie Axelsson (@maloki@elekk.xyz)

3. Daniel Bohrer, citizen of chaos.social (@daniel_bohrer@chaos.social)

4. Benjamin G Carlisle PhD, Postdoctoral researcher, QUEST (Berlin Institute of Health); administrator of scholar.social (@socrates@scholar.social)

5. Maëlan Chapovaloff, administrator of pipou.academy (@shad@pipou.academy)

6. Lorenz Diener, administrator of icosahedron.website (@halcy@icosahedron.website)

7. Alejandro Gaita-Ariño, Senior Researcher at the Institute of Molecular Science, Universitat de València (@agaitaarino@todon.nl)

8. Millia Gallé-Tessonneau, administrator of eldritch.cafe (@milia@eldritch.cafe)

9. Michael Gerdemann, owner and administrator of dizl.de (@nottulner@dizl.de)

10. Stéphane Guillou, Technology Trainer, The University of Queensland (Library) (@stragu@mastodon.indie.host)

11. Gwenfar, Puffin & Joe, administrators of sunbeam.city (@GwenfarsGarden, @puffinus_puffinus, @joecassels@sunbeam.city)

12. Vieno Hakkerinen, citizen of chaos.social (@txt_file@chaos.social)

13. Peter Hessler, owner and administrator of bsd.network (@phessler@bsd.network)

14. Host, administrator of tabletop.social (@host@tabletop.social)

15. Xan Indigo, Postdoctoral Researcher, Université Paris-Sud (@invaderxan@scholar.social)

16. Martin Kopischke, Head of Production at ifs international film school cologne (@kopischke@mastodon.social)

17. Tobias Kunze, owner and administrator of chaos.social (@rixx@chaos.social)

18. l4p1n, user of miaou.drycat.fr (@l4p1n@miaou.drycat.fr)

19. lawremipsum, administrator of mspsocial.net and attorney (@lawremipsum@mspsocial.net)

20. Leonie, administrator of koyu.space (@koyu@koyu.space)

21. Sam Lloyd, Cert HE (Open), administrator of computerfairi.es (@troubleMoney@computerfairi.es)

22. Bryce Alexander Lynch, Founder and Administrator, Virtual Adept Networks; Security Researcher, Special Circumstances, LLC (@drwho@hackers.town)

23. maple mavica syrup, BSc in Information Systems, owner and administrator of computerfairi.es (@mavica@computerfairi.es)

24. Matthew Meier, administrator of pettingzoo.co and faery.pub (@tyr@pettingzoo.co)

25. Erin Moon, Research Assistant, Electrical and Computer Engineering Department, University of Wisconsin–Madison; server administrator of social.mecanis.me (@er1n@social.mecanis.me)

26. Leah Oswald, owner and administrator of chaos.social (@leah@chaos.social)

27. Paul, administrator of kith.kitchen and software engineer at Natural History Museum, London (@paul@kith.kitchen)

28. Kim Reece (@kimreece@mathstodon.xyz)

29. Ros, user of weirder.earth and mastodon.xyz (@certifiedperson@weirder.earth)

30. Sascha, administrator of doesnt.undo.it (@laggard@doesnt.undo.it)

31. Isabelle Santos, postdoctoral researcher, UniGe (@moutmout@scholar.social)

32. self, administrator of lgbt.io (@self@lgbt.io)

33. Katt Sextant, vulpine.club member (@starkatt@vulpine.club)

34. Ana Silvia C. Silva, PhD in applied data science (@sissas@wandering.shop)

35. Kaito/Katie Sinclaire, owner of single-user instance is.a.qute.dog (@KS@is.a.qute.dog)

36. Blake C. Stacey, Research Assistant Professor, Department of Physics, University of Massachusetts Boston (@bstacey@icosahedron.website)

37. Sylvhem, administrator of eldritch.cafe (@Sylvhem@eldritch.cafe)

38. The_Gibson, admin of hackers.town and Chief Officer of BlackFireSec (@TheGibson@hackers.town)

39. Rylie James Thomas, administrator of gamemaking.social and makestuff.club (@rjt@makestuff.club)

40. Rey S. Tucker, administrator of vulpine.club (@rey@vulpine.club)

41. Wim Vanderbauwhede, professor in Computing Science, University of Glasgow, UK (@wim_v12e@octodon.social)

42. Walter Vannini, GDPR consultant and teacher of high-school mathematics (`@dataknightmare@octodon.social`)

43. David Wolfpaw, administrator of `tech.lgbt` (`@david@tech.lgbt`)

44. Liaizon Wakest, administrator of `social.wake.st` and `autonomous.zone` (`@liaizon@wake.st`)

A Relevant Articles from the 2019 *Codice etico e per l'integrità nella ricerca*

All quotations are from the official English version. From Article 16, "Feasibility, and social and environmental impact":

> Researchers assess the project feasibility, the ethical and legal aspects and, if necessary, the social requests and needs it satisfies. Should the project likely produce a significant impact on the objects of the research or, in general, on society, the environment or the biosphere, researchers shall responsibly examine the potential impact, providing details of these assessments in the appropriate documentation.

Article 26, "Informed Consent":

> Without prejudice to the principle of due respect for human dignity and autonomy, should the research entail the involvement of recruited participants, the research leader ensures that applicable norms on informed consent are respected, with special regard to incompetent subjects or, in any event, to individuals unable to give consent.

Article 27, "Storage and processing of personal data":

> 1. Storage and processing of personal data shall take place pursuant to applicable norms and University regulations. Participants in the research shall be provided with the names and contact details of the controller and processor of personal data.
> 2. Processing and storage of personal data of participants recruited for the study shall preferably be effected in codified or anonymous form. Should this not be possible due to the object of the research or its purpose, researchers shall scrupulously observe the provisions in force, in order to ensure due respect for the privacy of the persons involved.

Article 31, "Confidentiality":

> 1. The dissemination of research findings shall take place respecting the privacy of all persons involved.

2. If, due to scientific constraints, it is impossible to respect anonymity, personal data of research participants may be divulged only in conformity with their previous informed consent.

This open letter is released under the Creative Commons CC0 license.