The Medium is the Message: Toxicity Declines in Structured vs Unstructured Online Deliberations

Mark Klein[1,2](✉)

[1] Massachusetts Institute of Technology, Cambridge, MA 02139, USA
m_klein@mit.edu
[2] School of Collective Intelligence, University Mohammed VI Polytechnic, Ben Guerir, Morocco

Abstract. Humanity needs to deliberate effectively *at scale* about highly complex and contentious problems. Current online deliberation tools - such as email, chatrooms, and forums - are however plagued by levels of discussion toxicity that deeply undermine their utility for this purpose. This paper describes how a structured deliberation process can substantially reduce discussion toxicity compared to current approaches.

Keywords: collective intelligence · crowd-scale deliberation · toxicity

1 Introduction

Deliberation processes have changed little in centuries, perhaps even millennia. Typically, small groups of powerful stakeholders and designated experts craft solutions behind closed doors. Most people affected by the decisions have limited input, so important ideas and perspectives do not get incorporated, and there is often substantial resistance to *implementing* the ideas from those who were frozen from the process.

Humanity now however needs to deliberate effectively about highly complex, contentious, and existentially important problems – such as climate change, security, and poverty – where a small-circle process is no longer adequate. We need to find a way to effectively integrate the expertise and preferences of tens, hundreds or even thousands of individuals in our most consequential deliberations.

This paper addresses one important barrier to creating this capability: toxicity[1] in online deliberations. Online technology seems to represent our best hope for scaling up deliberations, but it has been plagued by debilitating levels of toxic comments. How can we fix that? As part of that discussion, we will cover:

- Goal: defining deliberation, and why scale is so important
- Challenge: the toxicity trap of existing deliberation technologies

[1] We define toxicity as the presence of rude, disrespectful, or unreasonable comments that are likely to make people leave a discussion.

R. Chbeir et al. (Eds.): MEDES 2023, CCIS 2022, pp. 374–381, 2024.
https://doi.org/10.1007/978-3-031-51643-6_27

- Solution: an introduction to deliberation mapping, a solution to online toxicity:
- Assessment: an evaluation of how well deliberation mapping reduces toxicity
- Conclusions: lessons learned and next steps

2 The Goal: Effective Deliberation at Scale

Let us define deliberation as the activity where groups of people (1) *identify* possible solutions for a problem, (2) *evaluate* these alternatives, and (3) *select* the solution(s) that best meet their needs (4).

Research from the field of collective intelligence has shown that engaging crowds in the way has the potential to unleash such powerful benefits as:

- *many hands*: the advent of cheap digital communication and ubiquitous personal computing has revealed the existence of a massive cognitive surplus: very large numbers of people with deep and diverse skill sets are eager to participate in collective tasks, driven by such non-monetary incentives as contributing to problem or communities they care about. (20) (22) Wikipedia is an excellent example of this.
- *casting a wide net:* frequently, solutions for difficult problems can be found by consulting outside of the usual small circle of conventional experts in that field (7). Innocentive is one example of a company that has been very successful exploiting this phenomenon.
- *idea synergy:* out-of-the-box solutions can often be achieved by bringing together many individuals and engaging them in *combining* and *refining* each other's ideas. The Matlab Coding Coopetition is a spectacular example of the power of this effect (12)
- *wisdom of crowds:* large numbers of suitably diverse, motivated and independent raters have been shown to produce assessment accuracy - e.g. for prediction and estimation tasks - that exceeds that of experts (21). Prediction markets are a powerful example of the value of this phenomenon.
- *many eyes:* our ability to detect possible problems in solution ideas increases dramatically by simply engaging more people in the task. This has been one of the key reasons for the success of such volunteer-created open-source software tools as Linux (the dominant operating system for supercomputers), Apache (the most widely-used web server), MySQL (the most widely-used relational DB) and the web toolkits used by Chrome, Firefox (the most popular web browsers in the world). These open source tools have decisively out-competed software developed by massive software companies with thousands of highly-paid engineers (16).

Engaging the relevant stakeholders in making decisions also has the great advantage of reducing the resistance and confusion that can occur when trying to actually *implement* the solutions developed by the deliberation engagement.

3 The Challenge: Limitations of Existing Technologies

While existing collective intelligence tools (i.e. email, wikis, chatrooms, forums, blogs, micro-blogs, and so on) have become very successful in some domains, they almost invariably produce poor outcomes when applied to large-scale deliberations on com-plex and contentious topics. This problem is inherent to the approach they take. These tools move conversations into the digital world, but they incorporate no model of what *kind* of conversations will lead crowds to quickly and efficiently find good solutions for complex problems. This frequently results in haphazard and highly inefficient de-liberations that produce large disorganized "comment piles" made up of frequently low-value content (Klein & Convertino, 2014).

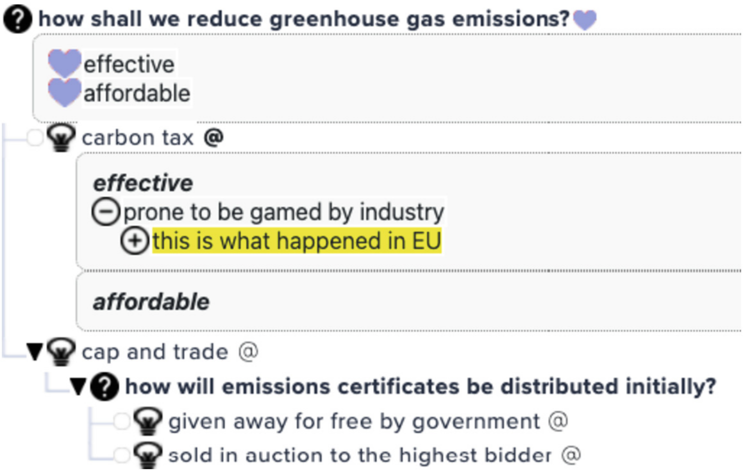
We will focus, in the paper, on one piece of this problem. Participants in conversation-centric tools frequently contribute toxic postings which enormously undercuts the willingness and ability of other participants to engage in thoughtful, meaningful, deliberations (2; 18) (23) (5) (11) (19) (17) (14) (6) (13) (1). All too often, such toxicity makes online discussions all but useless for deliberation, leading many organizations to either shut down their forums or invest in expensive and ethically fraught manual moderation of people's contributions to their forums. More recently, techniques for *automatically* detecting (and potentially filtering out) toxic posts have emerged. Perhaps the leading example of this is the Google Perspective API, which developed a set of toxicity-assessment rules by applying machine learning to a large corpus of manually-classified forum posts (8). While this approach represents a substantial advancement, it is far from perfect. The sarcastic phrase "I can see that you are a quote unquote *expert*" gets a very low toxicity score of 4/100, while a genuinely empathic comment like "Wow, it's terri-ble that someone called you a big fat jerk" gets a high toxicity score of 76/1000. It is also a *band-aid* solution, in the sense that it does nothing to address the underlying **cause** of the generation of toxic comments. Can we change the deliberation process in a way that prevents toxicity from happening in the first place?

4 A Solution to Toxicity: Deliberation Mapping

Deliberation mapping (Shum, Selvin, Sierhuis, Conklin, & Haley, 2006) is an alternative to unstructured online conversations that engages participants in co-creating logically-organized knowledge structures rather than conversation transcripts. As we will see below, the introduction of this structure fundamentally changes the participant incentives and results in a substantial reduction in toxicity.

The work reported in this paper uses a form of deliberation mapping called the "Deliberatorium" (9). It represents the simplest form of deliberation map that, in our experience, enables effective crowd-scale deliberation. Our map schema is built of "QuAACRs", i.e. *questions* to be answered, possible *answers* for these questions, *criteria* that describe the attributes of good answers, *arguments* that support or rebut an answer or argument, and *ratings* that capture the importance of questions and criteria, the value of answers, and the strength of arguments:

Deliberation maps have many important advantages over conversation-centric approaches. All the points appear in the part of the map they logically belong, e.g. all answers to a question are attached to that question in the map. It is therefore easy to find all the crowd's input on any given question, since it is collocated in the same branch. It's also easy to check if a point has already been contributed, and therefore to avoid *repeating* points, radically increasing the signal-to-noise ratio. Detecting and avoiding redundancy can in fact be mostly automated by the use of semantic similarity assessment



how shall we reduce greenhouse gas emissions?
effective
affordable
carbon tax @
effective
⊖prone to be gamed by industry
⊕this is what happened in EU
affordable
cap and trade @
how will emissions certificates be distributed initially?
given away for free by government @
sold in auction to the highest bidder @

Fig. 1. An example of a deliberation map.

tools based on text embedding technology (15). *Gaps* in the deliberation - e.g. questions without any answers, or answers without any arguments - are easy to identify, so we can guide crowd members to fill in these gaps and foster more complete coverage. Making arguments into first-class map entities implicitly encourages participants to express the evidence and logic for or against competing answers (3), and means that arguments can be critiqued individually. Users, finally, can easily collaborate to refine proposed solutions. One user can, for example, propose an answer, a second raise a question about how that answer can achieve a given requirement, and a third propose possible answers for that sub-question (see Fig. 1).

Why should this approach reduce toxicity? As was pointed out by media theorist Marshall McLuhan in his 1964 book *Understanding Media* (10), the nature of the discussion medium we use can have a profound impact on *what* we communicate. In a sense, as he points out, the medium *is* the message. How, then, do online discussion media shape what we say? In such tools, one of the key questions for participants is: how do I win the **attention war** as new posts pile on? Our inputs can easily be overlooked unless we frame them in ways that are likely to gather more attention. One guaranteed way to do that is to be more *extreme/toxic* than the others in the discussion. But if most people follow this individually rational strategy, the result is an upward toxicity spiral as contributors become more extreme in order to compete with other people using the same strategy.

Deliberation maps have different rules that in turn change the incentive structures (and thus typical behaviors) for contributors. Participants no longer need to engage in extremization in order to make themselves visible. Everybody's points on a given topic are co-located, right next to each other, and every unique idea appears just once, regardless of when or how often they were contributed. Deliberation maps make it immediately visible whether an individuals' postings have underlying (PRO or CON) arguments from the original poster and other crowd members. The "game" therefore changes *from* simply trying to get attention in a massive growing comment pile *to* creating points that people find compelling. In this context, less extremized and more carefully-argued points, we hypothesize, are instead more likely to receive positive evaluations. Based on this, we hypothesized that toxicity in deliberation maps will be significantly less than that in conventional (conversation-centric) forums.

5 Experimental Evaluation

We assessed the toxicity of the posts contributed in a random controlled trial consisting of two demographically matched experimental conditions of over 400 participants each:

- *Forum*: Participants used a forum (AKA threaded discussion) to submit posts as well as reply to other posts The posts and subsequent multiple levels of replies were viewed as an indented outline. Since users can contribute any kind of posts at any time, we considered this the "unstructured" condition.
- *Deliberatorium:* Participants used the Deliberatorium system, described above, to post questions answers and arguments in response to the newspaper articles. Since users are asked to contribute posts in a specific format (i.e. as questions answers and arguments in a logically-organized "map"), we considered this the "structured" condition.

Participants in each condition were asked to discuss, using their assigned tools, the content of the following eight newspaper articles (used with permission from the New York Times):

- Finding Compassion for 'Vaccine-Hesitant' Parents By Wajahat Ali
- We've All Just Made Fools of Ourselves—Again By David Brooks
- Why Are Young People Pretending to Love Work? By Erin Griffith
- New Zealand Massacre Highlights Global Reach of White Extremism By Patrick Kingsley
- The India-Pakistan Conflict Was a Parade of Lies. By Farhad Manjoo
- The West Doesn't Want ISIS Members to Return. Why Should the Syrians Put Up With Them? By Abdalaziz Alhamza 3/14/2019 at 10:52:22 pm
- Britain Is Drowning Itself in Nostalgia By Sam Byers 3/24/2019 at 4:34:26 pm
- If Stalin Had a Smartphone. By David Brooks 3/13/2019 at 0:19:42 am

The group participants were recruited using ads on a range of social media platforms including Facebook and others, and almost exclusively were based in the UK.

Neither of these conditions were moderated: so participants were free to take any tone they chose in their postings.

We used the Google Perspective API (https://www.perspectiveapi.com/) to assess the toxicity of the posts from the two conditions on a scale from 0 (non-toxic) to 1 (highly toxic). While, as noted above, the Perspective API is imperfect, it is the acknowledged state-of-the-art tool for this purpose and is widely used.

The average toxicity for the posts generated in the two conditions was as follows:

Platform	# posts	Average Toxicity
forum	915	0.19
deliberatorium	812	0.14

While the overall toxicity levels were relatively low in our community, the average toxicity of the forum posts was 30% higher than the deliberation map posts: this difference was highly significant statistically (p < 1.5 * 10^-10). We also found that high toxicity posts (i.e. with toxicity scores above 0.3) were twice as common in the forums than in the deliberation maps (Fig. 2):

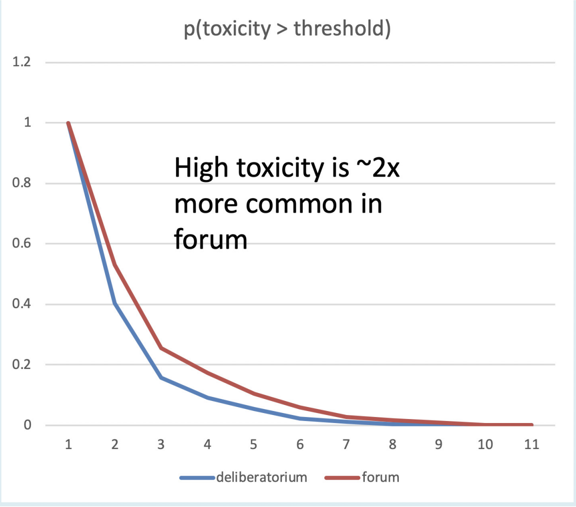


Fig. 2. Cumulative probability of posts above the given toxicity threshold, forum vs Deliberatorium.

6 Discussion

Toxicity has emerged as one of the major challenges for those who hope to enable useful crowd-scale online deliberations around complex and contentious topics. Our work has demonstrated that the level of toxicity in online discussions is deeply affected by the *way in which the discussions take place*. The structured nature of deliberation mapping, we believe, changes the rules of the game in a way that makes toxic comments no longer part of a winning strategy. Our data provides initial support for this hypothesis based on a carefully designed randomized control trial experiment involving a total of over 800 participants.

For future work, we would like to reproduce these experiments with communities and topics where the base toxicity level in the online forums is substantially higher, so we can assess the power of structuring conversations on reducing toxicity in more severely challenging contexts.

Acknowledgements. The author would like to acknowledge the contributions of Arnab Sircar and Paolo Spada to this work.

References

Almerekhi, H., Kwak, H., Jansen, B.J.: Investigating toxicity changes of cross-community redditors from 2 billion posts and comments. PeerJ Comput. Sci. **8**, e1059 (2022). https://peerj.com/articles/cs-1059/?via=indexdotco

Aroyo, L., Dixon, L., Thain, N., Redfield, O., Rosen, R.: Crowdsourcing subjective tasks: the case study of understanding toxicity in online discussions. In: Proceedings from World Wide Web Conference (2019)

Carr, C.S.: Using computer supported argument visualization to teach legal argumentation. In: Kirschner, P.A., Shum, S.J.B., Carr, C.S. (eds.) Visualizing argumentation: software tools for collaborative and educational sense-making, pp. 75–96. Springer-Verlag, Cham (2003). https://doi.org/10.1007/978-1-4471-0037-9_4

Eemeren, F.H.V., Grootendorst, R.: A Systematic Theory of Argumentation: The Pragma-dialectical Approach. Cambridge University Press, Cambridge (2003)

Hede, A., Agarwal, O., Lu, L., Mutz, D.C., Nenkova, A.: From toxicity in online comments to incivility in American news: proceed with caution. arXiv preprint: arXiv:2102.03671 (2021)

Jakob, J., Dobbrick, T., Freudenthaler, R., Haffner, P., Wessler, H.: Is constructive engagement online a lost cause? Toxic outrage in online user comments across democratic political systems and discussion arenas. Communication Research, 00936502211062773 (2022). https://doi.org/10.1177/00936502211062773

Jeppesen, L.B., Lakhani, K.R.: Marginality and problem-solving effectiveness in broadcast search. Organ. Sci. **21**(5), 1016–1033 (2010). http://dash.harvard.edu/bitstream/handle/1/3351241/Jeppesen_Marginality.pdf?sequence=2

Jigsaw: Reducing toxicity in large language models with perspective API (2023). https://medium.com/jigsaw/reducing-toxicity-in-large-language-models-with-perspective-api-c31c39b7a4d7

Klein, M.: How to harvest collective wisdom for complex problems: an introduction to the MIT Deliberatorium (2007). Retrieved

McLuhan, M.: Understanding Media: The Extensions of Man. Signet Books, New York (1964)

Mohan, S., Guha, A., Harris, M., Popowich, F., Schuster, A., Priebe, C.: The impact of toxic language on the health of Reddit communities (2017)

Gulley, N.: Patterns of innovation: a web-based MATLAB programming contest. In: CHI ‘01 Extended Abstracts on Human Factors in Computing Systems, pp. 337–338 (2001). https://doi.org/10.1145/634067.634266

Park, J.S., Seering, J., Bernstein, M.S.: Measuring the prevalence of anti-social behavior in online communities. Proc. ACM Hum.-Comput. Interact. **6**(CSCW2), 1–29 (2022).

Pelzer, B., Kaati, L., Cohen, K., Fernquist, J.: Toxic language in online incel communities. SN Soc. Sci. **1**, 1–22 (2021). https://doi.org/10.1007/s43545-021-00220-8

Ravichandiran, S.: Getting Started with Google BERT: Build and Train State-of-the-Art Natural Language Processing Models using BERT. Packt Publishing Ltd., Birmingham (2021). https://play.google.com/store/books/details?id=CvsWEAAAQBAJ&source=gbs_api

Raymond, E.: The cathedral and the bazaar. Knowl., Technol. Policy **12**(3), 23–49 (1999).

Rossini, P.: Toxic for whom? Examining the targets of uncivil and intolerant discourse in online political talk (2019)

Rossini, P.: Beyond toxicity in the online public sphere: understanding incivility in online political talk. Res. Agenda Digit. Polit., 160–170 (2020). https://www.elgaronline.com/display/edcoll/9781789903089/9781789903089.00026.xml

Salminen, J., Sengün, S., Corporan, J., Jung, S.-g., Jansen, B.J.: Topic-driven toxicity: exploring the relationship between online toxicity and news topics. PloS One **15**(2), e0228723 (2020). https://doi.org/10.1371/journal.pone.0228723

Shirky, C.: Here Comes Everybody: The Power of Organizing Without Organizations. Penguin, London (2009)

Surowiecki, J.: The Wisdom of Crowds. Anchor, New York (2005)

Tapscott, D., Williams, A.D.: Wikinomics: How Mass Collaboration Changes Everything. Portfolio Hardcover, Brentford (2006)

Xia, Y., Zhu, H., Lu, T., Zhang, P., Gu, N.: Exploring antecedents and consequences of toxicity in online discussions: a case study on Reddit. Proc. ACM Hum.-Comput. Interact. **4**(CSCW2), 1–23 (2020). https://doi.org/10.1145/3415179