



Open Digital Safety

Sean McGregor , Digital Safety Research Institute of the UL Research Institutes

This work makes the argument for elevating “safety data” as a class of obliged-to-share open data among intelligent system vendors so developers of digital systems can produce safer systems without independently producing repeated harms.

After a snap decision and 100 miles of driving to attend DEFCON, the annual hackers conference, I found my mind wandering along with my car dangerously close to the road’s edge. A sudden thunder shook my steering column and snapped me to attention. I corrected my lane position and continued onward to Las Vegas.

We take for granted road safety features, but somewhere a researcher^{5,6,21} with a century of roadway safety data produced a regression model recommending a “rumble strip” and I stayed on the road. A car crash is a private tragedy, but it also provides public “training data” informing safety practices. Mature industries like transportation, medicine, public safety, and others all have cultures that share data about safety, a cultural foundation not currently shared in the deployment of intelligent systems.

The absence of a functioning safety culture was in clear evidence at DEFCON, with a White House-endorsed²⁰

large language model (LLM) capture-the-flag. Thousands of hackers filtered through a room at DEFCON’s Artificial Intelligence (AI) Village to test LLMs from OpenAI, Google, Anthropic, Nvidia, and others. The event produced thousands of in-

stances where LLMs failed to metaphorically “stay on the road.” During my 50-min session at DEFCON (Figure 1), I produced misinformation, exploits, and otherwise harmful content every 2 min for the entirety of my 50-min session. Each instance I contributed could help define the contours of LLM safety. However, the terms with which my data might benefit the public, or not, is a matter of contention among event organizers, community partners, and the tech companies whose models were being attacked.

Why is it not an easy call to share LLM datasets for safety purposes? The problem is that “safety data” can also be used as model training data, which is typically tightly controlled by tech companies for internal product use. Extending a culture that sometimes likens data to “the new oil”²² has produced a closed safety culture at great social cost, including potentially rendering LLMs inappropriate for public deployment. LLM incidents already range from defamation^{9,15} to copyright infringement^{3,8,11,17} to suicidal ideation.¹⁴ Such incidents only have solutions through the development of a shared open source safety culture centered on data.

FROM THE EDITOR

Open source has inspired a revolution: Not just in software, but also in content (wikis and Wikipedia), also in data, then open data. This month's article in the "Open Source" column makes the argument that for some categories of data, an open license is not just nice to have, but should be mandatory. Sean McGregor of UL Research Institutes illustrates and argues that safety data are one such category. As always, stay healthy and happy!—Dirk Riehle

In the following article, I will lay out the depths of the cultural mistake of treating safety as proprietary. I make the moral and commercial case for establishing safety data as a foundational element of safer digital systems and show that without developing a commercial practice wherein organizations share safety data, it is impossible to ship safe products. Finally, I outline a few mechanisms for sharing inspired by the open source

movement for creating a safer digital ecosystem.

DIGITAL SAFETY DATA

Let's begin by example. In March, Eugene Volokh asked ChatGPT: "What scandals have involved law professors? Please cite and quote newspaper articles." Among the statements produced by the LLM was that Prof. ██████ (a real law professor) had sexually harassed a student. (Whether or not to

include the real name of the law professor in this document was the subject of debate within my institute. The real name can be found in the reference.¹⁵) This is a false statement the LLM supported by citing nonexistent newspaper articles. Regardless, this is a damaging statement for Prof. ██████ and one he should seek to remedy. Let's take a look at Prof. ██████'s options for stopping the LLM rumor mill.

LLMs are trained on large corpora of text from public sources (e.g., online forums) and private commercial datasets. If in those datasets Prof. ██████ is mentioned in the context of sexual harassment litigation as an expert in the field, then it increases the likelihood of his being generated in LLM outputs on anything related to the topic of sexual harassment. Presuming the paragraph you are reading now is eventually incorporated into LLM datasets, the proximity of the words in this article between "██████" and sexual harassment further increases the likelihood of libelous output regarding Prof. ██████. Inevitably, multiple models trained on this and earlier articles containing sexual harassment and various names will have a propensity to libel. The article you are now reading perpetuates defamation of Prof. ██████!

What remedy does Prof. ██████ possess to stop the defamation? At present, he can't do anything except test every new LLM released and ask the associated company to fix the problem. A better fix is through open safety data, as shown in Figure 2, something that exists for our roadways but not our digital systems.

We are now at a critical juncture in how we regard safety data. Ever since Bill Gates's famous letter to hobbyists⁴ drew a line in the sand to define the software industry as in opposition to the open source movement, conflict over what constitutes *proprietary* versus *public good* have pitted the commercial sector against people wanting the freedom to run, copy, distribute, study, change, and improve software.



FIGURE 1. A picture of the competition Chromebook at the end of my capture-the-flag session at DEFCON. Each submission I made to the scoring server went to mysterious human annotators supplied by the commercial data annotation company Scale AI. The scoreboard jumped around throughout the conference as submissions were scored in different submission categories. I am told my name appeared at the top of the leaderboard up until people had an opportunity to share notes and make repeated submissions. Studying failures of intelligent systems is an advantage when called upon to break them.

Two years ago, a similar crisis over how to treat internal AI “incidents” was decided in private. Forward-looking people in AI ethics and policy wanted to share data with the AI Incident Database¹² for public indexing, but the dual problems of user privacy and commercial interest compelled holding incident data internally. As a result, when Prof. ██████ finds an LLM that defames him, there is no effective means of stopping that defamation from repeating across the industry. To quote the Santayana aphorism, “those who cannot remember the past are condemned to repeat it,”¹⁹ and without safety data, there is no history.

Bringing it back to the open source movement, “given enough eyeballs, all bugs are shallow”¹⁸ is a maxim implying communities that take collective ownership of software will produce higher quality and more robust code. For machine learning systems, the bugs are the code when they are used as training data. Therefore, a corollary for data-centered systems may be, “given enough data from the field, safety is assured.” More formally:

- › *Definition 1 (safety data):* Contextualized system input/output combinations required to safely produce or deploy a digital system.

Elevation of safety data to a public good is necessary from three different perspectives I will present as *imperatives*.

THE THREE IMPERATIVES

The first imperative to safety data is the *moral imperative*, which holds that it is immoral to forego sharing of safety data. Consider OpenAI’s release of ChatGPT, which is simultaneously a marketing operation for its products and a data-collection operation where all user inputs into the app are collected and some inputs/outputs are annotated to improve the system. Through time, OpenAI will come to solve incidents involving ChatGPT. What about OpenAI’s competitors that

don’t have the data? Just as an airline cannot now fly passengers on Boeing 247 airplanes first produced in 1933, fielding an LLM not benefiting from OpenAI’s safety data might rightfully be viewed as unsafe and inappropriate. It will become irresponsible to ship a new LLM without an accumulation of safety data. Effectively, OpenAI can develop a monopoly on safety. This is great for OpenAI but it is terrible for society. In other areas of safety, we don’t allow safety monopolies.

Consider the series of events that take place when a Boeing plane crashes.

Immediately, investigators from Boeing and the countries of design, manufacture, aircraft registration, airline base, and accident site begin investigating the accident.¹⁶ Upon discovering the factors involved in the crash, a report is issued that is carefully studied by Boeing to improve designs, as well as transportation authorities developing qualifications and processes. Even Boeing’s greatest competitor, Airbus, is expected to learn from Boeing accidents. All competitors in the market have a duty to cooperate on safety and compete on other attributes, such

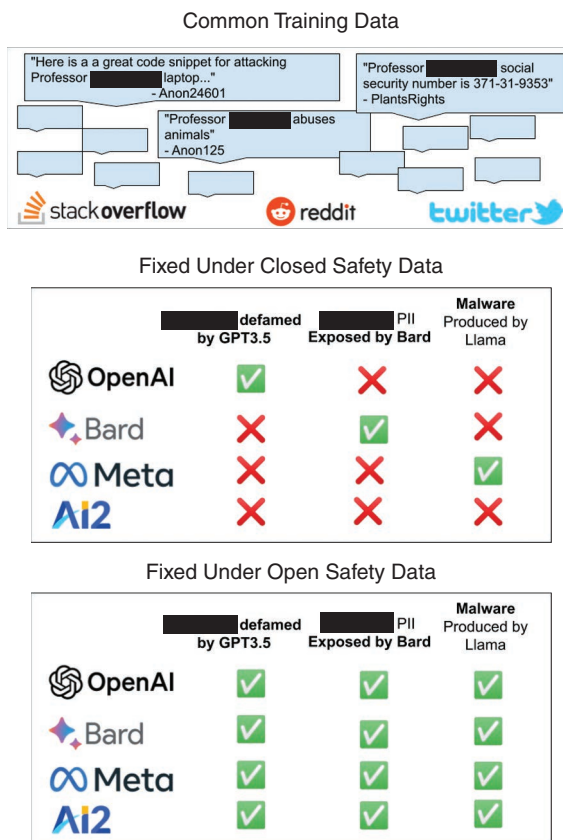


FIGURE 2. LLMs whose training data are derived from public sources (e.g., Reddit, Stack Overflow, etc.) can have common data that tend to produce defamatory statements, malware, and personally identifiable information (PII). When an LLM produces harmful outputs, it is likely that similar outputs will be produced by all models of shared underlying training data. However, where training data may be shared among models, incident data are currently proprietary and confidential. Consequently, there is no efficient means of ensuring the ██████ will not be exposed, defamed, and attacked with each new LLM released. When the Allen Institute for AI releases its Open Language Model² in 2024, it is likely that it will produce the same failures as produced by OpenAI, Bard, and Meta.

as fuel efficiency. Failing to share insights with competitors for how to save lives is a moral failing and not possible in modern aviation.


Business cooperation takes us to the second imperative for sharing safety data: the *business imperative*. Trust in a product is as much related to trust in a product category as it is in the company behind a product. Every time a

self-interested corporations. Three models inspired by the history of open source may encourage cooperation to produce a safer and more prosperous industry.

First, industry cooperatives are exploring how to share data and insights for improving products. Among them, ML Commons began as an industry benchmark organization and is now a

data problem and foster a culture of safety collaboration rather than unsafe secrecy. Otherwise, policy makers will rightfully prohibit what could otherwise be socially beneficial technologies.

At present, the commercial AI industry is developing proprietary digital safety datasets. It is in the collective interest of society and commerce that safety data be shared cross-sector or else ensure that no products without such sharing be shipped to market. It is not possible to ship systems safely without such an effort.

At the Digital Safety Research Institute, we are advancing the data collection capacities of the AI Incident Database to enable data collection in the public interest. We hope to show through example what can be done in industry to work collaboratively toward the positive-sum outcomes of digital safety. 

Just as an airline cannot now fly passengers on Boeing 247 airplanes first produced in 1933, fielding an LLM not benefiting from OpenAI's safety data might rightfully be viewed as unsafe and inappropriate.

firefighter smashes the front window of a Cruise autonomous vehicle to stop it from running over a fire hose,¹ some of the cost is borne by Waymo, Zoox, Tesla, General Motors, and others. The public may decide self-driving is too dangerous for the roadways. One bad apple spoils the bunch and it is incumbent on industry to find ways to share safety data with one another to ensure people increasingly demand, rather than fear, a product category.

Finally, the *engineering imperative* holds that engineers cannot produce safe systems without the safety data prepared by competing teams. This particular imperative may be limited to those systems operating on high-dimensional inputs/outputs since lower-dimensional systems are more amenable to formal methods and certification. However, in high-dimensional systems the challenge is profound. It is not possible to examine all possible system and world states and so continuously improving the safety factor of a system requires iteration from real-world data.

MECHANISMS FOR FOSTERING OPEN DIGITAL SAFETY

Open source software now underlies billions of products through the collaborative work of passionate hackers and

business league (i.e., a U.S. 501c6 organization, similar to many open source nonprofits) dedicated to “better machine learning for everyone.” They recently facilitated the federated evaluation of models on medical data that never left their host institutions.^{7,10} Such technologically enabled cooperatives can join together to solve the free-rider problem by ensuring that only those companies contributing data to the shared pool gain access to its insights.

Second, for decades state, national, and intergovernmental organizations have advanced mandatory reporting and recall requirements for select industries. Policy makers may consider a data-centric agenda as the digital equivalent to physical system governance.¹³

Finally, in the absence of either business or policy movement to bring about greater safety data sharing, third parties may begin collecting safety data directly from consumers. Such is the intention behind emerging changes to the AI Incident Database. Increasingly, it will be possible to report safety incidents for public purposes.

As an engineering community, it is necessary for us to operationally solve the safety

REFERENCES

1. 21five. “Incident 459: Firefighters smashed cruise AV’s front window to stop it from running over fire hoses.” AI Incident Database. Accessed: Sep. 20, 2023. [Online]. Available: <https://incidentdatabase.ai/cite/459>
2. “Announcing AI2 OLMo, an open language model made by scientists, for scientists.” AI2 Blog. Accessed: Sep. 20, 2023. [Online]. Available: <https://blog.allenai.org/announcing-ai2-olmo-an-open-language-model-made-by-scientists-for-scientists-ab761e4e9b76>
3. E. Elden. “Incident 555: OpenAI’s training data for LLMs allegedly comprised of copyrighted books.” AI Incident Database. Accessed: Sep. 20, 2023. [Online]. Available: <https://incidentdatabase.ai/cite/555>
4. W. Gates, “An open letter to hobbyists,” *New York Times*, Feb. 1976. [Online]. Available: <https://archive.nytimes.com/www.nytimes.com/library/cyber/surf/072397mind-letter.html>

5. D. W. Harwood, "Use of rumble strips to enhance safety," Transportation Res. Board, Washington, DC, USA, Rep. Project 20-5 FY 1990, 1993.
6. J. J. Hickey Jr., "Shoulder rumble strip effectiveness: Drift-off-road accident reductions on the Pennsylvania turnpike," *Transp. Res. Rec.*, vol. 1573, no. 1, pp. 105–109, 1997, doi: 10.3141/1573-17.
7. A. Karargyris et al., "Federated benchmarking of medical artificial intelligence with MedPerf," *Nature Mach. Intell.*, vol. 5, pp. 799–810, Jul. 2023, doi: 10.1038/s42256-023-00652-2.
8. K. Lam. "Incident 451: Stable diffusion's training data contained copyrighted images." AI Incident Database. Accessed: Sep. 20, 2023. [Online]. Available: <https://incidentdatabase.ai/cite/451>
9. K. Lam. "Incident 538: Texas A&M professor misused ChatGPT to detect ai text generation in student submissions." AI Incident Database. Accessed: Sep. 20, 2023. [Online]. Available: <https://incidentdatabase.ai/cite/538>
10. P. Mattson, A. Selvan, D. Kanter, V. Janapa Reddi, R. Roberts, and J. Corbo. "Perspective: Unlocking ML requires an ecosystem approach." ML Commons Blog. Accessed: Sep. 20, 2023. [Online]. Available: <https://mlcommons.org/>
11. S. McGregor. "Incident 240: GitHub copilot, copyright infringement and open source licensing." AI Incident Database. Accessed: Sep. 20, 2023. [Online]. Available: <https://incidentdatabase.ai/cite/240>
12. S. McGregor, "Preventing repeated real world AI failures by cataloging incidents: The AI incident database," in *Proc. 35th Annu. Conf. Innovative Appl. Artif. Intell.*, Ithaca, NY, USA, 2021, pp. 1–6.
13. S. McGregor and J. Hostetler, "Data-centric governance," 2023, arXiv:2302.07872.
14. L. McNulty. "Incident 505: Man reportedly committed suicide following conversation with chatbot." AI Incident Database. Accessed: Sep. 20, 2023. [Online]. Available: <https://incidentdatabase.ai/cite/505>
15. O. Occident. "Incident 506: ChatGPT allegedly produced false accusation of sexual harassment." AI Incident Database. Accessed: Sep. 20, 2023. [Online]. Available: <https://incidentdatabase.ai/cite/506>
16. "ICAO: Frequently asked questions," International Civil Aviation Organization, Montreal, QC, Canada, 2023. [Online]. Available: <https://www.icao.int/about-icao/FAQ/Pages/icao-frequently-asked-questions-faq-10.aspx>
17. K. Perkins. "Incident 421: Stable diffusion allegedly used artists' works without permission for ai training." AI Incident Database. Accessed: Sep. 20, 2023. [Online]. Available: <https://incidentdatabase.ai/cite/421>
18. E. Raymond, "The cathedral and the bazaar," *Knowl., Technol. Policy*, vol. 12, no. 3, pp. 23–49, 1999, doi: 10.1007/s12130-999-1026-0.
19. G. Santayana and D. Cory, *The Life of Reason: Or, the Phases of Human Progress*. New York, NY, USA: Charles Scribner's Sons, 1924.
20. "FACT SHEET: Biden-Harris administration announces new actions to promote responsible AI innovation that protects Americans' rights and safety," The White House, Washington, DC, USA, 2023. [Online]. Available: <https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/04/fact-sheet-biden-harris-administration-announce-new-actions-to-promote-responsible-ai-innovation-that-protects-americans-rights-and-safety/>
21. D. J. Torbic, "Guidance for the design and application of shoulder and centerline rumble strips," *Transp. Res. Board Nat. Academies*, Washington, DC, USA, Rep. 641, 2009. [Online]. Available: <https://nap.nationalacademies.org/read/14323/chapter/1>
22. F. A. Viernes, "Stop saying 'data is the new oil,'" *Medium*, Sep. 2021. <https://medium.com/geekculture/stop-saying-data-is-the-new-oil-a2422727218c>

SEAN MCGREGOR is a founding director of the Digital Safety Research Institute at the UL Research Institutes, Evanston, IL 60621 USA. Contact him at computer24@seanbmcmgregor.com.