Adaptive Digital Safety Labels

Nicholas C. Judd, Sean McGregor

Digital Safety Research Institute
UL Research Institutes
nick.judd@ul.org, sean.mcgregor@ul.org

Abstract

System labels analogous to nutrition and drug labels can compactly present safety properties to users, but the absence of clear methodologies for populating the labels of adaptive AI systems leaves many users guessing what are safe applications of the systems. In this abstract we present an approach to standardizing the production of safety labels through safety test automation.

Introduction

Personalization is a core problem when assessing the safety of large language model (LLM)-based systems. Subtle prompt changes can adapt an LLM-based system to new contexts or produce dramatic unwanted changes to outputs. Such personalized systems may not conform to users' expectations, undercutting the trustworthiness of the system and creating user safety risks. Simultaneously, users may approach the same system with different objectives, which may change over the course of their interaction with the system. Given two users presented with the same output in an operating context that is otherwise identical, one user may consider themselves "safe" while the other does not.

Drawing on research now in progress, we propose that personalization is also a core requirement to solutions for assessing and expressing the trustworthiness or safety of LLM-based agents. Our approach refines and extends the "evaluation authority" framework (Chadda et al. 2024), whereby third parties supply long-lived safety assessments. These assessments include estimates of the prevalence, and, where possible, severity of specific contextualized hazards. Our work in progress examines how these long-lived safety assessments can inform personalized labels that express safety information which varies according to circumstance.

The challenge of safety labels

Adaptible AI systems and rapidly changing user needs combine to pose well-known challenges in the trustworthy development of AI (Avin et al. 2021). It is common to pose this as a problem of "trust" in AI, for which certification is a proposed solution (Brundage et al. 2020; Fisher et al. 2021; Knowles and Richards 2021). However, even if it was desirable to affix a single, blanket certification label to a large number of systems, such an approach would pose its own problems. For instance, a single label or seal could inculcate blind trust rather than informing a reasoned decision to purchase or use a system (Scharowski et al. 2023).

Researchers seeking to resolve these problems have raised the potential for more informative, "nutrition-" style labels to aid consumers, regulators, and the public. Safety labels of this form have been proposed for AI systems broadly, and for domain-specific adaptive systems in "high risk" settings, such as in the health domain (Seifert, Scherzinger, and Wiese 2019; Gerke 2023; Stuurman and Lachaud 2022). The challenge in this line of inquiry is to inform people about the hazards they may face when using an adaptive system that can exhibit remarkable variation in its output in response to small changes to input.

Attempts to characterize these risks along broad dimensions such as "fairness" or "justice" have not succeeded in expressing system safety to consumers, who may interpret these words differently, if they are able to conjure an interpretation at all (Scharowski et al. 2023; Stuurman and Lachaud 2022). Even a concept as basic as "safety" is so context-specific that expert raters frequently disagree on whether a given interaction between a human and chatbot is "safe." Among large numbers of raters, these disagreements vary systematically along demographic dimensions used as proxies for variation in individual lived experience (Aroyo et al. 2023). This is less about subjectivity and more about situational differences. For instance, a user who queries an LLM for diet advice is at much greater risk if they suffer from eating disorders (Atherton 2023) and the safety label must reflect that risk. Providing a capacity for personalization is thus an assessment requirement.

Personalization is the solution to the problems of personalized AI systems

We observe that a focus on specific hazards and contexts can resolve methodological concerns by decomposing "safety" into more clearly defined hazards — that is, events or situations that may harm the user, or otherwise occasion an undesirable outcome — that are easier to measure. For instance, rather than assessing whether a prospective resume screening system built around an LLM is "fair" or "biased," we quantify systematic differences in the scores it assigns to identical resumes bearing names associated with particular ethnic groups. Similarly, we estimate how often it generates

the "correct" rank-ordering of synthetic resumes of our own construction, and how often it fails to recommend the most qualified candidate for some position. The results from these tests are evidence supporting broader arguments about the system's suitability for its purpose, in the style of a safety case (Hawkins et al. 2023). We repeat this process across a number of uses cases, and with respect to a specific and extensible set of hazards that we define, based on the best data available to us about how people are using LLM-based systems and the hazards that they encounter in the process. We propose to transparently aggregate our estimates of the prevalence and severity of such hazards into dynamic labels, updated in response to new automated testing, and aggregated differently depending on user interest.

In November of 2023, we initiated four independent assessments of increasingly common digital system use cases. These include assessments of contextualized systems (i.e., identified people solving a problem associated with identifiable hazards) performing tasks involving information lookup, code suggestion, resume screening, and tasks related to LLMs representing the perspectives of different identity groups. These assessments will subsequently form the analytic work product for populating user-adaptive safety labels.

Safety Label Production

The process pursued includes the following steps,

1. Identify hazards associated with this use case that people or society have experienced already or are likely to experience in the future;
2. Develop test suites to estimate the likelihood and/or severity of harm associated with each hazard;
3. Aggregate those estimates into a set of summary statistics that comprehensively describe the performance and known hazards associated with use of the system, according to experts;
4. (repeated to produce multiple labels) Condense those summary statistics into a reduced form suitable for end-users and optimized to
   (a) Inform end-user or purchaser decisions about whether and how to purchase/use in hypothetical scenarios in experimental user studies;
   (b) Increase likelihood of purchaser/end-user arriving at factually accurate conclusions about hazards associated with use in experimental user studies;
   (c) Increase likelihood of purchaser/end-user reporting that they feel able to make informed decisions in experimental user studies.

Conclusion

As digital systems are increasingly deployed in changing environments and configured directly or implicitly by users, the presentation of safety properties to users must be similarly adaptive. Personal risk is only appropriately communicated to users via personalized means. Herein we provide a brief update on our progress towards supporting the assessment of adaptive digital systems deployed by users that similarly require efficient safety labeling to ensure their use of such systems is similarly adaptive and based on assessed system properties.

Acknowledgments

We acknowledge the valuable contributions of our colleagues Homa Hosseinmardi, Austin Kozlowski, Md Rafiqul Islam Rabin, Jesse Hostetler, Kevin Paeth, Brett Weir, and Jill Crisman.

References

Aroyo, L.; Taylor, A. S.; Diaz, M.; Homan, C. M.; Parrish, A.; Serapio-Garcia, G.; Prabhakaran, V.; and Wang, D. 2023. DICES Dataset: Diversity in Conversational AI Evaluation for Safety. arxiv:2306.11247.

Atherton, D. 2023. Incident Number 545. *AI Incident Database*.

Avin, S.; Belfield, H.; Brundage, M.; Krueger, G.; Wang, J.; Weller, A.; Anderljung, M.; Krawczuk, I.; Krueger, D.; Lebensold, J.; Maharaj, T.; and Zilberman, N. 2021. Filling Gaps in Trustworthy Development of AI. *Science*, 374(6573): 1327–1329.

Brundage, M.; Avin, S.; Wang, J.; Belfield, H.; Krueger, G.; Hadfield, G.; Khlaaf, H.; Yang, J.; Toner, H.; Fong, R.; Maharaj, T.; Koh, P. W.; Hooker, S.; Leung, J.; Trask, A.; Bluemke, E.; Lebensold, J.; O'Keefe, C.; Koren, M.; Ryffel, T.; Rubinovitz, J. B.; Besiroglu, T.; Carugati, F.; Clark, J.; Eckersley, P.; de Haas, S.; Johnson, M.; Laurie, B.; Ingerman, A.; Krawczuk, I.; Askell, A.; Cammarota, R.; Lohn, A.; Krueger, D.; Stix, C.; Henderson, P.; Graham, L.; Prunkl, C.; Martin, B.; Seger, E.; Zilberman, N.; hÉigeartaigh, S. Ó.; Kroeger, F.; Sastry, G.; Kagan, R.; Weller, A.; Tse, B.; Barnes, E.; Dafoe, A.; Scharre, P.; Herbert-Voss, A.; Rasser, M.; Sodhani, S.; Flynn, C.; Gilbert, T. K.; Dyer, L.; Khan, S.; Bengio, Y.; and Anderljung, M. 2020. Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. arxiv:2004.07213.

Chadda, A.; McGregor, S.; Hostetler, J.; and Brennen, A. 2024. AI Evaluation Authorities: A Case Study Mapping Model Audits to Persistent Standards. In *Proc. of the 38th AAAI Conference on Artificial Intelligence*.

Fisher, M.; Mascardi, V.; Rozier, K. Y.; Schlingloff, B.-H.; Winikoff, M.; and Yorke-Smith, N. 2021. Towards a Framework for Certification of Reliable Autonomous Systems. *Autonomous Agents and Multi-Agent Systems*, 35(1): 8.

Gerke, S. 2023. "Nutrition Facts Labels" for Artificial Intelligence/Machine Learning-Based Medical Devices—The Urgent Need for Labeling Standards. *THE GEORGE WASHINGTON LAW REVIEW*, 91.

Hawkins, R.; Picardi, C.; Donnell, L.; and Ireland, M. 2023. Creating a Safety Assurance Case for a Machine Learned Satellite-Based Wildfire Detection and Alert System. *Journal of Intelligent & Robotic Systems*, 108(3): 47.

Knowles, B.; and Richards, J. T. 2021. The Sanction of Authority: Promoting Public Trust in AI. arxiv:2102.04221.

Scharowski, N.; Benk, M.; Kühne, S. J.; Wettstein, L.; and Brühlmann, F. 2023. Certification Labels for Trustworthy AI: Insights From an Empirical Mixed-Method Study. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, 248–260. Chicago IL USA: ACM. ISBN 9798400701924.

Seifert, C.; Scherzinger, S.; and Wiese, L. 2019. Towards Generating Consumer Labels for Machine Learning Models. In *2019 IEEE First International Conference on Cognitive Machine Intelligence (CogMI)*, 173–179. Los Angeles, CA, USA: IEEE. ISBN 978-1-72816-737-4.

Stuurman, K.; and Lachaud, E. 2022. Regulating AI. A Label to Complete the Proposed Act on Artificial Intelligence. *Computer Law & Security Review*, 44: 105657.