Beyond Safety: Toward a Value-Sensitive Approach to the Design of AI Systems

Alexander J. Fiannaca
Google Research
Seattle, WA, USA
afiannaca@google.com

Cynthia L. Bennett
Google Research
New York, NY, USA
clbennett@google.com

Shaun K. Kane
Google Research
Boulder, CO, USA
shaunkane@google.com

Meredith Ringel Morris
Google Research
Seattle, WA, USA
merrie@google.com

Abstract

As modern, pre-trained ML models have proliferated in recent years , many researchers and practitioners have made significant efforts to prevent AI systems from causing harm. This focus on safety is critical, but a singular focus on safety can come at the exclusion of considering other important stakeholder values and the interactions between those values in the AI systems we build. In this position paper, we propose that the AI community should incorporate ideas from the Value-Sensitive Design framework [9, 11] from the Human-Computer Interaction community to ensure the needs and values of all stakeholders are reflected in the systems we build. We share observations and reflections from our experiences working on AI-supported accessibility technologies and with members of various disability communities to illustrate the tensions that sometimes arise between safety and other values.

1 Introduction

The foundation models of modern artificial intelligence [4] are pre-trained on vast amounts of information and consequently are generalizable to a wide range of tasks, resulting in AI systems that are more powerful than ever before. As the development of these systems has flourished, the opportunity for these systems to cause harm has grown [24], giving rise to significant concerns around the safety of AI models. As an example, the recently released text-to-image diffusion model Stable Diffusion has already been appropriated for generating harmful and pornographic content due to its wide release and lack of safety filters, resulting in concern from the research community and the public [6, 25].

Amid such incidents, significant effort within the research community has been devoted to developing methods to ensure the safety of AI systems [7]. Of note, the NeurIPS community proposed a code of ethics [1] that calls out safety of an AI system as an ethical concern and proposes an expansive view of safety, defining it as "protection from harm where this extends to include prospective negative impact on a person's physical, emotional or psychological well-being."

In practice, it is challenging to design for safety given that concepts of safety and risk often vary across cultures and use contexts. For example, Bennett et al. [2] found that U.S.-based blind screen reader users from minoritized backgrounds did not want AI systems to attempt to perceive people's gender in images due to the potential to misgender a person. However, in our current research, we

Preprint. Under review.

have found the opposite for blind women in India who want AI to identify perceived gender of bystanders when seeking assistance due to concerns of physical safety and cultural taboos around approaching male strangers. Further, safety systems can themselves cause harms, as was seen with the case of a father who was recently reported to the police by an automated safety system for possession of child pornography when he took photos of his son's skin infection to send to his pediatrician [13].

These examples illustrate tensions between safety and competing concerns commonly encountered when designing AI systems. To explore these tensions, we draw on the Value-Sensitive Design (VSD) framework from the Human-Computer Interaction community [9, 11]. VSD provides a framework through which designers can identify the range of stakeholder values and needs for a specific system (i.e., values and needs of end users, developers, organizations, etc.). Rather than having designers decide what issues are most important to address, adopting a VSD approach allows us to bring different stakeholders impacted by these systems to the foreground and elicit their particular values; by doing so, we may avoid dehumanizing or abstracting away from particular groups.

In this position paper, we explore some examples of values that can be in tension with safety when designing AI systems and discuss how design decisions with respect to these values can affect users. We draw these examples from our experiences as accessibility researchers, as we have found that individuals with disabilities often need to balance several values including safety, autonomy, and privacy when choosing to use specific technologies. We conclude with guiding questions to scaffold future research on promoting safety alongside other important values.

2 Examples of Value Tensions

Safety is a critical value embedded in the design of AI systems. However, there are many other stakeholder values that may be in tension with safety. Using accessibility as a case study in this section, we explore examples of where these tensions may arise and how they can impact stakeholders.

2.1 Personal Expression

In the field of augmentative and alternative communication (AAC) for people with mobility and speech impairments, recent work has explored the possibility of using AI models to autocomplete sentences for users who have difficulty communicating [5, 21]. However, as many AI models have filters that prevent them from generating text discussing specific topics, users of these systems could be prevented from discussing topics that the model creators deem inappropriate. Kane et al.'s [14] interview study with people with ALS who relied on AAC-mediated speech found that authenticity in self-expression was a top value for this demographic, to the degree that users would undertake additional work to communicate in-line with their expressive goals. Kane et al. [14] explored how communication barriers limited users' ability to express themselves authentically and, in some cases, led to the users withdrawing from social interaction. In our current work with people with mobility impairments who use AAC devices, we have observed that safety filters on large language models (LLMs) designed to prevent the model from outputting offensive or harmful language directly limit the users' ability to express themselves in situations where the LLMs filter out the text they are intending to write (e.g., restricting users' ability to use curse words or to discuss "taboo" interests such as smoking).

For communication mediated by AAC devices, it is critical that communication systems support the full breadth of human communication and not merely a limited "safe" subset of language (e.g., AAC devices must support the ability for users to have highly sensitive and unfiltered conversations with medical providers, to use intimate language with their romantic partners, etc.). Given that notions of what is considered offensive or harmful are dependent upon both culture and context, developers of AI models should be careful to examine how the tradeoff between safety and the ability of users to freely express themselves is embedded in the systems they develop; a value-sensitive design approach can help model developers reflect on this nuance as pertains to their target deployment context.

2.2 Autonomy

The integration of AI algorithms into the design of AAC devices is also a useful example for illustrating the tension between the values of safety and user autonomy. The degree to which a system can affect the autonomy of a user is a core consideration in HCI and design [10], and is particularly

salient in the context of users with disabilities. Fiannaca et al. [8] showed that the impact of an AAC feature's design on the user's autonomy when using the system to communicate was a core concern for users. Through this lens of user autonomy, the current design of LLM safety filters can be seen as restricting the autonomy of AAC users by putting limits on their ability to freely and accurately input utterances in the AAC device's UI without an added technological burden. Another tension relating to autonomy that we have seen in our current work with AAC users is the desire to have final ownership (approval, veto, and/or edit power) over any communications generated in full or in part by a model; similar issues arose in Goodman et al's [12] work on the use of LLMs to support writing by people with dyslexia, in which the system designers were careful to consider and measure the extent to which the automated writing scaffolds preserved user agency.

People with disabilities are often willing to accept a higher level of risk given the tradeoff of abilities gained through the use of technology [3] (the risk in this example being a risk of offense to interlocutors), and deserve to be afforded the agency to make this risk-reward calculation for themselves. While AI developers and their organizations may prioritize safety as the paramount value embedded in the systems they are building, such a design decision may not be acceptable to users who are disempowered by the status quo. Tanis and Lewis [22] explored this issue and noted that when decisions around accepting the risk inherent in using AI are made by individuals other than the users themselves (e.g. designers, developers, etc.), those users are denied the dignity of making those decisions. From the context of accessibility, denying users the dignity of making these value decisions themselves may further the long history of infantilization of people with disabilities [15, 18]. Incorporating a consideration of this value tension when designing LLMs and other AI models could lead to developing more flexible safety-related filters that can be tuned up or down to match the needs of the particular use case. However, applying differing levels of safety filtering for different users or situations raises complex ethical questions, presenting us with an opportunity to critically examine the tradeoffs between safety-related risks and benefits in a manner dependent on the application scenario. In essence, such an approach would embrace the nuance in designing AI systems for different use cases rather than applying a one-size-fits-all lens to safety and risk mitigation.

2.3 Privacy

Privacy has become a growing concern for many users as AI technologies have become more ubiquitous. However, the social acceptability of potentially invasive AI technologies is highly dependent on the context in which they are used [17]. For example, Profita et al. [19] showed that observers of a person using a head-mounted camera (Google Glass) found the use of the camera more socially acceptable when the person using the camera was perceived to be blind (i.e., wearing dark glasses and carrying a white cane), raising questions around enforcing disclosure of one's disability status as justification to use technology that may infringe on the privacy of other stakeholders, along with what such disclosure is expected to look like.

In our previous discussion of autonomy, we used disability as an illustrative example that highlights a particular risk/reward curve of autonomy vs. safety, though it is important to note that many user groups will highly prize autonomy. Furthermore, in the context of privacy, requiring users to reveal sensitive demographic information (e.g., disability status) in order to access more permissive models would clearly be in conflict with privacy and raise myriad ethical concerns, in addition to being impractical [17, 19]. This highlights the importance of offering flexibility that can be set by an individual user, rather than gated on group membership (though the latter approach may be appropriate for simply-defined groups such as children under a certain age).

Additionally, preferences around privacy and views of the acceptability of technology use may differ between users of the technology and indirect stakeholders whose privacy is impacted by that technology use. In the case described by Profita et al. [19], blind users were seeking scene descriptions of inaccessible infrastructure. Regarding people descriptions, which raise more salient privacy issues, Bennett et al. [2] reported some users' concerns around the fairness of systems that made assumptions about minoritized people's gender identity or race, though sighted people commonly make such assumptions about passersby (often unconsciously). An automated scene description tool for blind users potentially infringes on bystander privacy by codifying such demographic assumptions; however, the privacy protection is in tension with the end-user's physical safety and awareness of their surroundings based on ample research that suggests there are contexts, like navigating in unfamiliar

areas, where blind people generally want more information about those in their vicinity [2, 19]. Given these complex tensions among the privacy value judgements of various stakeholders and between privacy, safety, and autonomy, a value-sensitive design approach would allow AI system developers to consider how the potentially conflicting stakeholder views of complex values are reflected in their system.

3 A Value-Sensitive Design Approach to the Development of AI Models

Given the existence of tensions between the value of safety and many other stakeholder values when developing systems that use AI, we propose that researchers and practitioners integrate aspects of the Value-Sensitive Design framework into their design processes. A key strength of Value-Sensitive Design is that it offers processes for discovering and resolving these value tensions, or disagreements among stakeholders' value definitions and prioritizations. For example, the *Value Dams and Flows* method from VSD [16], maps value agreements into technical requirements and assesses the severity of different value tensions if a design favors a particular stakeholder's perspective. In this way, different possibilities may be tested to develop models that honor disparate values (e.g., mitigating egregious safety concerns while supporting user autonomy for users with disabilities). Returning to our case study, applying a VSD lens during the design process may lead developers of an AI model to design safety systems that provide explicit controls for application developers to tune the degree of safety filtering to the specific use case of the application the model is integrated into. Given such controls, developers of an organization-sponsored chatbot publically available on the internet may tune the safety system to provide strong safety protections to prevent the chatbot from generating harmful content, whereas developers of AAC devices may relay control over the safety filters to the end user in support of their values of autonomy and free expression.

Similar to its utility for researchers and developers of AI systems, VSD also may scaffold explainability, so users may understand what values guided design decisions, and potentially make adaptations in-line with their own values. However, currently, explainability and accessibility research have rarely intersected, risking that explainability approaches themselves may be inaccessible and therefore not useful for people with disabilities to understand model limitations due to safety concerns. Also, because users with disabilities may be "outliers" in the system [17, 23], it is important that systems are able to explain their behavior even when that behavior has unexpected causes. We argue that explainability techniques are fundamental to the deployment of Value-Sensitive AI systems, since no AI will ever be perfect and no system designer can anticipate nor predetermine how to handle "unknown unknown" tensions that will arise among competing values, even when following a VSD approach. Explainable AI systems provide end-users with information that can empower them to bridge this "last mile" gap in VSD for AI.

Finally, we note that while we have used accessibility as a case study to demonstrate the complexity of various stakeholder values that interact with and are often in tension with the value of safety when designing AI systems, we expect that similar value tensions and the issues they give rise to exist in many other use cases and across many other stakeholder communities. Furthermore, by examining these value tensions, we see how the current expansive definition of safety (e.g., [1]) lacks the nuance necessary to have honest design discussions about tradeoffs between safety and other important values. There may be room to do work on better understanding the relative importance of potential harms for different user groups and scenarios, similar to previous work done to classify types of harm in social media [20].

4 Conclusion

In this paper, we introduced the Value-Sensitive Design framework from the Human-Computer Interaction community as a useful framework for considering how the various stakeholder values and needs of systems built with AI may interact and sometimes conflict. We described the current focus of the AI research community on the safety of AI models, and then drew upon examples from the accessibility literature to demonstrate stakeholder values that come into tension with the current focus on safety, including the values of personal expression, autonomy, and privacy. We proposed that the community of AI practitioners should integrate aspects of the VSD framework into their design process (and incorporate explainability techniques to supplement and support VSD) in order to ensure the systems we develop respect the values of the diverse set of stakeholders of these systems.

References

[1] S. Bengio, A. Beygelzimer, K. Crawford, J. Fromer, I. Gabriel, A. Levendowski, D. Raji, and M. Ranzato. Provisional draft of the neurips code of ethics. *NeurIPS*, May 2021. URL https://openreview.net/pdf?id=zVoy8kAFKPr.

[2] C. L. Bennett, C. Gleason, M. K. Scheuerman, J. P. Bigham, A. Guo, and A. To. "it's complicated": Negotiating accessibility and (mis)representation in image descriptions of race, gender, and disability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445498. URL https://doi.org/10.1145/3411764.3445498.

[3] J. P. Bigham and P. Carrington. Learning from the front: People with disabilities as early adopters of ai. *Proceedings of the 2018 HCIC Human-Computer Interaction Consortium*, 2018.

[4] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, and P. Liang. On the opportunities and risks of foundation models, 2021. URL https://arxiv.org/abs/2108.07258.

[5] S. Cai, S. Venugopalan, K. Tomanek, A. Narayanan, M. R. Morris, and M. P. Brenner. Context-aware abbreviation expansion using large language models, 2022. URL https://arxiv.org/abs/2205.03767.

[6] S. Cole. This ai tool is being used to make freaky, machine-generated porn. https://www.vice.com/en/article/xgygy4/stable-diffusion-stability-ai-nsfw-ai-generated-porn, Aug 2022. (Accessed on 09/23/2022).

[7] E. Dinan, G. Abercrombie, A. S. Bergman, S. Spruit, D. Hovy, Y.-L. Boureau, and V. Rieser. Anticipating safety issues in e2e conversational ai: Framework and tooling, 2021. URL https://arxiv.org/abs/2107.03451.

[8] A. Fiannaca, A. Paradiso, M. Shah, and M. R. Morris. Aacrobat: Using mobile devices to lower communication barriers and provide autonomy with gaze-based aac. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, page 683–695, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450343350. doi: 10.1145/2998181.2998215. URL https://doi.org/10.1145/2998181.2998215.

[9] B. Friedman. Value-sensitive design. *Interactions*, 3(6):16–23, dec 1996. ISSN 1072-5520. doi: 10.1145/242485.242493. URL https://doi.org/10.1145/242485.242493.

[10] B. Friedman. User autonomy: Who should control what and when? a chi 96 workshop. *ACM SIGCHI Bulletin*, 30(1):26–29, 1998.

[11] B. Friedman, P. Kahn, and A. Borning. Value sensitive design: Theory and methods. *University of Washington technical report*, 2:12, 2002.

[12] S. M. Goodman, E. Buehler, P. Clary, A. Coenen, A. Donsbach, T. N. Horne, M. Lahav, R. Macdonald, R. B. Michaels, A. Narayanan, M. Pushkarna, J. Riley, A. Santana, L. Shi, R. Sweeney, P. Weaver, A. Yuan, and M. R. Morris. Lampost: Design and evaluation of an ai-assisted email writing prototype for adults with dyslexia. *arXiv preprint arXiv:2207.02308*, 2022.

[13] K. Hill. A dad took photos of his naked toddler for the doctor. google flagged him as a criminal. - the new york times. https://www.nytimes.com/2022/08/21/technology/google-surveillance-toddler-photo.html, Aug 2022. (Accessed on 09/26/2022).

[14] S. K. Kane, M. R. Morris, A. Paradiso, and J. Campbell. "at times avuncular and cantankerous, with the reflexes of a mongoose": Understanding self-expression through augmentative and alternative communication devices. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, page 1166–1179, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450343350. doi: 10.1145/2998181.2998284. URL https://doi.org/10.1145/2998181.2998284.

[15] J. Mankoff, G. R. Hayes, and D. Kasnitz. Disability studies as a source of critical inquiry for the field of assistive technology. In *Proceedings of the 12th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '10, page 3–10, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605588810. doi: 10.1145/1878803.1878807. URL https://doi.org/10.1145/1878803.1878807.

[16] J. K. Miller, B. Friedman, G. Jancke, and B. Gill. Value tensions in design: The value sensitive design, development, and appropriation of a corporation's groupware system. In *Proceedings of the 2007 International ACM Conference on Supporting Group Work*, GROUP '07, page 281–290, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595938459. doi: 10.1145/1316624.1316668. URL https://doi.org/10.1145/1316624.1316668.

[17] M. R. Morris. Ai and accessibility. *Commun. ACM*, 63(6):35–37, may 2020. ISSN 0001-0782. doi: 10.1145/3356727. URL https://doi.org/10.1145/3356727.

[18] M. R. Nario-Redmond, A. A. Kemerling, and A. Silverman. Hostile, benevolent, and ambivalent ableism: Contemporary manifestations. *Journal of Social Issues*, 75(3):726–756, 2019. doi: https://doi.org/10.1111/josi.12337. URL https://spssi.onlinelibrary.wiley.com/doi/abs/10.1111/josi.12337.

[19] H. Profita, R. Albaghli, L. Findlater, P. Jaeger, and S. K. Kane. The at effect: How disability affects the perceived social acceptability of head-mounted display use. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 4884–4895, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450333627. doi: 10.1145/2858036.2858130. URL https://doi.org/10.1145/2858036.2858130.

[20] M. K. Scheuerman, J. A. Jiang, C. Fiesler, and J. R. Brubaker. A framework of severity for harmful content online. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2), oct 2021. doi: 10.1145/3479512. URL https://doi.org/10.1145/3479512.

[21] J. Shen, B. Yang, J. J. Dudley, and P. O. Kristensson. Kwickchat: A multi-turn dialogue system for aac using context-aware sentence generation by bag-of-keywords. In *27th International Conference on Intelligent User Interfaces*, IUI '22, page 853–867, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391443. doi: 10.1145/3490099.3511145. URL https://doi.org/10.1145/3490099.3511145.

[22] E. S. Tanis and C. Lewis. Artificial intelligence and the dignity of risk. *SIGACCESS Access. Comput.*, 125 (7), mar 2020. ISSN 1558-2337. doi: 10.1145/3386296.3386303. URL https://doi.org/10.1145/3386296.3386303.

[23] S. Trewin. Ai fairness for people with disabilities: Point of view, 2018. URL https://arxiv.org/abs/1811.10670.

[24] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

[25] K. Wiggers. This startup is setting a dall-e 2-like ai free, consequences be damned | techcrunch. https://techcrunch.com/2022/08/12/a-startup-wants-to-democratize-the-tech-behind-dall-e-2-consequences-be-damned/, Aug 2022. (Accessed on 09/26/2022).