Troubling Taxonomies in GenAI Evaluation

Anonymous Author(s)

1 Models of the social world in GenAI evaluations

Recent scholarship has started to make explicit the normative values and commitments of GenAI and Machine Learning (ML) practices [1] and evaluation [2]. In this provocation, we extend this line of inquiry by arguing that we need to attend to the implicit values and assumptions reflected in how *societal impacts*[1] are conceptualised and constructed through ML evaluations. Doing so reveals that the work of assessing and managing societal impacts of GenAI is best conceptualised as a governance challenge.

Evaluating GenAI's societal impacts requires a model of how these impacts manifest [3]. This model, often implicit, enables interpretation of how GenAI systems interact with people and social structures; the model constructs particular social factors as capable, and deserving, of measurement. We ask: what is the model of societal impacts reflected in existing efforts to evaluate GenAI systems? One avenue to understand this model is to look at taxonomies of societal impacts [e.g., 4–6] which provide conceptual infrastructure for societal impact evaluations of GenAI [for an alternate approach see 7].

2 Taxonomies of societal impacts

Taxonomies of societal impacts exist for a range of GenAI technologies and components, including foundation models [8], text-to-image models [9], large language models [6], speech generation models [10], AI agents [11], and GenAI or algorithmic systems more generally [4, 5]. These tools enable researchers and practitioners to think systematically about the potential consequences of deploying GenAI technologies. In doing so, taxonomies direct the attention of GenAI evaluators, and provide structure for GenAI evaluations. Taxonomy development is, therefore, ontological work that has far-reaching consequences for the way researchers and practitioners understand the relationship between GenAI technologies, people, and society [12, 13]. Indeed, the development of bespoke taxonomies for different GenAI technologies implies a conceptualisation of societal impacts that centres technology as the primary causal determinate of harms and positions technology developers as the critical actors in impact evaluations (for contrast, imagine a range of taxonomies for different social contexts).

One way in which taxonomies of societal impacts direct attention is by navigating the tension between abstraction and contextualisation, which is present throughout AI development [14]. GenAI components tend to be understood within an abstract space that disregards the social context of their development [3]. Social harms, meanwhile, are understood as being deployment, i.e. context, dependent [e.g., 4]. Shelby et al. [5], in their taxonomy, attempt to navigate this tension by distinguishing between harms that originate with computational components of AI systems, and harms that originate in their deployment. Yet, GenAI systems (e.g., ChatGPT) are deployed extremely broadly, cutting across vast swathes of different social contexts. In this context, taxonomies direct attention towards forms of harm that can be tested or detected at the abstract level of the GenAI model, within the broader software evaluation paradigm in which GenAI deployment occurs [e.g., GPT4, 15].

Taxonomies of societal impacts also direct attention towards prediction and trade-offs, centering the site of GenAI development rather than application or deployment contexts. Taxonomies are often framed as predictive tools, enabling practitioners to forecast risks of GenAI deployment [e.g., 6, 9]. Yet, in attempting to develop an exhaustive schema of potential impacts, taxonomies invite comparison, and trade-offs, across disparate categories of societal impacts [16], such as "trust in media and information", "community erasure", and "intellectual property and ownership" [4]. Implicit in this

[1] Societal 'impacts' is the phrasing generally adopted in responsible AI literature, which we follow. 'Impact', however, is suggestive of immediacy and collisions. Social 'outcomes' may be preferable. This encourages thinking about long term and second- and third-order outcomes of introducing AI systems into society.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

organisation is a conceptualisation of societal impacts as modular, independent, and commensurable, with GenAI developers positioned as arbitrators in determining which impacts to address, and how.

3 Conceptualising the societal impacts of GenAI

A conceptualisation of societal impacts that centres GenAI technologies and GenAI developers may be useful, in terms of producing a discourse on societal impacts and risks that is tractable within GenAI, but should be approached with caution. Critical questions to consider include: what factors should be centered when thinking about GenAI's societal impacts? what are the limits of societal impact prediction? how should evaluators balance different forms of societal impacts? To begin responding to these questions, we offer three premises for rethinking the relationship between GenAI, people, and society.

Societal impacts should be understood as application- and context-specific [17, 3] and indeterminate [18]. Failure to do so produces an understanding of societal impacts that is universalising and self-fulfilling; the work of evaluating societal impacts becomes the work of extending patterns of social relations from one place to many [19]. Societal impacts of GenAI should be thought of at the system level, with the GenAI system situated in a particular social context [cf., regarding model explanations, 20]. Impacts manifest when a model is integrated into a sociotechnical system, and implemented in a specific social setting [5]. Context-specificity makes predictions about societal impacts difficult and unverifiable. Consequently, taxonomies of societal impacts are inherently partial, always incomplete.

Some societal impacts should be understood as incommensurable [21]. The scaling of large multilingual models to include many languages, including Indigenous languages, illustrates this phenomenon [22, 23]. Such scaling is motivated by the assumption that language technologies should be accessible to everyone in their first language [24], which leads to model evaluations focused on identifying and rectifying performance disparities across languages. Yet, how should evaluators reconcile issues of disparate performance with issues of Indigenous data sovereignty, given one strategy to improve GenAI performance is to collect more language data? Navigating these trade-offs is particularly problematic in situations where the objectives of GenAI developers may diverge from those of local communities. Some Australian Aboriginal and Māori communities prioritise managing cultural knowledge, including language data, to support intergenerational transmission rather than expanding access to language technology [25, 26]. In contexts like these, while taxonomies of societal impacts can help GenAI practitioners identify a broad range of impacts, GenAI practitioners are not well-positioned to balance competing impacts–these are value-laden decisions that require community leadership. Taxonomies can support such leadership by enabling practitioners to identify relevant stakeholders associated with different societal impacts [9].

Finally, questions of societal impacts are questions about social power. Taxonomies of societal impacts enable evaluators to decide what to include (and exclude) in their evaluations. This legitimises particular concerns or forms of impact as salient to GenAI. As evaluation practices mature and become standardised, they gain efficacy, in terms of their capacity to enforce the values and assumptions they reflect [27]. The dominance of cost-benefit analysis in environmental impact evaluation, for example, supports a capitalist and extractive epistemology, in which the worth of the environment is expressed in monetary terms [28]. Efforts to standardise societal impact evaluation—worthy as they are—should therefore be understood as sociopolitical efforts that can reify, or resist, particular social orders. Who determines which societal impacts to focus on matters.

4 For a governance-first approach

The conceptualisation of societal impacts sketched above suggests a redirection of efforts, away from evaluations of potential harms, and towards a governance-first approach to GenAI evaluation. If societal harms are contextually contingent and indeterminate, then anticipatory evaluations may not be as effective at identifying and mitigating impacts as robust governance and monitoring of GenAI deployment led by stakeholders or governments. Reflecting this, a governance-first approach would demand accountability to, participation of, and deliberation within, stakeholders or communities impacted by GenAI deployments [29]–for example, to determine how to balance disparate impacts. Taxonomies and other evaluation tools can serve as useful inputs to robust governance and accountability processes [12]. However, without first establishing sustainable, representative governance structures (or engaging with those that already exist) these tools risk generalising predictions of harms across diverging contexts, equating incommensurable impacts, and ultimately serving the interests of GenAI researchers and developers rather than affected communities.

References

[1] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. The Values Encoded in Machine Learning Research. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 173–184, Seoul Republic of Korea, June 2022. ACM. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533083.

[2] Ben Hutchinson, Negar Rostamzadeh, Christina Greer, Katherine Heller, and Vinodkumar Prabhakaran. Evaluation Gaps in Machine Learning Practice. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1859–1876, Seoul Republic of Korea, June 2022. ACM. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533233.

[3] Donald Martin, Vinodkumar Prabhakaran, Jill Kuhlberg, Andrew Smart, and William S. Isaac. Extending the Machine Learning Abstraction Boundary: A Complex Systems Approach to Incorporate Societal Context. *arXiv preprint arXiv:2006.09663*, 2020. doi: 10.48550/ARXIV.2006.09663.

[4] Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Canyu Chen, Hal Daumé, Jesse Dodge, Isabella Duan, Ellie Evans, Felix Friedrich, Avijit Ghosh, Usman Gohar, Sara Hooker, Yacine Jernite, Ria Kalluri, Alberto Lusoli, Alina Leidinger, Michelle Lin, Xiuzhu Lin, Sasha Luccioni, Jennifer Mickel, Margaret Mitchell, Jessica Newman, Anaelia Ovalle, Marie-Therese Png, Shubham Singh, Andrew Strait, Lukas Struppek, and Arjun Subramonian. Evaluating the Social Impact of Generative AI Systems in Systems and Society, 2023.

[5] Renee Shelby, Shalaleh Rismani, Kathryn Henne, Ajung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 723–741, Montreal QC Canada, August 2023. ACM. ISBN 9798400702310. doi: 10.1145/3600211.3604673.

[6] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Taxonomy of Risks posed by Language Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 214–229, Seoul Republic of Korea, June 2022. ACM. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533088. URL `https://dl.acm.org/doi/10.1145/3531146.3533088`.

[7] Gloire Rubambiza, Phoebe Sengers, Hakim Weatherspoon, and Jen Liu. Seam Work and Simulacra of Societal Impact in Networking Research: A Critical Technical Practice Approach. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–19. ACM, 2024. ISBN 9798400703300. doi: 10.1145/3613904.3642337.

[8] Andrés Domínguez Hernández, Shyam Krishna, Antonella Maia Perini, Michael Katell, Sj Bennett, Ann Borda, Youmna Hashem, Semeli Hadjiloizou, Sabeehah Mahomed, Smera Jayadeva, Mhairi Aitken, and David Leslie. Mapping the individual, social and biospheric impacts of Foundation Models. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, volume 7, pages 776–796, Rio de Janeiro Brazil, June 2024. ACM. doi: 10.1145/3630106.3658939. URL `https://dl.acm.org/doi/10.1145/3630106.3658939`.

[9] Charlotte Bird, Eddie Ungless, and Atoosa Kasirzadeh. Typology of Risks of Generative Text-to-Image Models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 396–410, Montr\'{e}al QC Canada, August 2023. ACM. ISBN 9798400702310. doi: 10.1145/3600211.3604722. URL `https://dl.acm.org/doi/10.1145/3600211.3604722`.

[10] Wiebke Hutiri, Orestis Papakyriakopoulos, and Alice Xiang. Not My Voice! A Taxonomy of Ethical and Safety Harms of Speech Generators. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 359–376, Rio de Janeiro Brazil, June 2024. ACM. doi: 10.1145/3630106.3658911. URL `https://dl.acm.org/doi/10.1145/3630106.3658911`.

[11] Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, Michelle Lin, Alex Mayhew, Katherine Collins, Maryam Molamohammadi, John Burden, Wanru Zhao, Shalaleh Rismani, Konstantinos Voudouris, Umang Bhatt, Adrian Weller, David Krueger, and Tegan Maharaj. Harms from Increasingly Agentic Algorithmic Systems. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 651–666, Chicago IL USA, June 2023. ACM. ISBN 9798400701924. doi: 10.1145/3593013.3594033. URL https://dl.acm.org/doi/10.1145/3593013.3594033.

[12] Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf. Assembling Accountability: Algorithmic Impact Assessment for the Public Interest, 2021. URL https://datasociety.net/wp-content/uploads/2021/06/Assembling-Accountability.pdf.

[13] Geoffrey C. Bowker and Susan Leigh Star. *Sorting Things out: Classification and Its Consequences*. MIT Press, 2000. ISBN 0-262-52295-0.

[14] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, pages 59–68, 2019. ISBN 978-1-4503-6125-5. doi: 10.1145/3287560.3287598.

[15] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and ...Barret Zoph. GPT-4 Technical Report, March 2024. URL http://arxiv.org/abs/2303.08774. arXiv:2303.08774 [cs].

[16] Mary Douglas and Aaron Wildavsky. *Risk and Culture: An Essay on the Selection of Technological and Environmental Dangers*. University of California Press, 1983. ISBN 978-0-520-90739-3. doi: 10.1525/9780520907393.

[17] Lucy Suchman, Jeanette Blomberg, Julian E. Orr, and Randall Trigg. Reconstructing Technologies as Social Practice. *American Behavioral Scientist*, 43(3):392–408, November 1999. ISSN 0002-7642, 1552-3381. doi: 10.1177/00027649921955335.

[18] Brian Wynne. Uncertainty and environmental learning: Reconceiving science and policy in the preventive paradigm. *Global Environmental Change*, 2(2):111–127, July 1992. ISSN 09593780. doi: 10.1016/0959-3780(92)90017-2.

[19] Donna Haraway. Situated knowledges: The science question in feminism and the privilege of partial perspective. *Fem. Stud.*, 14(3):575–599, 1988. ISSN 0046-3663. doi: 10.2307/3178066.

[20] Andrew Smart and Atoosa Kasirzadeh. Beyond model interpretability: socio-structural explanations in machine learning. *AI & SOCIETY*, September 2024. ISSN 0951-5666, 1435-5655. doi: 10.1007/s00146-024-02056-1. URL https://link.springer.com/10.1007/s00146-024-02056-1.

[21] Nicolas Espinoza. Incommensurability: The Failure to Compare Risks. In *The ethics of technological risk*. Earthscan London, 2009. URL https://api.taylorfrancis.com/content/books/mono/download?identifierName=doi&identifierValue=10.4324/9781849772990&type=googlepdf.

[22] Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. Scaling speech technology to 1,000+ languages. *arXiv [cs.CL]*, May 2023. URL http://arxiv.org/abs/2305.13516.

[23] Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 1082–1117, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.61. URL https://aclanthology.org/2023.acl-long.61.

[24] Steven Bird. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.313. URL http://dx.doi.org/10.18653/v1/2020.coling-main.313.

[25] Ned Cooper, Courtney Heldreth, and Ben Hutchinson. "It's how you do things that matters": Attending to process to better serve indigenous communities with language technologies. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 204–211, 2024. URL https://aclanthology.org/2024.eacl-short.19.pdf.

[26] Te Mana Raraunga. Our charter. https://www.temanararaunga.maori.nz/tutohinga, April 2016. URL https://www.temanararaunga.maori.nz/tutohinga. Accessed: 2023-6-16.

[27] Michèle Lamont. Toward a comparative sociology of valuation and evaluation. *Annual review of sociology*, 38:201–221, 2012. doi: 10.1146/annurev-soc-070308-120022.

[28] Langdon Winner. *The Whale and the Reactor: The Search for Limits in a Technological Age*. University of Chicago Press, Chicago, 1986.

[29] Bogdana Rakova and Roel Dobbe. Algorithms as social-ecological-technological systems: an environmental justice lens on algorithmic audits. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 491, New York, NY, USA, June 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594014. URL https://doi.org/10.1145/3593013.3594014.