Using Scenario-Writing for Identifying and Mitigating Impacts of Generative AI

Kimon Kieslich
University of Amsterdam
k.kieslich@uva.nl

Nicholas Diakopoulos
Northwestern University
nad@northwestern.edu

Natali Helberger
University of Amsterdam
n.helberger@uva.nl

Abstract

Impact assessments have emerged as a common way to identify the negative and positive implications of AI deployment, with the goal of avoiding the downsides of its use. It is undeniable that impact assessments are important - especially in the case of rapidly proliferating technologies such as generative AI. But it is also essential to critically interrogate the current literature and practice on impact assessment, to identify its shortcomings, and to develop new approaches that are responsive to these limitations. In this provocation, we do just that by first critiquing the current impact assessment literature and then proposing a novel approach that addresses our concerns: Scenario-Based Sociotechnical Envisioning.

1 A Critique of Current Impact Assessment Methods

Power. *Who* is in charge of identifying and managing impacts? In other words, what are the underlying power relations in impact assessment? Impact assessment is political, and different impacts are prioritized according to the goals and social or economic priorities of the entity conducting the assessment. Whoever decides, also influences which impacts to look at, which values to prioritize, and how value conflicts should be resolved. Scholars have criticized the current regulatory and practical landscape, stating that most existing impact assessments ought to be performed by (technical) experts, regulators, technologists and academics, and citizen representatives, leaving out perceptions of ordinary citizens, and/or marginalized voices [9, 31]. A look at the regulatory landscape, for instance the EU AI Act and the Digital Service Act (DSA), underlines this criticism: These regulations, which aim to strike a balance between innovation and regulation, place the responsibility for assessing "reasonably foreseeable risks" largely in the hands of technology providers [9, 11, 24]. This could lead to the issue “that corporations selectively focus on those risks and mitigation measures that are least disruptive for their business goals” [13].

Inclusion. Most current impact assessments rely either on literature review of scholarly literature [3, 16, 28, 29, 33], corporate authored reports [33, 34], or expert judgments [6, 30]. Yet, research has shown that expert judgments can be biased and fail to recognize impacts that lie outside their lived experience [23], and that "individuals and communities affected by algorithmic systems are often the foremost experts in the potential harms they regularly encounter" [21]. This critique is particularly relevant to general purpose technologies, as end-users (e.g., through their usage patterns) will strongly influence the impact of these technologies. Similarly, impacts are perceived differently by different affected groups as their contexts and lived realities differ. Thus, the diversity of perspectives as well as the contextual embedding of a technology matter for a comprehensive and inclusive impact assessment [25]. Although the active involvement of affected people and communities is recognized by many scholars in toolkits [19], frameworks [21, 22, 27] and studies [10], there is still a lack of implementation on a larger practical scale and in regulation [10].

Quantification. Another critical issue is the heavy reliance on metrics, i.e. quantifiable measures of impact. Some impacts are simply easier to measure than others, as they can be traced to (material)

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

objects that can be scaled and quantified [13, 26]. Other impacts, however, are much more difficult and costly to measure (e.g. human rights). While there are checklists and scales that attempt to measure such impacts, they ultimately depend on the judgment of the assessor and their embedding in existing power structures. In addition, many impacts only become visible in the socio-technical interaction between humans and generative AI. Thus, in the quest for quantification, complex phenomena are simplified, taken out of context, or data may be incomplete [12, 24, 32]. Thus, the reliance on only quantitative metrics is insufficient to account for the complexity of the social world [24].

Anticipating Unknowns. An important distinction in the literature on impact assessment is between known and unknown [12]. But regulations such as the DSA or the AI Act, only mandates the identification of "known" and "reasonably foreseeable" risks. Accordingly, most impact assessments rely on established measures, but fail to identify potential impacts that are not yet known. Anticipating unknown, future impacts becomes less a matter of evaluating quantifiable evidence and more a matter of thinking in terms of different possible future scenarios. This shift from identifying known impacts to anticipating future impacts also has important implications for the mitigation of impacts: Instead of designing interventions to "fix" known impacts, anticipatory risk management becomes a choice between more and less desirable futures, and identifying the responsible actors and intervention points needed to realize one and move away from the other.

2 Scenario-based sociotechnical envisioning

In response to the shortcomings of current impact assessments for generative AI, we have developed a novel approach: *Scenario-based Sociotechnical Envisioning (SSE)* (see Table 1 for an overview of our contribution). SSE involves three elements. Following the scenario design and planning literature [1, 7, 8] (1) *Scenario-based* refers to the emphasis on written narratives: these narratives are the core outcome of applying the method, though downstream analysis methods can also develop metrics based on these descriptions. Narratives make complex issues visible to the assessor and thus emphasize individual contextualized experiences. (2) *Sociotechnical* emphasizes the need to assess the human interaction of people and technology and responds to the criticism of the lack of context in existing impact assessment frameworks. Especially in generative AI, the sociotechnical part plays a huge role, as users have endless options for prompting generative AI – which also entails a huge variety of different impacts. (3) The *envisioning* element emphasizes the need to think prospectively and identify potential impacts that are not yet known. Based on anticipatory governance [5, 14], this aspect highlights the need to illuminate future pathways of technological interactions with individuals and society. SSE can be applied in a survey or workshop setting. The core of SSE is a writing task that participants are asked to complete on their own. SSE explicitly aims for an inclusive sampling, including stakeholders of varied expertise and background such as technology developers or professional deployers of the technology, but also end-users (incl. marginalized groups). Participants of SSE studies are introduced to the technology, including information about the capabilities, limitations, and trends of the technology, and quality criteria for a well-written story. Participants will then write a story that outlines their projection of the future impacts of generative AI and are also asked to *reflect* on the story they wrote. For instance, respondents can be asked to elaborate on their value beliefs underlying the identification of impacts, but also to develop mitigation strategies, or evaluate the effectiveness of existing mitigation strategies such as policy proposals.

The resulting stories can stand alone as future projections of human-machine interactions and their implications. They can be informative for scientists, policy makers, or auditors because they contain readily accessible and vivid imaginings of possible future pathways. These narratives can be powerful by themselves because they convey values, reflect identities, and trigger agency [20]. By collecting a large number of scenarios, SSE can also be used to develop a socio-technical impact assessment framework that complements the existing literature. Using qualitative thematic analysis and axial coding, the impacts outlined in the stories can be identified; this step also entails the identification of previously under-represented or even unknown impacts. These impacts can further be quantified by counting the mention of impacts – however, in retaining quotes from the stories, the illustrative character of the stories is retained. Depending on the granularity of the impact assessment framework and the sample size, SSE also allows to statistically trace specific risk perceptions back to respondents' sociodemographic characteristics or attitudes. It is also possible to sample for different expertise and then compare the emerging risk frameworks with the lived experience of these groups. We developed this approach over the course of three empirical, peer-reviewed papers [2, 17, 18].

Limitations

Structural. SSE does not produce metrics that can be standardized. While academics argue that there is a need for more qualitative methods to enrich impact assessment and management, industry and standards bodies prefer to rely on fixed, quantifiable metrics that give them more power to define risks themselves and to operate cost-effectively [9, 11]. Relying on the good will of business alone won't be enough to bring SSE to scale. It will require investment from policymakers who mandate that methodologies like SSE be an integral part of impact assessment and management processes.

Resources. Depending on the purpose of SSE, it can be costly. Particularly when used in a survey design, users of SSE will need to pay respondents for their participation in the study. In large-scale designs - such as those needed to conduct impact assessments - the sample size must be large enough to ensure a holistic assessment of impact. In addition, users will need to account for the time spent by SSE users in analyzing the scenarios. Though, we believe that investment in rigor methods to protect citizens from detrimental impacts of generative AI must not necessarily be cheap - first and foremost, they should safeguard affected people.

Inclusion and expert bias. SSE requires its users to make a variety of decisions prior to application. (1) Whose narratives should be assessed with SSE? In principle, SSE can be conducted with a wide range of stakeholders, including marginalized groups. Highlighting voices from the margins is particularly important given the detrimental effects that generative AI can have on these groups. However, when conducting SSE with marginalized groups, it is important to engage in equitable participatory design, that is, to fully acknowledge the lived experiences and contributions of these communities [15]. When engaging with marginalized groups, it is important to build trust, understand their historical context and allow for alternative solutions (e.g., mitigation strategies) [15]. Therefore, when explicitly applying SSE with these groups, it is recommended that multiple researchers and collaborators with different cultural backgrounds are involved in the study design and data analysis. (2) While SSE follows the idea of participatory AI to make the lived experiences of communities visible and to bring them into the broader discussion of AI impact assessment, the approach can still leave communities out: For example, some marginalized groups may not be able or willing to engage in SSE [4]. (3) Scenarios are biased towards the instructions that SSE users give to respondents. SSE users need to think carefully about what and how much information they provide to participants. A bias in the instruction material can lead to an equally biased thematization of impacts.

Broader Impact Statement

SSE is an approach that responds to calls for new perspectives and methodologies that go beyond quantifiable impact metrics to identify previously overlooked issues and amplify the perspectives of typically underrepresented populations. We hope to spark a lively discussion about the limitations of current impact assessments, and to highlight the need for more qualitative impact assessments for generative AI technologies. We note that we don't aim to replace current impact assessment, but rather to enrich the current landscape with alternative approaches that aim to capture previously overlooked impacts and illuminate the contextual nature of impacts.

SSE can be a cornerstone to enrich current impact assessment practices. We argue for its use to uncover unknown future impacts of generative AI technology. We also highlight its potential to uncover the lived experiences of affected groups. The application of SSE on a larger scale can make impact assessment practices more contextualized and inclusive, and help to provide mitigation strategies that are based on the needs of affected groups and communities. In this way, SSE contributes to good AI practice.

Funding

The funding for this research was provided by UL Research Institutes through the Center for Advancing Safety of Machine Intelligence.

References

[1] Muhammad Amer, Tugrul U. Daim, and Antonie Jetter. 2013. A review of scenario planning. *Futures : the journal of policy, planning and futures studies* 46 (Feb. 2013), 23–40. https://doi.org/10.1016/j.futures.2012.10.003 ISBN: 0016-3287.

[2] Julia Barnett, Kimon Kieslich, and Nicholas Diakopoulos. 2024. Simulating Policy Impacts: Developing a Generative Scenario Writing Method to Evaluate the Perceived Effects of Regulation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. 82–93.

[3] Charlotte Bird, Eddie L. Ungless, and Atoosa Kasirzadeh. 2023. Typology of Risks of Generative Text-to-Image Models. http://arxiv.org/abs/2307.05543 arXiv:2307.05543 [cs].

[4] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the People? Opportunities and Challenges for Participatory AI. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. ACM, Arlington VA USA, 1–8. https://doi.org/10.1145/3551624.3555290

[5] Philip A. E. Brey. 2012. Anticipatory Ethics for Emerging Technologies. *Nanoethics* 6, 1 (April 2012), 1–13. https://doi.org/10.1007/s11569-012-0141-7 ISBN: 1871-4757.

[6] Zana Buçinca, Chau Minh Pham, Maurice Jakesch, Marco Tulio Ribeiro, Alexandra Olteanu, and Saleema Amershi. 2023. AHA!: Facilitating AI Impact Assessment by Generating Examples of Harms. http://arxiv.org/abs/2306.03280 arXiv:2306.03280 [cs].

[7] Lena Börjeson, Mattias Höjer, Karl-Henrik Dreborg, Tomas Ekvall, and Göran Finnveden. 2006. Scenario types and techniques: Towards a user's guide. *Futures* 38, 7 (2006), 723–739. https://doi.org/10.1016/j.futures.2005.12.002 ISBN: 0016-3287.

[8] John M Carroll. 2003. *Making use: scenario-based design of human-computer interactions*. MIT press.

[9] Julie E Cohen and Ari Azra Waldman. 2023. Introduction: Framing Regulatory Managerialism as an Object of Study and Strategic Displacement. *Law & Contemp. Probs.* 86 (2023), i. Publisher: HeinOnline.

[10] Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini. 2022. Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 1571–1583. https://doi.org/10.1145/3531146.3533213

[11] Luciano Floridi and Andrew Strait. 2020. Ethical Foresight Analysis: What it is and Why it is Needed? *Minds and Machines* 30, 1 (March 2020), 77–97. https://doi.org/10.1007/s11023-020-09521-y

[12] Raphaël Gellert. 2020. *The Risk-Based Approach to Data Protection* (1 ed.). Oxford University PressOxford. https://doi.org/10.1093/oso/9780198837718.001.0001

[13] Rachel Griffin. 2024. What do we talk about when we talk about risk? Risk politics in the EU's Digital Services Act. https://dsa-observatory.eu/2024/07/31/what-do-we-talk-about-when-we-talk-about-risk-risk-politics-in-the-eus-digital-services-act/

[14] David H. Guston. 2013. Understanding 'anticipatory governance'. *Social Studies of Science* 44, 2 (2013), 218–242. https://doi.org/10.1177/0306312713508669

[15] Christina Harrington, Sheena Erete, and Anne Marie Piper. 2019. Deconstructing Community-Based Collaborative Design: Towards More Equitable Participatory Design Engagements. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–25. https://doi.org/10.1145/3359318

[16] Mia Hoffmann and Heather Frase. 2023. *Adding Structure to AI Harm*. Technical Report. Center for Security and Emerging Technology. https://cset.georgetown.edu/publication/adding-structure-to-ai-harm/

[17] Kimon Kieslich, Nicholas Diakopoulos, and Natali Helberger. 2024. Anticipating impacts: using large-scale scenario-writing to explore diverse implications of generative AI in the news environment. *AI and Ethics* (May 2024). https://doi.org/10.1007/s43681-024-00497-4

[18] Kimon Kieslich, Natali Helberger, and Nicholas Diakopoulos. 2024. My Future with My Chatbot: A Scenario-Driven, User-Centric Approach to Anticipating AI Impacts. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 2071–2085. https://doi.org/10.1145/3630106.3659026

[19] Tobias D. Krafft, Katharina A. Zweig, and Pascal D. König. 2022. How to regulate algorithmic decision-making: A framework of regulatory requirements for different applications. *Regulation & Governance* 16, 1 (Jan. 2022), 119–136. https://doi.org/10.1111/rego.12369

[20] Vivien Lowndes. 2016. Narrative and storytelling. In *Evidence-Based Policy Making in the Social Sciences* (1 ed.), Gerry Stoker and Mark Evans (Eds.). Bristol University Press, 103–122. https://doi.org/10.46692/9781447329381.007

[21] Jacob Metcalf, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish. 2021. Algorithmic impact assessments and accountability: the co-construction of impacts. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021). https://doi.org/10.1145/3442188.3445935

[22] Emanuel Moss, Elizabeth Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf. 2021. Assembling Accountability: Algorithmic Impact Assessment for the Public Interest. *SSRN Electronic Journal* (2021). https://doi.org/10.2139/ssrn.3877437

[23] Priyanka Nanayakkara, Jessica Hullman, and Nicholas Diakopoulos. 2021. Unpacking the expressed consequences of AI research in broader impact statements. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 795–806.

[24] Frank Pasquale. 2023. Power and Knowledge in Policy Evaluation: From Managing Budgets to Analyzing Scenarios. *Law and Contemporary Problems* 86, 3 (2023).

[25] P Marijn Poortvliet, Martijn Duineveld, and Kai Purnhagen. 2016. Performativity in action: How risk communication interacts in risk regulation. *European Journal of Risk Regulation* 7, 1 (2016), 213–217. Publisher: Cambridge University Press.

[26] Michael Power. 2004. *The risk management of everything: rethinking the politics of uncertainty* (1. publ ed.). Demos, London.

[27] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, Barcelona Spain, 33–44. https://doi.org/10.1145/3351095.3372873

[28] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. 2023. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. http://arxiv.org/abs/2210.05791 arXiv:2210.05791 [cs].

[29] Peter Slattery, Alexander K Saeri, Emily A C Grundy, Jess Graham, Michael Noetel, Risto Uuk, James Dao, Soroush Pour, Stephen Casper, and Neil Thompson. 2024. The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence. (2024). https://doi.org/10.13140/RG.2.2.28850.00968 Publisher: Unpublished.

[30] Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Hal Daumé III, Jesse Dodge, Ellie Evans, Sara Hooker, Yacine Jernite, Alexandra Sasha Luccioni, Alberto Lusoli, Margaret Mitchell, Jessica Newman, Marie-Therese Png, Andrew Strait, and Apostol Vassilev. 2023. Evaluating the Social Impact of Generative AI Systems in Systems and Society. http://arxiv.org/abs/2306.05949 arXiv:2306.05949 [cs].

[31] Bernd Carsten Stahl, Josephina Antoniou, Nitika Bhalla, Laurence Brooks, Philip Jansen, Blerta Lindqvist, Alexey Kirichenko, Samuel Marchal, Rowena Rodrigues, Nicole Santiago, Zuzanna Warso, and David Wright. 2023. A systematic review of artificial intelligence impact assessments. *Artificial Intelligence Review* (March 2023). https://doi.org/10.1007/s10462-023-10420-8

[32] Jeroen van der Heijden. 2021. Risk as an approach to regulatory governance: An evidence synthesis and research agenda. *Sage Open* 11, 3 (2021), 21582440211032202. Publisher: SAGE Publications Sage CA: Los Angeles, CA.

[33] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from Language Models. http://arxiv.org/pdf/2112.04359v1http://arxiv.org/abs/2112.04359v1https://arxiv.org/pdf/2112.04359v1.pdf

[34] Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, and William Isaac. 2023. Sociotechnical Safety Evaluation of Generative AI Systems. http://arxiv.org/abs/2310.11986 arXiv:2310.11986 [cs].

Appendix

Table 1: SSE Solution for Shortcomings of Current IAs

Shortcoming of IAs	SSE Solution	Description
Depoliticization; Lack of democratic accountability (Power)	Politicization; Strengthening of democratic accountability	Politicization leads to a stronger demand for accountability. Inclusion of multiple stakeholder (incl. laypersons) creates more democratic accountability.
Lack of inclusion; Expert-biased	Inclusive sampling; Facilitating Diversity	Inclusive sampling leads to the amplification and visibility of different voices. Gauging the subjectivity of many affected populations and acknowledging expertism of different crowds.
Focus on quantification	Keeping qualitative character of impacts	Impacts are identified and qualitatively described. Narratives convey meaning, sociocultural contexts and make compelling arguments that cannot be expressed in numbers.
Exclusion of unknown unknowns	Identification of unknown unknowns	AI technologies might entail novel risks that are yet unknown in common assessment frameworks. SSE can reveal and add them to existing IAs.