CAIDP Paper Submission
Answering to the prompt: Implementation of the Universal Guidelines on AI
Author: Cornelia Evers
Wordcount: 1563 words

The launch of Open AI's Chatgpt in autumn 2022 along with a plethora of concurring generative AI large language models (LLMs) intensify regulatory attention of lawmakers around the world. They grapple to develop AI governance models that align such systems with democratic principles such as accountability, due process, or transparency. Whilst algorithmic harms such as discriminatory and disparate impacts on marginalised communities have been exposed by scholars and activists since the mid-2010s, LLMs such as Chatgpt contribute to mainstreaming the visibility of these harms across broad audiences (see Eubanks, 2018, O'Neil, 2016, Benjamin, 2019). This widespread public attention accelerates the pressure on AI actors to affront ethical shortcomings associated with wide-scale AI deployment as the path of unfettered innovation reveals its limits. Important progress is underway, with a proliferation of 'ethical' or 'responsible' AI principles endorsed by public and private AI actors alike since the late 2010s. Such frameworks pioneered by the G20's or the OECD's AI Principles provide non-legally binding requirements such as "human-centred values and fairness" or "transparency and explainability" to guide how organisations develop and deploy AI systems across their lifecycle (OECD, 2019).

However, the processes necessary to enable such achievements depend on the consistent advocacy and civic engagement of civil society organisations such as the Center for AI and Digital Policy. Through fostering public discourse, developing policy commentary and meaningful frameworks such as the Universal Guidelines on AI (UGAI), they play a key role in aligning AI with ethical values. The UGAI's deserve particular merit for the role they play in consolidating such frameworks in national law and guidelines through tireless advocacy and exchange with multidisciplinary domain experts. Its twelve guidelines reflect the most important key principles for ensuring an ethical AI lifecycle. Most requirements mirror the core themes developed in comparable AI principles. Nevertheless, its requirement on "assessment and accountability" distinguishes it from other AI principles. Most documents allude to the principle of accountability, yet without linking it to an *ex-ante* assessment requirement as the UGAI do. Maintaining that "an AI system should be deployed only after an

adequate evaluation of its purpose and objectives, its benefits, as well as its risks" implies that some form of risk- or impact assessment shall be performed prior to AI deployment (CAIDP, 2018). The CAIDP clarifies that this might mean not developing an AI system at all if the risks "concerning Public Safety and Cybersecurity" are deemed to be too high (ibid.). This notion of *ex-ante* assessment and the possibility of halting innovation for the sake of principles deemed to trump the projected benefits associated with AI development are not revolutionary, yet more often proclaimed than practiced. This may change thanks to the new wave of AI policy which focuses on ensuring compliance with requirements such as the UGAIs through risk-based algorithmic impact assessment frameworks. Recent examples such as UNESCO's Ethical AI Impact Assessment or Canada's compulsory Algorithmic Impact Assessment Tool show how policymakers envision such increased accountability measures (Government of Canada, 2023; UNESCO, 2023). Amidst this trend, the UGAI's "assessment and accountability" requirement can be seen as a precursor to this development.

This paper illuminates in how far the requirement may serve as a basis for algorithmic impact assessments, particularly in fostering a culture of ethical reflection points across the AI lifecycle and crucial points that need to be considered in this emerging standard setting process. I first focus on how *ex-ante* assessments must effectively include possibilities of refusal, which may raise tensions with economic imperatives various AI actors prioritise. Second, I emphasise the importance of making the ongoing standard setting process open and participatory to ensure that impact assessments are not regarded as a silver bullet compliance tool, more focused on checking boxes and calculating risks rather than confronting the material impacts of algorithmic harms.

Impact Assessments are defined as structured processes "for considering the implications, for people and their environment, of proposed actions, whilst there still an opportunity to modify (or even, if appropriate, abandon) the proposals" (IAIA as quoted by Stahl et al., 2023). In the context of AI, this means to assess the potential impacts of an AI system on a variety of fields including human rights, ethics, data protection, safety and cybersecurity and environmental impact (ibid.).

The here presented idea of performing the assessment at a moment where it is still possible to modify a planned action (e.g., development or deployment of an AI system) matches the “assessment and accountability” requirement as expressed in the UGAI. Whilst being able to abandon such plans upon shortcomings in an *ex-ante* impact assessment (e.g., high risk of disparate impacts on certain individuals or disproportionate environmental impact) is necessary for accountable conduct, it necessitates broad community engagement and consultation to be an effective guardrail. As the planning and design of AI systems is primarily concentrated in the hands of the private sector, which may regard impact assessments through a narrow corporate risk lens, counterbalancing forces are necessary in such processes. Limiting the determination of algorithmic risks to provider and client stakeholders will be insufficient. Whilst recently proposed Algorithmic Impact Assessment frameworks sometimes foresee the consultation and exchange with potentially impacted communities, the modalities of the latter seldom seem sufficient for enabling their meaningful contribution to decision-making. Current Algorithmic Impact frameworks from government agencies in the Netherlands or the United Kingdom include not only the identification of affected stakeholders but also the obligation to determine what benefits and risks these groups may associate with a given AI system (Government of the Netherlands, 2023; United Kingdom Information Commissioner’s Office, n.d.). Despite identifying impacted communities and such that may be impacted by proxy is part of the challenges associated with conducting such assessments in the first place, it will not be enough for public and private experts to imagine which harms certain communities may be exposed to if a certain AI system is chosen to be deployed. Meaningfully including those community members in impact assessment processes may increase the chances of more critically assessing whether there is a need to use AI to address a certain task at all. Gangadharan’s (2019) “politics of refusal” is a key concept to explore in this regard. She contends that when marginalised communities choose to refuse certain technologies, this does not need to mean a blanket, full-scale technology rejection but rather a form of “informed refusal” which allows to reimagine “new ways of being and relating to one another in a technologically mediated society” (Gangadharan, 2019, p.113). In the context of algorithmic impact assessments, this means to create environments that effectively allow AI actors to formulate counternarratives and express “informed refusal” when considering the suitability of an AI system for deployment.

Moreover, meaningful engagement of impacted communities may reduce the anticipated risk that impact assessments lead to an abstract construct of algorithmic harm that is rationalised and numerically treated rather than based on the lived experiences of algorithmic harms (Metcalf et al., 2023). Blending abstract assessments with community experience and expertise should be a consideration for decisionmakers developing algorithmic impact assessment tools. This will require to make important design choices regarding the modalities of how impact assessments are conducted. Choosing between consultation and notice, co-creation or selective community engagement models will determine how well public input can be translated into how AI system are assessed and deemed appropriate for use in a given context. Other more seemingly unimportant choices about how results of assessments are presented may also influence the stringency with which organisations will consider public voices. Whether impacts are assessed through binary Yes and No questions, ticking boxes of categories or having to provide open-ended answers may influence not only what information may later be accessible to reviewers, auditors or the public but also how it may be perceived within the organisation. For example, ticking a box stating that public concerns have been heard and considered is not comparable to having to provide a summary of stated concerns, how they are evaluated and potentially mitigated. Finally, it is important to acknowledge the high financial costs and long timelines meaningful public engagement incurs upon organisations, which may be an inhibitor for them to choose such means of engagement. This may be further sustained by primarily corporate ambitions to avoid public scrutiny that may not be easily fixable within a firm's public image, particularly if it might entail the non-development or use of a technology. There are counter examples of Big Tech companies such as Microsoft that openly advertise AI applications they chose not to release after internal risk assessments, yet such examples only underwent internal scrutiny (Microsoft, 2023). Opening such processes to a broader public surely would produce diverging assessments. Whilst it may not be realistic to assume that the current standard-setting process will revolutionise already entrenched modes of limited public engagement in shaping our digital worlds, making it a key consideration in the design of algorithmic impact assessments may nevertheless be fruitful.

Overall, the ensuing standard setting process of algorithmic impact assessment tools should include the possibility of refusal, as articulated by the “assessment and accountability” requirement expressed in the UGAI and a sincere involvement of community voices. Policymakers and civil society actors such as the Center for AI and Digital Policy involved in shaping the modalities of AI policy development will play an important role in this process and the UGAI build a strong basis for this after its five years of existence.

Bibliography

- **Benjamin, R**. (2019). Race after Technology: Abolitionist Tools for the New Jim Code. Chapter 1: Engineered Inequity, 2: Default discrimination, p. 59, p.80-84. *Medford: Polity Press*.
- **Center for AI and Digital Policy.** (2018). Universal Guidelines for AI. Website. https://www.caidp.org/events/oct2023-dc-ugai/
- **Eubanks, V.** (2018). Automating Inequality : How High-Tech Tools Profile, Police, and Punish the Poor. Macmillan. https://us.macmillan.com/books/9781250074317/automatinginequality
- **Gangadharan, S.** (2019). Digital Exclusion: A Politics of Refusal. Retrieved September 30, 2023, from http://eprints.lse.ac.uk/103076/1/Gangadharan_digital_exclusion_politics_of_refusal_accepted.pdf
- **Government of Canada.** (2023). Algorithmic Impact Assessment Tool. Website. https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html
- **Government of the Netherlands.** (2023). AI Impact Assessment. Website. https://www.government.nl/documents/publications/2023/03/02/ai-impact-assessment
- **Metcalf, J., Moss, E., Watkins, E. A., Singh, R., & Elish, M. C.** (2021). Algorithmic Impact Assessments and Accountability. *Proceedings of the 2021*

ACM Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3442188.3445935

- **Microsoft.** (2023). Governing AI: A blueprint for the future. Report. https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW14Gtw
- **O'Neil, C.** (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Google Books. https://books.google.fr/books/about/Weapons_of_Math_Destruction.html?id=CxD-DAAAQBAJ&redir_esc=y
- **OECD.(2019).** Recommendation of the Council on Artificial Intelligence. OECD Legal Instruments. https://oecd.ai/en/ai-principles
- **Stahl, B. C., Antoniou, J., Bhalla, N., Brooks, L., Jansen, P., Lindqvist, B., Kirichenko, A., Marchal, S., Rodrigues, R., Santiago, N., Warso, Z., & Wright, D.** (2023). A systematic review of artificial intelligence impact assessments. *Artificial Intelligence Review*, *56*(11), 12799–12831. https://eprints.whiterose.ac.uk/197705/
- UNESCO. (2023). Ethical Impact Assessment: A tool of the Recommendation of the Ethics of Artificial Intelligence. Website. https://www.unesco.org/en/articles/ethical-impact-assessment-tool-recommendation-ethics-artificial-intelligence?TSPD_101_R0=080713870fab2000c4b30646a292283645f6e977e7db5992ac9e963cc852a4cf292391d68baace6708f087749a143000f1dae990995a0d7eb069f3ccf9ab7cfb629cc34185322c59b859648707217669eb275366b69a653227bc0f6502d5dcc8
- **United Kingdom Information Commissioner's Office.** (n.d.). AI and Data Protection Risk Toolkit. Website. https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/ai-and-data-protection-risk-toolkit/