The Intersection of Disclosure and Scale: A Trauma-Informed Approach to AI

Amanda Dahl
Leverhulme Centre for the Future of Intelligence
University of Cambridge
akd57@cam.ac.uk

17 November 2024

Introduction

In Kazuo Ishiguro's "Klara and the Sun," Klara, an artificial intelligence, reflects that despite her efforts to accurately understand her human charge, "something would have remained beyond my reach” (Ishiguro, 2021, p. 338). This introspection highlights the fundamental challenge facing AI in the domain of caregiving - the gap between technical accuracy and true reflection of human complexity.

Socio-technical imaginaries of the future of care, like Ishiguro’s, hold a promise that someday we’ll achieve human-level artificial intelligence which is able to anticipate our every need, understand our deepest desires and provide us with the comfort and emotional support we crave, forever banishing loneliness and helplessness to a vestigial trauma of generations past. But if we consider the historical forces and economic drivers behind AI's current rise, we must question how it might evolve from a technology rooted in epistemologies of positivism and control to some version which deeply understands human frailties and complexities, providing this (utopian) version of care – rather than a materialisation that has achieved accuracy but still missed the mark, as did Klara.

In this essay, I consider Lucy Suchman's assertion that AI acts as a "disclosing agent for assumptions about the human" (Suchman, 2006, p. 226) and how this concept might be applied for harm prevention to AI situated in care. Examining AI through the lens of control, trauma and care allows us to unpack the underlying assumptions being disclosed. As AI systems are deployed at scale and pace, the logics and biases embedded in their design and development can have impacts on individual and societal experiences of trauma.

This analysis extends Suchman’s concept of AI as a "disclosing agent" into the caregiving domain, integrating emerging thoughts from the intersection of history, philosophy, and sociology of technology. By employing Jon Agar’s exploration of how "technologies…intervene between scales" (Agar, 2020, p.381), I propose the

Disclosure-Scale-Trauma framework. This novel approach suggests that AI systems in caregiving simultaneously: (a) disclose underlying assumptions about human nature and care; (b) intervene across multiple scales of human experience; and (c) modulate trauma responses at both individual and systemic levels. Through this lens, we can better understand AI's dual potential for control and care, revealing fundamental assumptions about what constitutes effective caregiving.

AI as a disclosing agent

Given that AI is a material expression of its developers, it is a technology imbued with human values and strivings. Thus, in line with Suchman, AI reveals the human characteristics embedded as artefacts of the many small decisions involved in it’s creation; from choices governing data collection and training features, to determining which version of objectivity prevails (Wiggins & Jones, 2023; Hagerty et al., 2023; Forsythe, 1994). Suchman contends that efforts to create humanlike AI, whether it be embracing embodiment as the key to doubling human qualities, or mimicking emotion for affect, generally result in dim simulacra of small tranches of humanity, rather than a convincingly humanlike machine.

When we examine the dominant paradigm of technology, born of military-industrial origins and, as Zuboff (2019) points out, shaped by Hayekian individuality and “impatient money” (p. 40), this disappointingly unhuman interpretation of humanity is hardly surprising. On examination of the forces which have shaped AI, we see embedded emphasis on command, control, extractivism, positivism and purported objectivity, leading to the production of harms and leaving us unable to fully grasp its impacts until we “trace the scars [it] carves into the flesh of our daily lives” (Zuboff, 2019, p. 19).

While techno-optimists, like Ray Kurzweil (2024), suggest that future advancements will enable AI to fully replicate human abilities, such views often overlook emotional understanding, with disdain for diversions “by anything as misleading as emotion or empathy” (McQuillan, 2022, p.93). Additionally, the notion of technological neutrality assumes AI is merely a tool shaped by its users. However, as Kranzberg (1986) affirms, technology isn’t neutral, and these counter arguments underscore the importance of critical assessment of AI's role in caregiving, highlighting the need for ethical accountability and trauma-informed practices.

AI intervenes between scales…of trauma

Agar (2020) suggests that a key characteristic of technologies is their ability to intervene between scales: thermodynamic scales (like a refrigerator), scales of luminosity (like lighting) and other scales yet to be determined for newer technologies like AI. The concept of the scale is useful for us to consider how technological achievements, in terms of their impact on different scales, can have transformative effects on human lives and society.

Alongside Suchman’s concept of disclosure, we can apply Agar’s ideas of scale to the tension between control and care as manifested in AI, particularly thinking about how that tension can both exacerbate and alleviate trauma. Trauma, a byproduct of this harm, can be understood as a deeply impactful experience that disrupts the sense of self, safety, and well-being (Herman, 1992). It manifests on various scales, from individual psychological distress to collective societal trauma resulting from large-scale events like war or natural disasters (Gray et al., 2004).

While Agar primarily focuses on physical scales, it is possible to extend the concept of scale so as to encompass psychological and emotional dimensions, along the lines of Seaver (2021) who concludes that care and scale are orthogonal. It is for this reason that I would like to add another dimension, in the hopes of arriving at vectors of harm prevention.

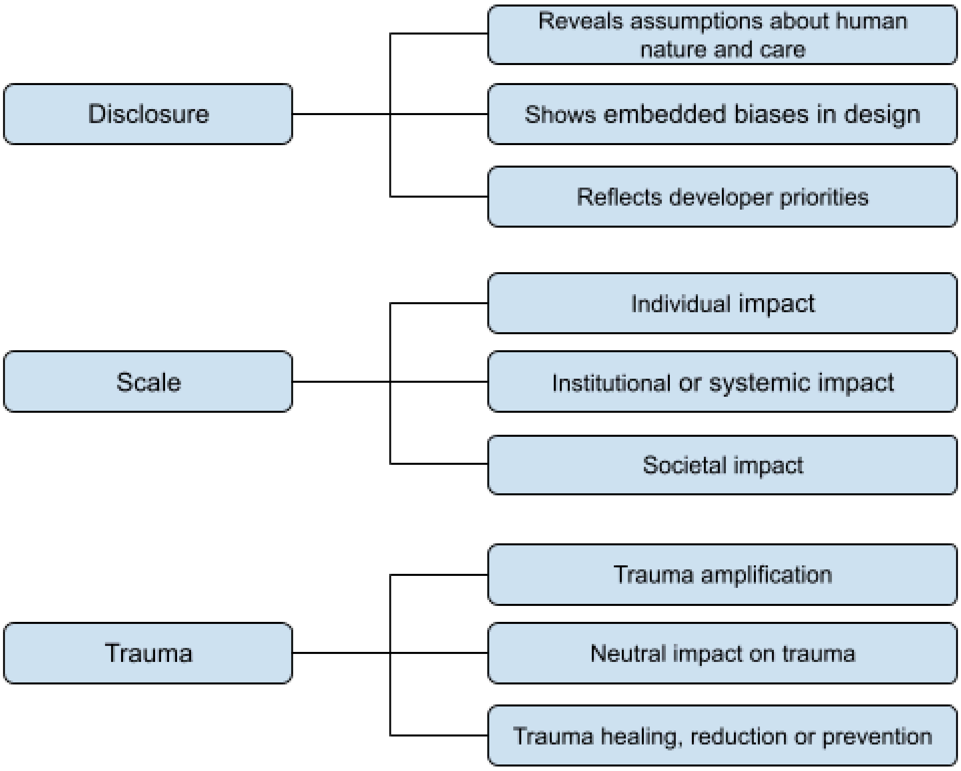
The Disclosure-Scale-Trauma Framework

It can be difficult to articulate what doesn’t hit the mark with AI that is designed for care, what remains “beyond reach” as Klara would say. As MacDorman (2024) suggests, the eeriness and ‘uncanny valley’ effect can arise from perceptual factors related to the physical appearance of robots, implying that discomfort with AI in caregiving roles might result from a mismatch between human-like appearance and our expectations based on that appearance. However, in the case of disembodied AI like chatbots, as suggested by Chhabra et al. (2024), user discomfort may be attributed to poor semantic understanding, lack of customisation, lack of humanisation, and poor competency. Frustration happens because users may find interactions with disembodied AI to be impersonal and inadequate in addressing their specific needs or understanding the context of their queries.

Seeking to pin down the reasons why AI might go beyond frustration and tip into causing harm or trauma, the Disclosure-Scale-Trauma framework is an attempt to create a theoretical framework that integrates the concepts of disclosure from Suchman, scale from Agar, and trauma-informed studies to analyse AI systems, particularly in contexts of care.

Figure 1

Disclosure-Scale-Trauma Framework


Disclosure
Reveals assumptions about human nature and care
Shows embedded biases in design
Reflects developer priorities
Scale
Individual impact
Institutional or systemic impact
Societal impact
Trauma
Trauma amplification
Neutral impact on trauma
Trauma healing, reduction or prevention

Note. Illustration of the Disclosure-Scale-Trauma framework. Own work.

The Disclosure-Scale-Trauma framework suggests that AI systems simultaneously disclose assumptions about human nature, intervene across multiple scales, and modulate trauma responses.

- *Disclosure dimension*: Examines the underlying assumptions about care and control that AI systems reveal through their design and implementation.
- *Scale dimension:* Explores how AI impacts different levels, from individual experiences to societal structures, either amplifying or mitigating effects.
- *Trauma dimension:* Considers how AI can alleviate or exacerbate trauma through diagnosis, misdiagnosis, or exclusionary practices.

For instance, an AI system developed to improve family court outcomes (Hall, 2024) that detects victim-blaming language amongst judges (disclosure), impacts both individual judgements and institutional protocols (scale), and reduces risks of re-traumatising vulnerable individuals (trauma).

By analysing AI systems through this integrated lens, the framework highlights inherent biases and the need for inclusive, trauma-informed approaches.

Case Study:
Voice Analysis for PTSD Diagnosis through the Disclosure-Scale-Trauma Lens

Autonomous weapons systems reveal assumptions about psychological distance and moral responsibility in warfare, while also affecting trauma scales. This impact appears not only in trauma to civilians under drone surveillance (Pong, 2022), but also in psychological effects on operators. Despite physical safety, operators experience trauma, re-traumatisation, and PTSD from constant exposure to war imagery. Their distress is compounded by the moral burden of remote killing, accountability ambiguity (Elish, 2016), and potential dehumanisation of targets (Mbembe & Meintjes, 2003).

What happens when operators experience trauma as a result of repeated exposure to remote killing? In the United States, the Veterans Administration (VA) has achieved progress in access to care for veterans, including LGBTQ+, women, and unhoused Veterans (Kaboli & Shimada, 2023). However, the economic case to employ AI automation in care has become increasingly strong (Oliver, 2008).

This case study examines AI developed by MITRE and deployed by the VA for automated PTSD detection (Scott, 2024). The underlying technology used in current PTSD detection generally uses biometrics (Bourla et al.,2018) and gender-based voice analysis to detect fluctuations which could indicate if the subject is at risk (García-Valdez et al., 2024). There is no clear indication that the diagnostic AI deployed by MITRE/VA diverts from this norm, therefore I am analysing it accordingly using the Disclosure-Scale-Trauma framework. Should MITRE disclose more detail about their models, it would merit further analysis.

Disclosure

The PTSD voice analysis systems reveal assumptions about the nature of trauma and human identity. Epistemologically, they are premised on the notion that trauma can be reduced to quantifiable vocal characteristics, disclosing a positivist view of psychological experiences. Furthermore, by categorising PTSD patterns strictly along binary gender lines (García-Valdez et al., 2024), these technologies disclose and reinforce normative assumptions about gender, excluding non-binary identities. This aligns with Suchman's contention that AI often replicates narrow, familiar assumptions about humanity.

Scale

The systems intervene across multiple levels with far-reaching effects. At an individual scale, they impact personal diagnoses, access to treatment, and can enforce gender categorisation, thus shaping personal experiences of trauma. Institutionally, such technologies influence healthcare protocols, resource allocation decisions, and embed biases within medical infrastructures. On a societal scale, they perpetuate exclusionary narratives and reify reductive gender binary norms surrounding trauma, particularly within veteran mental healthcare contexts.

Trauma

These voice analysis systems risk exacerbating trauma through misdiagnosis and invalidation. While aimed at PTSD diagnosis, their rigid gender assumptions may cause them to miss or discount the trauma experiences of non-binary individuals. This poses risks of re-traumatisation through misgendering and the psychological toll of impersonal, invalidating diagnostic processes. Moreover, by reinforcing gender-based discrimination and excluding non-binary trauma narratives, such technologies can have systemic traumatising effects on marginalised communities.

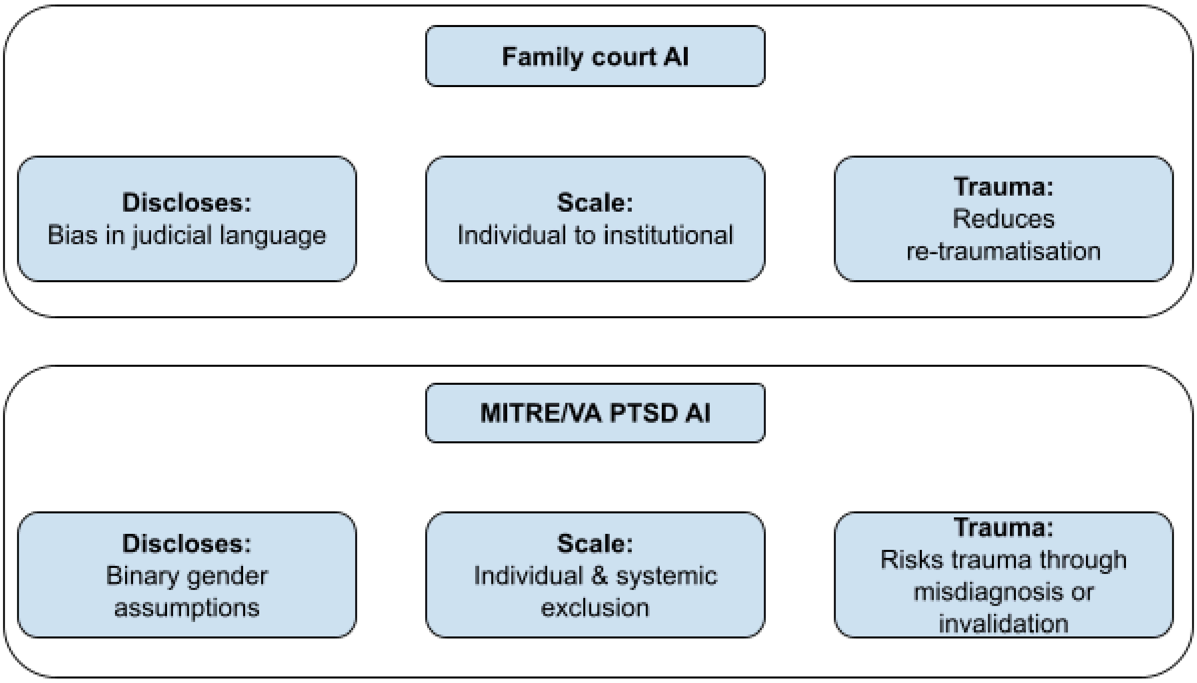
MITRE's PTSD AI technology and the gender-based voice trauma analysis systems exemplify the challenges posed by a lack of diversity in treatment design and an emphasis on care as a site of profit (Nadasen, 2023). By examining these systems through the Disclosure-Scale-Trauma framework, we can articulate the case for trauma-informed and ethically accountable AI that acknowledges the complexity of human experience.

Healing trauma through AI

AI technology is not merely a collection of algorithms, but rather multifaceted systems with interconnected layers (Crawford & Joler, 2018), with mildly-attributed data compiled from numerous sources, and using routines which grow together over time. Because of this, it is easy for developers to deny their role in the system believing they only play a small part. But, as Zuboff (2019) points out, “technology is not and never can be a thing in itself, isolated from economics and society,” (Zuboff, 2019, p. 19) instead the responsibility of developers extends beyond narrow technical implementation to include a deep understanding of their position within complex socio-technical systems. They must regularly evaluate their design choices (Chen et al, 2022) and consider their outputs holistically.

Figure 2

The Disclosure-Scale-Trauma framework applied


Family court AI
Discloses:
Bias in judicial language
Scale:
Individual to institutional
Trauma:
Reduces re-traumatisation
MITRE/VA PTSD AI
Discloses:
Binary gender assumptions
Scale:
Individual & systemic exclusion
Trauma:
Risks trauma through misdiagnosis or invalidation

Note. Applying the framework highlights the contrasts between the two examples. Own work.

Technologies can be designed to mitigate the effects of trauma, actively working to alleviate harm. Kind (2020) describes a "third wave" of ethical AI that goes beyond abstract principles and technical fixes, focussing instead on practical measures for rebalancing power. This wave is characterised by a greater emphasis on action, accountability, and the development of tools for auditing and assessing AI systems. It also highlights the importance of decolonising AI, recognising the ways in which historical and ongoing colonialism can shape technological development and deployment and can include divesting from carceral technologies and investing in community-led solutions (Hamid, 2020).

There is an intentionality in these use cases that was not seen in previous “waves” of AI, where consideration is taken for epistemologies of trauma and ethnographic methods for trauma-informed design are consciously incorporated into practices. Applying the

Disclosure-Scale-Trauma, as shown in Figure 2 above can help highlight where intentionality and awareness can reinforce “third wave” ethical AI practices.

Conclusion

Returning to Klara's insight that “something would have remained beyond my reach,” we can now understand this limitation not simply as a technical boundary, but as an illustration of scale and disclosure. Klara's inability to fully replicate humanity represents the challenge of scale – the gap between computational processes and human experience cannot be bridged simply by increasing processing power or datafication. What remains “beyond reach” exists at a scale that, perhaps, cannot be mediated by technology.

Suchman shows us that AI serves as a disclosing agent at scale, reflecting deeply ingrained assumptions about the human condition and societal priorities. In advocating for balance between control and care, going beyond the binary, this essay aims to contribute to ongoing conversations about AI’s role in human well-being. To harness AI's potential as an instrument of care, developers must integrate trauma-informed practices and prioritise ethical accountability. As AI evolves, its position on the spectrum between control and care becomes increasingly crucial, requiring practitioners to commit to developing systems that minimise trauma and contribute to positive systemic change.

References

Abercrombie, G., Vitsakis, N., Jiang, A., & Konstas, I. (2024). Revisiting annotation of online gender-based violence. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024* (pp. 31–41).

Agar, J. (2020). What is technology? *Annals of Science, 77*(3), 377–382. https://doi.org/10.1080/00033790.2019.1672788

Bourla, A., Mouchabac, S., El Hage, W., & Ferreri, F. (2018). e-PTSD: An overview on how new technologies can improve prediction and assessment of Posttraumatic Stress Disorder (PTSD). *European Journal of Psychotraumatology, 9*(1), Article 1424448. https://doi.org/10.1080/20008198.2018.1424448

Brown, L. S. (2004). Feminist paradigms of trauma treatment. *Psychotherapy: Theory, Research, Practice, Training, 41*(4), 464–471. https://doi.org/10.1037/0033-3204.41.4.464

Chhabra, S., Kaushal, V., & Girija, S. (2024). Determining the causes of user frustration in the case of conversational chatbots. *Behaviour & Information Technology*. Advance online publication. https://doi.org/10.1080/0144929X.2024.2362956

Chen, J. X., McDonald, A., Zou, Y., Tseng, E., Roundy, K. A., Tamersoy, A., Schaub, F., Ristenpart, T., & Dell, N. (2022). Trauma-informed computing: Towards safer technology experiences for all. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1–20). https://doi.org/10.1145/3491102.3517475

Crawford, K., & Joler, V. (2018). Anatomy of an AI system: The Amazon Echo as an anatomical map of human labor, data and planetary resources. AI Now Institute and Share Lab. https://anatomyof.ai

Elish, M. C. (2016). Moral crumple zones: Cautionary tales in human-robot interaction. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2757236

Forsythe, D. E. (1994). STS (Re)Constructs anthropology: A reply to Fleck. *Social Studies of Science, 24*(1), 113–123. https://www.jstor.org/stable/370294

García-Valdez, A. A., Román-Godínez, I., Salido-Ruiz, R. A., & Torres-Ramos, S. (2024). Identifying PTSD sex-based patterns through explainable artificial intelligence in biometric data. *Network Modeling Analysis in Health Informatics and Bioinformatics, 13*(1). https://doi.org/10.1007/s13721-024-00485-y

Gray, M. J., Maguen, S., & Litz, B. T. (2004). Acute psychological impact of disaster and large-scale trauma: Limitations of traditional interventions and future practice recommendations. *Prehospital and Disaster Medicine, 19*(1), 64–72. https://doi.org/10.1017/S1049023X00001497

Hagerty, A., Aranda, F., & Jemio, D. (2023). Predictive puericulture in Argentina: The Plataforma Tecnológica de Intervención Social and the reproduction of Latin eugenics. *BJHS Themes, 8*, 221–235. doi:10.1017/bjt.2023.2

Hall, R. (2024). Family Court judges use victim-blaming language in domestic abuse cases, finds AI Project. *The Guardian*. https://www.theguardian.com/law/2024/oct/08/family-court-judges-victim-blaming-language-domestic-abuse-cases-ai-project

Hamid, S. T. (2020). Abolishing carceral technologies. *Logic Magazine*. https://logicmag.io/care/community-defense-sarah-t-hamid-on-abolishing-carceral-technologies/

Herman, J. L. (1992). Complex PTSD: A syndrome in survivors of prolonged and repeated trauma. *Journal of Traumatic Stress, 5*(3), 377–391. https://doi.org/10.1002/jts.2490050305

Ishiguro, K. (2021). *Klara and the Sun*. Faber.

Kaboli, P. J., & Shimada, S. L. (2023). A decade of focus on and improvement in access to care in the Veterans Health Administration. *Journal of General Internal Medicine, 38*(3), 801–804. https://doi.org/10.1007/s11606-023-08208-1

Kind, C. (2020, August 23). The term "ethical AI" is finally starting to mean something. *VentureBeat*.https://venturebeat.com/2020/08/23/the-term-ethical-ai-is-finally-starting-to-mean-something/

Kranzberg, M. (1986). Technology and History: “Kranzberg’s Laws”. *Technology and Culture 27*(3). 545. https://doi.org/10.2307/3105385.

Kurzweil, R. (2024). *The singularity is nearer: when we merge with AI.* Chicago/Turabian

MacDorman, K. F. (2024). Does mind perception explain the uncanny valley? A meta-regression analysis and (de)humanization experiment. *Computers in Human Behavior: Artificial Humans, 2*(1), Article 100065. https://doi.org/10.1016/j.chbah.2024.100065

Mbembé, J-A., & Meintjes, L. (2003). Necropolitics. *Public Culture, 15*(1), 11–40. https://doi.org/10.1215/08992363-15-1-11

McQuillan, D. (2022). *Resisting AI: An Anti-Fascist Approach to Artificial Intelligence*. Bristol University Press.

Nadasen, P. (2023). *Care: The highest stage of capitalism*. Haymarket Books.

Oliver, A. (2008). Public-sector health-care reforms that work? A case study of the US Veterans Health Administration. *The Lancet, 371*(9619), 1211–1213. https://doi.org/10.1016/S0140-6736(08)60528-0

Pong, B. (2022). The art of drone warfare. *Journal of War & Culture Studies, 15*(4), 377–387. https://doi.org/10.1080/17526272.2022.2121257

Seaver, N. (2021). Care and scale: Decorrelative ethics in algorithmic recommendation. *Cultural Anthropology, 36*(3), 509–537. https://doi.org/10.14506/ca36.3.11

Scott, T. (2024). What does trauma sound like? AI tool mines voice cues to detect PTSD. *MITRE*.https://www.mitre.org/news-insights/impact-story/what-does-trauma-sound-like-ai-tool-to-detect-ptsd

Suchman, L. (2006). *Human-machine reconfigurations*. Cambridge University Press.

Wiggins, C. H., & Jones, M. L. (2023). *How data happened: A history from the age of reason to the age of algorithms*. W.W. Norton & Company.

Zuboff, S. (2019). *The age of surveillance capitalism*. Hachette Book Group.