Weizenbaum's legacy in the era of public interest AI

Anne Gerdes[1]

[1] *University of Southern Denmark, Universitetsparken 1, 6000 Kolding, Denmark*

Abstract

Revitalizing Weizenbaum's work, this paper presents an overall account of how AI threatens human autonomy and authenticity, paving the way for computer calculations and standardization to overshadow human judgment. Against this backdrop, negotiating AI's role in society requires finding ways to include the voices of socially marginalized groups. We have not yet accomplished that as researchers or citizens. Consequently, public interest AI shall continuously strive to voice marginalized groups' life experiences and concerns regarding AI and digitalization. In this context, participatory approaches are often advocated for empowerment. However, such approaches may inadvertently amplify inequity, as socially marginalized groups might lack the capability and motivation to engage in such activities disconnected from their daily realities. To achieve meaningful and enduring improvements for marginalized groups, public interest AI should enhance research that contextualizes AI challenges within the lived experiences of those most adversely affected by AI-driven interventions. Additionally, public interest AI should establish a comprehensive framework for evaluating, documenting, and sharing successful projects and interventions that have resulted in successful, lasting improvements for marginalized groups.

Keywords

Weizenbaum, participatory design, authenticity, autonomy

1. Introduction

Since the 70s, in the field of Human Computer Interaction (HCI), participatory design has democratized system development [1, 2]. Similarly, highlighting ethics, value sensitive design (VSD) [3-6] has emphasized stakeholder engagement and the need to "front load" ethics [7] to proactively include attention to values in the design of technologies [3, 8]. In contemporary research on AI-driven digitalization, approaches, such as the FATML community, provide tools to enhance transparency, mitigate algorithmic bias, and document the quality of data sets and models [9, 10]. Ethics by Design or Responsible AI frameworks [11-14] seek to anticipate ethical challenges in AI design and deployment. Yet, besides the early days of participatory design (70s/80s), these approaches lack reflections concerning how design choices are affected by the socio-technological patterns underneath.

For instance, (VSD) aims to design technologies that reflect what matters to people, with a strong emphasis on ethics and morality [3, 4, 8]. VSD seeks to anticipate ethical considerations early in the design process, ensuring that stakeholder values are central [15]. Likewise, AI for Social Good Value Sensitive Design (AI4SG-VSD) aligns with the EU's ethical guidelines on AI, using value hierarchies to guide design decisions and enhance understanding of the ethical challenges in AI development [16]. However, VSD lacks a critical reflection on how underlying socio-technological patterns influence design choices. While VSD has led to commendable projects involving vulnerable groups [17, 18], it tends to overlook broader power dynamics and political conflicts beyond the immediate stakeholders or design context. As a result, VSD sometimes falls into a "depoliticized scholasticism"[19]. Friedman et al. [20] have recently acknowledged the need to consider power dynamics in VSD.

Additionally, digitalization is not inevitable. Framing research questions solely within a technical context, such as relying on "-by Design" approaches to solve digitalization issues, leads to technological solutionism. Moreover, the interests of socially marginalized groups are frequently echoed by proxies like NGOs, hence not reflecting a first-hand perspective. And, as [21] and [22] point out, marginalized groups often lack the strategic knowledge to influence societal agendas effectively.

Revitalizing Weizenbaum's insights, this paper begins by outlining the broader societal challenges posed by AI's impact on human autonomy and authenticity. It argues that the erosion of these qualities has allowed datafication and standardization to increasingly override human judgment (sec. 2) and

* Corresponding author.
gerdes@sdu.dk
0000-0002-2991-5074

discusses AI's amplification of societal structural disparities (sec. 3). While the AI ethics research community has embraced citizen and user engagement methods, it has mostly overlooked the challenge of involving marginalized groups in AI projects as well as the potential drawbacks of participatory approaches, as such approaches might inadvertently increase inequity. The paper concludes (sec. 4) by suggesting that public interest AI should establish a comprehensive framework to facilitate the reproducibility of research results and enable benchmarking against one another. This approach will help assess how participatory methods contribute to lasting improvements and ensure that successful projects and interventions are effectively communicated and compared.

2. Revisiting Weizenbaum – authenticity and human encounters

Sometimes, a compelling parable overshadows the core message. When referring to Joseph Weizenbaum's book "Computer Power and Human Reason" [23], most people think of Eliza and the human tendency to anthropomorphize technology. This is a pity. Surely, Weizenbaum was astonished by people's reactions to the Eliza program [24], which, in his first experiment, was given a script designed to mimic a Rogerian psychotherapist. Weizenbaum saw how people exaggerated the capabilities of a simple computer program and anthropomorphized DOCTOR (the psychiatrist version of Eliza) entering dialogues as if they were talking to a psychiatrist. Also, psychiatrists expressed their enthusiasm and "seriously believed the DOCTOR program could grow into a nearly automatic form of psychotherapy" [23]. As a computer scientist, Weizenbaum knew from experience that programmers form ties to their machines. However, he had not expected that "extremely short exposures to a relatively simple computer program could induce powerful delusional thinking in quite normal people." [23]. Nevertheless, and more importantly, Weizenbaum viewed his experience with Eliza and DOCTOR as a sign of deeper societal problems, noting that "[o]ne socially significant question I thus intend to raise is over the proper place of computers in the social order." [23]. He feared dehumanization and a "mechanical conception of man" [23]. He noted how some psychiatrists saw DOCTOR as a promising tool for efficient psychiatric treatment and even described the therapeutic task in terms of computer metaphors as being about information processing. Yet, finding out what constitutes significant issues for me is not something I can do in isolation with Eliza echoing my words in a Rogerian data pattern. Taylor [25] emphasizes the dialogical feature of our existence as a condition for the ideal of authenticity and notes that authenticity requires "openness to horizons of significance" [25], i.e., "a pre-existing horizon of significance, whereby some things are worthwhile and others less so, and still others not at all" [25]:

> Otherwise, I can define my identity only against the background of things that matter. But to bracket out history, nature, society, the demands of solidarity, everything what I find in myself, would be to eliminate all candidates for what matters. Only if I exist in a world in which history, or the demands of nature, or the needs of my fellow human beings, or the duties of citizenship, or the call of God, or something else of this order *matters* crucially, can I define an identity for myself that is not trivial. Authenticity is not the enemy of demands that emanate from beyond the self; it supposes such demands.
>
> [25]

Being true to myself involves grasping my original way of being – "I am called upon to live my life in this way, and not in imitation of anyone else's" [25]. Hence, I must be in contact with myself to be true to my originality. This does not imply a slide towards subjectivism in the shape of fulfillment as "just-me-dialoguing-with-Eliza." In Taylor's words, authenticity must be understood in the background of the dialogical feature of our condition, emphasizing the demands of our relations with others – "self-definition in dialogue" [25]. Hence, Taylor argues that we define and form our identity through dialogical interactions with significant others in our surroundings, and those significant others cannot be automatic therapists for individualized self-realization via self-indulging activities [25].

Thus, the computational invasion of the realm of therapy threatens human autonomy and authentic human encounters. Nevertheless, according to Bassett [26], the Rogerian script behind the DOCTOR version of Eliza might also be viewed as a tool facilitating self-actualization, which implies that Weizenbaum's dichotomic clash between computer power versus human autonomy is too simplistic because a socio-technic relation might generate a therapeutic space for "self-realized autonomy":

> Perhaps we might say that, if 'ELIZA' was code, then 'Eliza' was the comfort found in the machine, by

humans, who built a different kind of relationship with 'her' that exceeded what the procedures of code offered, precisely because code came into contact with human thought.

[26]

Similarly, Verbeek [27] views technologies as mediating devices that actively shape human understanding and actions. While technological artifacts may lack human-like intentions, Verbeek operates with distributed or hybrid intentionalities, whereby technology can be attributed intentionality as playing a directing role in human actions and experiences:

Technologies "in themselves" cannot be free, but neither can human beings. Freedom is a characteristic of human-technology associations. On the one hand, technologies help to constitute freedom by providing the material environment in which human existence takes place and takes its form. And on the other hand, technologies can form associations with human beings, which become the places where freedom is to be located. Technological mediations can create the space for moral decision making.

[27]

However, socio-technically formed relationships [26] or Verbeek's post-phenomenological interpretation of human-technology interactions [27] do not imply that Weizenbaum was wrong in arguing that reliance on computer capabilities has led to the loss of authentic human interactions and a devaluation of human judgment. Consequently, it matters how and with what we enter into dialogue. Echoing Turkle, the question is not what technology will be like in the future but rather what we will become as we increasingly form relationships with and through technology. Thus, Turkle cautions that the advent of the *robotic moment* is approaching. This signifies the point where we opt for risk-free robotic interactions instead of authentic but demanding human encounters [28]:

At the robotic moment, we have to be concerned that the simplification and reduction of relationships are no longer something we complain about. It may become what we expect, even desire. [..] We have invented inspiring and enhancing technologies, and yet we have allowed them to diminish us. The prospect of loving, or being loved by, a machine changes what love can be.

[28]

Likewise, social media platforms set a scene for social relationships scaffolded around an attention economy [29] in which metric values fertilize self-esteem:

The sweeping success of Facebook, Instagram, Twitter & co. testifies to the enormous demand for possibilities of earning attention, which, in turn, testifies to the significance of self-esteem as a motive of business behaviour. The attention measured in terms of page views, likes and followers corresponds to the social product measured in terms of money income.

[29]

Hence, within the attention economy, our data is no longer the raw material driving Big Tech's business model. Instead, we are the raw material in a surveillance capitalism [30], shaped by big tech, making us addicted to their products. Not only do they predict our attitudes, behaviors, preferences, and desires, they also shape them. Thus, self-determination regarding something as vital as our identity and self-understanding is put under pressure.

In summary, the contemporary technological landscape makes it hard to cultivate authentic ends, as reliance on computer capabilities has diminished not only authentic human interactions but also devalued human judgments. Thus, Weizenbaum was correct in observing that "[w]hat emerges as the most elementary insight is that, since we do not now have any ways of making computers wise, we ought not now to give to computers tasks that demand wisdom [23]. Nevertheless, the following section outlines the drawbacks of precisely doing *that* by prioritizing computational calculations over human judgment in public digitalization efforts.

3. Revisiting Weizenbaum – datafication and the lack of inclusion

Weizenbaum's analysis of the consequences of viewing the computer as indispensable and technological advancement as unavoidable [23] reflects the current challenges in AI-driven digitalization, as illustrated in the quotations below:

The computer becomes an indispensable component of any structure once it is so thoroughly integrated with the structure, so enmeshed in various vital substructures that it can no longer be factored out without fatally impairing the whole structure. (…) It is not true that the American banking system or the

stock and commodity markets or the great manufacturing enterprises would have collapsed had the computer not come along “just in time”. It is true that the specific way in which these systems actually developed in the past two decades, and are still developing, would have been impossible without the computer. It is true that, were all computers to suddenly disappear, much of the modern industrialized and militarized world would be thrown into great confusion and possible utter chaos. The computer was not a prerequisite to the survival of modern society in the post-war period and beyond; its enthusiastic, uncritical embrace by the most “progressive” elements of American government, business, and industry quickly made it a resource essential to society’s survival in *the form* that the computer itself had been instrumental in shaping.

[23]

Problems in the postwar decades might have encouraged political and social innovation. Instead, the computer came “just in time” and was used to “conserve America’s social and political institutions” [23]. Following up, Weizenbaum offers insights that parallel contemporary discussions concerning how the ongoing datafication across sectors disvalues the parts of a human practice that cannot be quantified:

The computer has thus begun to be an instrument for the destruction of history. For when society legitimates only those "data" that are "in one standard format" and that "can easily be told to the machine", the history, memory itself, is annihilated.

[23]

Correspondingly, Crawford [31] shows how we design AI tools that shape the world according to what can be computerized. Noting that "the theory fits what the tools could do," she describes how Ekman's widely disputed theory of facial expressions made its way into the flourishing industry of emotion detection systems. Ekman proposes the existence of basic emotions, which are universally expressed through six facial expressions. These come in handy when labeling image datasets to train emotion detection models. Consequently, Ekman’s theory is easily applicable in the field of computer vision. Against this context, Crawford exemplifies how emotion detection tools amplify structural inequalities due to skewed data and classification systems that reinforce historical and cultural suppressive schemes and favor that which can be easily formalized:

They [affect recognition tools] take us back to the phrenological past, where spurious claims were made, allowed to stand, and deployed to support existing power systems. The decades of scientific controversies around the idea of inferring distinct emotions from human faces underscores a central point: the one-size-fits-all recognition model is not the right metaphor for identifying emotional states.

[31]

As an additional illustration of the persistent issue of inherent structural biases, Buolamwini highlights the matrix of domination referring to AI health tools for skin cancer detection, which underperform when assessing individuals with darker skin as these tools have been trained on datasets that primarily comprise individuals with lighter skin tones. She concludes that “A sociotechnical view requires that we think not only of datasets but also of the social conditions that led to a privileging of white skin in dermatology and how medical apartheid manifests” [32]. Likewise, in a Danish context, discussions questioning the absence of pictures of colored people in medicine textbooks are recent, even though dermatological examination of dark skin is more complicated than that of light skin.

Similarly, in dismantling the American welfare system's implementation of predictive risk modeling systems and profiling tools, Eubanks advocates for profound changes:

It will take more than high-tech tweaks to bring down the institutions we have built to profile, police, and punish the poor. It will take profound changes to culture, politics, and personal ethics.

[33]

As an essential first step, Eubanks encourages us to stop framing poverty in a manner that stigmatizes poor people. She coins the term "poverty profiling" and exemplifies how a model "confuses parenting while poor with poor parenting" [33].

Consequently, marginalized groups ought to be included and seen as a part of the solution rather than constantly being objectified as constituents of the problem. For instance, in discussing inequity in health and including attention to digitalization as well, [34] argues that labeling people as “vulnerable” tends to place the focus on issues in an individualistic framework, potentially assigning blame to them for their situation (“victim blaming”). Conversely, terms like “marginalized” or “exposed” patients indicate that the underlying problem complex is systemic, thereby emphasizing structurally driven inequalities.

We should draw lessons from the challenges experienced by marginalized groups, who are often overlooked despite being the ones most adversely impacted by public digitalization. We need to find ways to give voice to their first-hand perspectives and develop emancipatory approaches that move beyond doing "parachute research" [35], characterizing researchers immersing themselves in people's lives and experiences without striving to create lasting improvements for those affected. Against this backdrop, the growing field of design justice [36] recognizes community-based traditions. It focuses on how design might reproduce or challenge the matrix of domination, i.e., white supremacy, heteropatriarchy, capitalism, and colonialism. Design justice incorporates the principles of intersectional feminism into design theory and practice, aiming to cater to the diverse needs and experiences of various social groups. Social design tries to make changes as a form of social innovation, emphasizing dialogue and participation in redesigning social systems and living and working environments to create lasting changes for citizens and vulnerable groups [37]. In a technology setting, Bondi, Xu [6] seek to reformate AI for social good by focusing on "elevating the capabilities of those members who are most marginalized", arguing that this can be achieved through participatory approaches including "those most affected throughout the design, development, and deployment process" [6].

Within participatory design and HCI, there is a long-standing tradition of facilitating democratic user involvement in design activities [1, 38], aiming to give voice to users to empower them and improve their working conditions. Despite the successes of these pioneering approaches, one should not underestimate the challenges related to enacting participatory design, as noted by [39], who argue that users often view user involvement in research IT-design projects as a burden on top of their existing tasks. Even when researchers succeed in creating design spaces that offer mutual gains for researchers and users, this might still not result in lasting changes.

Individuals facing the most damaging consequences of digitalization are wrestling with challenging life conditions. Dedicating resources to engaging in, e.g., discussions about the role of AI in the healthcare sector, let alone participating in co-design research activities concerning AI, is not a priority. In most cases, marginalized groups are, if included at all, typically represented by proxies, e.g., NGOs voicing issues on their behalf with the risk of turning the design space into a political battlefield by being too eager to represent the perspective of "their" specific group. Or, as it is often seen in healthcare, patients' perspectives on AI are described by patients with a background in professional healthcare [40]. Inadvertently, approaches for user involvement and co-design might increase inequity [34]. Negotiating AI's role in society requires finding sustainable ways to include marginalized groups. We have not yet accomplished that as researchers or citizens.

4. Conclusion

Weizenbaum's legacy continues to be crucial in critiquing the dominance of computer power over human judgment. He demonstrates how the contemporary technological landscape makes it challenging to cultivate authentic goals, as reliance on computer capabilities has weakened genuine human interactions as well as devalued human judgment, ultimately paving the road for viewing computers as indispensable and technological advancement as unavoidable. His work resonates with contemporary discussions about AI's impact on human autonomy, authenticity, and its amplification of structural inequalities. Unfortunately, the public discussion on AI's role in society is often overshadowed by techno-solutionism and sensationalism, fueled by AI hype, distracting informed and inclusive discussions about AI's role in society.

Against this backdrop, public interest AI must continuously refine its approaches to voice marginalized groups' life experiences and concerns regarding AI, facilitating the voices of those most negatively impacted by AI interventions. Various participatory design approaches have successfully been redesigning social systems or technologies within contexts that are immediately meaningful in communities of marginalized groups. However, participatory approaches also risk unintentionally increasing inequities, given that marginalized groups may lack the necessary resources and incentives to engage, especially when activities concern challenges related to digitalization and AI in the public sector, as these may seem detached from their daily realities. Additionally, working on fostering lasting improvement for marginalized groups should be complemented by the development of a robust, comprehensive framework for measuring, collecting, documenting, and disseminating projects and interventions that have resulted in lasting positive changes for these communities.

References

1. Greenbaum, J. and M. Kyng, *Design at work: Cooperative design of computer systems*. 1991: Lawrence Erlbaum Associates Inc.
2. Bødker, S. and M. Kyng, *Participatory Design that Matters—Facing the Big Issues.* ACM Trans. Comput.-Hum. Interact., 2018. **25**(1): p. Article 4.
3. Friedman, B., D. Hendry, and A. Borning, *A Survey of Value Sensitive Design Methods.* Foundations and Trends® in Human–Computer Interaction, 2017. **11**(2): p. 63-125.
4. Friedman, B. and D. Hendry, *Value sensitive design: shaping technology with moral imagination*. 2019, Cambridge, Massachusetts: The MIT Press.
5. Umbrello, S. and I. van de Poel, *Mapping value sensitive design onto AI for social good principles.* AI and Ethics, 2021. **1**(3): p. 283-296.
6. Bondi, E., et al., *Envisioning Communities: A Participatory Approach Towards AI for Social Good*, in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 2021, Association for Computing Machinery: Virtual Event, USA. p. 425–436.
7. van den Hoven, J., *ICT and Value Sensitive Design*, in *The Information Society: Innovation, Legitimacy, Ethics and Democracy In honor of Professor Jacques Berleur s.j*. 2007, Springer US: Boston, MA. p. 67-72.
8. Friedman, B., P.H. Kahn Jr, and A. Borning, *Value sensitive design and information systems.* Human-Computer Interaction and Management Information Systems: Foundations, 2006: p. 348-372.
9. FATML. *FATML -Fairness, Accountability, and Transparency in Machine Learning*. 2022 [cited 2022 07-22-2022]; Available from: https://www.fatml.org/.
10. Mitchell, M., et al. *Model Cards for Model Reporting*. NEW YORK: ACM.
11. Dignum, V., *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. 1. 2019 ed. 2019, Cham: Springer International Publishing. 1-127.
12. Brey, P., Dainow, B. *Ethics by design and ethics of use in AI and robotics. The SIENNA project - Stakeholder-informed ethics for new technologies with high socioeconomic and human rights impact*. 2021; Available from: https://www.sienna-project.eu/digitalAssets/915/c_915554-l_1-k_sienna-ethics-by-design-and-ethics-of-use.pdf.
13. Floridi, L., et al., *AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations.* Minds and machines (Dordrecht), 2018. **28**(4): p. 689-707.
14. Gerdes, A., *A participatory data-centric approach to AI Ethics by Design.* Applied artificial intelligence, 2022. **36**(1).
15. van den Hoven, J., *ICT and Value Sensitive Design*, in *The Information Society: Innovation, Legitimacy, Ethics and Democracy In honor of Professor Jacques Berleur s.j.* 2007, Springer US: Boston, MA. p. 67-72.
16. Umbrello, S. and I. van de Poel, *Mapping value sensitive design onto AI for social good principles.* AI Ethics, 2021. **1**(3): p. 283-296.
17. Plass, J.L., et al. *RAPUNSEL: Improving self-efficacy and self-esteem with an educational computer game*. in *Proceedings of the 17th international conference on computers in education*. 2009.
18. Woelfer, J.P., et al., *Improving the safety of homeless young people with mobile phones: values, form and function*, in *CHI '11: CHI Conference on Human Factors in Computing Systems*. 2011, Association for Computing Machinery: Vancouver, BC, Canada. p. 1707-1716.
19. Winner, L., *Upon Opening the Black Box and Finding It Empty: Social Constructivism and the Philosophy of Technology.* Science, Technology, & Human Values, 1993. **18**(3): p. 362-378.
20. Friedman, B., et al., *Eight grand challenges for value sensitive design from the 2016 Lorentz workshop.* Ethics and Information Technology, 2021. **23**(1): p. 5-16.
21. Estlund, D. and D.M.M. Estlund, *Democratic Authority : A Philosophical Framework*. 2007, Princeton, UNITED STATES: Princeton University Press.
22. Washington, A. and R. Kuo. *Whose side are ethics codes on?: power, responsibility and the social good*. 2020. ACM.
23. Weizenbaum, J., *Computer power and human reason: From judgment to calculation*. 1976, San Francisco, Calif., Freeman.
24. Weizenbaum, J., *ELIZA—a computer program for the study of natural language communication between man and machine.* Commun. ACM, 1966. **9**(1): p. 36–45.
25. Taylor, C., *The Ethics of Authenticity*. 1991: Harvard University Press.
26. Bassett, C., *The computational therapeutic: exploring Weizenbaum's ELIZA as a history of the present.* AI & Society, 2019. **34**(4): p. 803-812.

27. Verbeek, P.-P., *Moralizing technology: Understanding and designing the morality of things*. 2011: University of Chicago press.
28. Turkle, S., *Alone Together: Why We Expect More from Technology and Less from Each Other*. 2011, New York, UNITED STATES: Basic Books.
29. Franck, G., *Vanity Fairs - Another View of the Economy of Attention*. 2020, Switzerland: Springer International Publishing AG.
30. Zuboff, S., *The age of surveillance capitalism: the fight for the future at the new frontier of power*. 2019, London: Profile Books.
31. Crawford, K., *Atlas of AI: power, politics, and the planetary costs of artificial intelligence*. 2021, New Haven: Yale University Press.
32. Buolamwini, J., *Unmasking AI: My Mission to Protect What Is Human in a World of Machines*. 2023, NY, United States: Penguin Random House.
33. Eubanks, V., *Automating inequality: how high-tech tools profile, police, and punish the poor*. First ed. 2018, New York, N.Y: St. Martin's Press.
34. Bruun, J., Bjarne, et al., *Involvering af borgere i sårbare positioner*. 2023.
35. Gebru, T. *Oxford Handbook on AI Ethics Book Chapter on Race and Gender*. 2019. DOI: 10.48550/arxiv.1908.06165.
36. Costanza-Chock, S., *Design justice: Towards an intersectional feminist framework for design theory and practice.* Proceedings of the Design Research Society, 2018.
37. Sachs, A. and C. Banz, *Social design: participation and empowerment*. 2018, Zürich: Lars Müller Publishers.
38. Bødker, S., et al. *Co-operative Design—perspectives on 20 years with 'the Scandinavian IT Design Model'*. in *proceedings of NordiCHI*. 2000.
39. Bossen, C., C. Dindler, and O.S. Iversen, *Impediments to user gains: experiences from a critical participatory design project*, in *Proceedings of the 12th Participatory Design Conference: Research Papers - Volume 1*. 2012, Association for Computing Machinery: Roskilde, Denmark. p. 31–40.
40. Fasterholdt, I., et al., *Model for ASsessing the value of Artificial Intelligence in medical imaging (MAS-AI).* International Journal of Technology Assessment in Health Care, 2022. **38**(1): p. e74.