Vulnerability, Trust and AI

AHMER ARIF*, University of Texas at Austin, United States
OS KEYES, University of Washington, United States

In this paper, we are working out a position on the interrelations between vulnerability, trust and distrust in AI research. Vulnerability is an important aspect of trust because it is critical to understanding issues of power and dependency that characterise many trust-relationships. Building on the literature, our work considers how vulnerability is conceptualised in research on AI trust more broadly. Drawing on Annette Baier's contributions, enlarged and developed recently by Gilson and Mackenzie, we argue for the treatment of vulnerability as a potential source of positive change rather than as a de-facto negative state that should be avoided. Illustrating our argument with examples from both within and without the AI Trust literature, we suggest some implications of viewing relations of trust as a product of mutual vulnerability for AI researchers. We conclude by describing some ways in which our argument is incomplete and needs further development.

ACM Reference Format:
Ahmer Arif and Os Keyes. 2022. Vulnerability, Trust and AI. In *Proceedings of Workshop on Trust and Reliance in AI-Human Teams at CHI 2022 (TRAIT).* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

> *"Trust is one of those mental phenomena attention to which shows us the inadequacy of attempting to classify mental phenomena into the 'cognitive', the 'affective', and the 'conative'. Trust, if it is any of these, is all three."*
> —Annette Baier, "Trust and its Vulnerabilities" [4, p.132]

Trust is a difficult concept to define, as the above quote by Annette Baier suggests. Yet Baier is also the author of one of the most influential definitions of trust, one deployed widely within HCI broadly and AI specifically [6, 10, 15]; that trust is "accepted vulnerability to another's possible but not expected ill will (or lack of good will) toward one"[3, p.99].

We (the authors) came to reflect on this definition as part of a broader project inquiring into the metaphysics of trust as it is conceptualised in misinformation and AI work. One particular facet that drew our attention was this concept of 'vulnerability', which is simultaneously under-theorised in much of the literature within HCI, and yet—as Baier's definition makes clear—essential to an understanding of trust.

In this paper, we are working out the consequences of taking vulnerability more seriously as an aspect of trust, particularly for an HCI audience. We start by elaborating on the concept of vulnerability. We will then unpack some implications for a vulnerability-centred understanding of trust for HCI audiences. We will then conclude with some reflections on how our position is incomplete and can benefit from future work.

*Both authors contributed equally to this research.

TRAIT, 30 April, New Orleans, LA, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 VULNERABILITY

Bair's definition of trust emphasises the toleration of vulnerability because in trusting others we make ourselves vulnerable to their possible betrayal or abuse of that trust. But how do we understand vulnerability? Vulnerability is conventionally seen as a state of being susceptible, exposed, or in danger. For example, the Oxford dictionary defines vulnerability as "the quality or state of being exposed to the possibility of being attacked or harmed, either physically or emotionally"[25].

Erin Gilson, Judith Butler and myriad other philosophers have critiqued this understanding of vulnerability—as a de facto negative state to be avoided—by pointing to the positive dimensions of vulnerability [7, 8, 14]. Gilson argues, convincingly, that vulnerability should instead be understood as a condition of "ambivalent potential"[13, p.310], one that, in Catriona Mackenzie's words, "opens us to being affected in both positive and negative ways"[22, p.624]. Although vulnerability makes us susceptible to suffering, it is also the source of some of the most profound emotions and human relationships. In the context of trust and AI, this ambiguity can be demonstrated with the example of trusting an algorithmic system that recommends employment opportunities; while one's vulnerability to its decisions is a risk (since it might steer the user wrong), it is also an opportunity (since it might offer the user unexpected and exciting sites of employment).

Moreover, vulnerability is inevitable, just like trust more broadly. In a world where each of us live in relation to each other—not only on a one-to-one basis, but woven into broader practices, infrastructures and systems—we are each always vulnerable. What matters, then, is not the avoidance of vulnerability but the avoidance of harmful forms of vulnerability that are rooted in morally dysfunctional situations, or oppressive structural arrangements. Such *situational vulnerability* can engender powerlessness, insecurity, and compromised agency [22]. Moreover, such vulnerabilities can make a person susceptible to inescapably risky trust relationships. A person who is socially marginalised or disadvantaged might know they are vulnerable to ill treatment, but be forced to tolerate that vulnerability to gain access to valuable opportunities that might be otherwise unavailable (at the risk of compounding existing vulnerabilities and enabling abuses of trust).

It is precisely this ambiguity that makes a considered understanding of vulnerability important, particularly in efforts to understand and establish trust in AI systems. The *inevitability* of vulnerability directs our attention away from overly rationalistic conceptions of trust and towards its affective dimensions. Trust or distrust become a matter of whether we feel optimistic or pessimistic about our potential vulnerability to others and the risks emerging from that vulnerability [16]. Similarly, *situational* vulnerability places issues of power and powerlessness at the centre of trust. In a situation of trust involving the same degree of vulnerability and risk, a person with social privilege and security will be less likely to regard this risk as a threat, whereas a person with social disadvantages and low security will be more likely to live in "continual awareness of her own vulnerability" [17]. Moreover, as Baier notes, sometimes distrust of the socially powerful is warranted: "If the network of relationships is systematically unjust or systematically coercive, then it may be that one's status within that network will make it unwise of one to entrust anything to those persons whose interests, given their status, are systematically opposed to one's own" [3, p.127]. Along these lines, we now turn to existing literature in AI to explore how vulnerability appears in the context of algorithmic trust.

3 VULNERABILITY IN AI

Vulnerability has a paradoxical place in research on AI trust. On the one hand—if we understand vulnerability as a precondition of positive change—the field is dependent on it. Work on trust in AI matters, implicitly and explicitly, because good things can come from that trust—from engagement in a way that enables change. Correspondingly, when users do not trust a system, they lose out on

those possible alterations and futures. In that respect, vulnerability is very much the foundation and impetus of research on AI trust.

But if we look at the other half of vulnerability—at the possible negative consequences of being vulnerable—we see a somewhat different picture. There are certainly myriad efforts to talk about trust (as the existence of this workshop suggests!), and many of them even point to Baier's vulnerability-centred definition, but much of the time the research question is some variant on: what attributes (of a system, or of a user) influence the formation of trust? How do we make a system trusted? These are important questions—but they are also questions that elide vulnerability; that ignore the possible negative consequences to a user of a system being trusted when it *should not be*.

In some respects this elision is somewhat inevitable; it is hard to take vulnerability seriously and simultaneously take those questions as being of paramount importance. If we centre vulnerability, we centre, in part, researchers" moral responsibilities: we centre the need not to ensure users trust a system but that a system *deserves to be trusted*. This is far harder to address—not only because the consequences of an action will always exceed what we can know [1], but also because the consequences of trusting an algorithmic system are not always (or usually) in the hands of the developers and researchers "on the ground".

To be sure, some research genuinely does examine the possible consequences of "over-trusting"—of making oneself vulnerable to a system and then experiencing negative consequences. A good example is Yang *et al.*'s work [28], which explores notions of "overtrust"—where the vulnerability that stems from trusting an algorithm that should not be trusted results in harm. Yang *et al.*'s paper, and work like it, represent a promising beginning. But their elision of vulnerability—in common with much of the work in this area, and so not a critique of them specifically—and so the elision of changes *to the self*, means that explorations of negative consequences often centre on the visible; the material; the *immediate*.

A good demonstration of other factors—and the broader implications of vulnerability for the consequences of (misplaced) trust—can be found in Mary F.E. Ebeling's "Healthcare and Big Data". The story began when Ebeling became pregnant, something she shared with her doctors. Sadly, two to three months in, she miscarried. Miscarriages are troubling from the get-go, and Ebeling speaks of her "crushing grief" [9, p.5] at the discovery—but her experience became even more complicated due to the consequences of her sharing her pregnancy with her doctor. That information was, unknown to her, sucked into marketing databases aimed at advertising baby-related purchases to new parents. Because her miscarriage did not register in the same way, the result was what she refers to as a haunting; a "marketing baby", a "data phantom". A stream of coupons, flyers and magazines, delivered to her door, advertising purchases for her new, dead, child:

> *"since I lost my baby in March 2011, I have received more than eighty separate email solicitations, social media advertisements, phone calls, mailed boxes of baby formula and diaper samples, magazines, baby photography offers, baby clothes, and direct-marketing flyers advertising everything from savings bonds to cord-blood banking. Much of the unsolicited mail I receive features softly lit photographs of dewy skinned babies, so freshly scrubbed I can almost smell the baby powder through the image, who beckon to me to buy Enfamil formula or a $1,200 Bugaboo All-Terrain stroller. The bulk of the direct mail offers, however, are for children's life insurance. I find these marketing offers particularly ghoulish."*[9, p.6]

The cause of all of this was misplaced trust; was vulnerability towards a medical system that, unbeknownst to her (and, as she later documents, her doctor) streamed her information not only to electronic medical records but also to marketing algorithms and databases as well. And while the

consequences of this certainly include a loss of her trust in that marketing system, they stretch far wider, in ways that a focus on engagement (or not) as a binary signal of trust cannot capture. We must also wonder: What medical information will she withhold, or be wary of disclosing, as a result of this? What are the long-term consequences for her health from having to make that determination on an ongoing basis? What are the implications for her future trust in medical and algorithmic processes more broadly? And how much more dangerous would the answers to these questions be if Ebeling was not, as she notes, a trained sociologist—one with the access, money and time to trace out the routes her pregnancy data had taken?

The consequences, in other words, touch on more than a simple matter of whether Ebeling trusts the tool—they are also deeply interwoven with (and potential catalysts for) broader relations of vulnerability. In particular, Ebeling's experiences play a role in what Mackenzie refers to as "pathologies of trust and distrust"—the emotional and experiential contexts in which one makes determinations of trust or mistrust. Trust and distrust are not isolated phenomena; rather, they "magnify and tend to be self-perpetuating" [17, p.956], particularly in the context of broader states of precarity and vulnerability. What this highlights is that people neither come to nor leave moments of trust as individualised blank slates—rather, they are shaped by pre-existing vulnerabilities, and past and present experiences of (mis)trust and (in)justice.

These experiences cannot be individualised; rather, they are constituted by broader relations of power. We can see this by returning to the example of job-search algorithms—systems that try to match a user to various employment opportunities. Finding employment is undertaken not just in an increasingly precarious world, but in a world where that precarity strikes differently at different people; it is shaped by longstanding inequalities around race, gender, sexuality and disability. These inequalities, and their depth, means that violations of trust around employment algorithms do not only cause immediate material harm. Rather, they reinforce pathologies of suspicion around harmed parties" status in society, while—with the expense in time and money of disputing the outcomes of processes—reducing those parties" ability to demand accountability, and perhaps restore a measure of trust. A reduction of these consequences and contexts to the user not finding a job, and so not trusting the system that promised to find them one, misses this.

4 RESEARCHER RELATIONS TO VULNERABILITY

Thus far, we have discussed the role of vulnerability in trust—and, in particular, the way that the consequences of vulnerability (whether they are good or bad, and to what degree) are heavily shaped by questions of power and precarity.

By now some of our more patient readers may be getting restive. "Yes, yes," they mutter, "it's all very well to say that trust is a matter of a certain orientation or affective stance towards vulnerability. And we are willing to grant you the point, more or less, that pre-existing power and precarity need to be considered when designing for human-AI trust. But what new questions does this raise? Should we start tabulating how people's varying backgrounds create situational vulnerabilities (in both the positive and negative senses), and how to address those vulnerabilities?" We applaud this instinct and recognise that there are already researchers heavily invested in mining such questions [2, 12]. Let us not forget that HCI is oriented towards problem-solving: we want to take action and fix things [5]!

However, we want to make the case that promoting trust requires asking not only questions about vulnerability, but also asking questions from a *different position*. If we acknowledge that vulnerability intersects with broader pathologies and relations, then we also have to accept that we, the researchers, are also inescapably enmeshed within those pathologies and relations. As such, we must confront the peculiar dilemmas of being committed to deepening trust while also sorting

through ontological and epistemological frameworks or practices that may be incommensurable with those very goals.

Consider, for example, positivist frameworks and traditions that reinforce the myth of the unattached researcher who can steer clear of entangling or compromising loyalties. This detachment can make it harder to think and communicate clearly about the politics of intellectual work on trust—a politics that will, of necessity, partly take place within structures of higher education and technology companies. Overlooking this politics can lead well-meaning interventions about promoting trust in AI to actually reinforce mistrust when users bring in the political picture in ways that those behind the intervention had artificially separated.

What is needed in other words, is not only a different set of questions, but a different set of relations around those questions—one that recognises the existing mistrust and vulnerability, not just in AI but often in AI *researchers*. In this line of thought, we are always already entangled, mediated, interdependent; "non-complicit" is not only an unrealised condition but an *unrealisable* one. Some of these pre-existing entanglements prove more helpful or enabling than others, while others limit or constrain, but the condition of being entangled is not a choice. What is a choice is how we recognise and practice those entanglements. At the present time, it seems self-evident that users are vulnerable to us; that our relations take a form in which researcher actions have an impact on them. But if we understand vulnerability as the state of being responsive to, or capable of being changed by, the other party—we can ask: are researchers vulnerable to users?

Our answer is "no", for a variety of reasons. Speaking generally, the positivist frameworks mentioned above not only encourage avoidance of vulnerability but valorise it; the scientific ideal is, precisely, to be unconcerned with the consequences of one's pursuit of truth. More organisationally, large systems—from university systems to corporate entities to the complex infrastructural chains between a theorist and a user—provide further insulation. Consider when an AI system (such as a self-driving car system) causes harm—who is forced to confront that harm? To wrestle with their relation to it, and the consequences it has for the shape their life takes moving forward? The victim, certainly; the driver of the car, absolutely. But the researcher—who experiences that harm solely as a slight tick in the false positive or false negative rate—is insulated. In our effort to minimise uncomfortable forms of vulnerability, we have—aided by access to structures of power that many of our users lack—also closed off its transformative potential. We have, as Tutenel and Heylighen remark of HCI approaches to vulnerability more broadly, treated it as a "problem to be solved...[something to] design away"[27].

How can researchers be vulnerable with users in ways that are useful and potentially transformative? A conceptual starting point might come from Nagar's recent work on *radical vulnerability*, a framework of relationality that emerged from working with a collective-action movement involving marginal farmers and landless labourers in Uttar Pradesh in India [23]. A full accounting of this framework is beyond the scope of this paper, but we want to briefly invoke it here as a stimulus to thought. One of its core ideas is that researchers can build trust by embracing vulnerability rather than eliding it. Vulnerability becomes radical when we are critically open about it and collectively surrender to it to forge knowledge-making relationships that are more rooted in solidarity. These relations require academics to forgo the compartmentalisation of research and activism by accepting the politics of forging more just sociotechnical systems, politics that can be accompanied by difficult refusals. On a deeper level, these relations require academics to be "not merely travelling to the Othered worlds that form the basis of our knowledge claims"[24], but to open themselves to being entwined with those others in ways that can transform them—becoming a "we" with our users that struggles to overcome intense distrust or suspicion at times.

These ideas may seem distant and impractical for building trust in AI, but we can see connections to existing lines of work within HCI research and practice. There are certain affinities, for example,

between this direction and feminist HCI and participatory design approaches[19, 20]. The difference from participatory design—and the affinity to feminist methods—is that radical vulnerability is centred not on collaboration changing what is built, but collaboration changing the builders. It carries a more capacious vision of transformation—one that makes room for more honesty in how we go about research. It is, after all, hard to argue that mistrust is unjustified if we refuse for ourselves the very vulnerability that transformation is premised on—transformation that we promise to users, and often demand they accept as the outcome of new technologies. The work of translating radical vulnerability into methodological approaches is vitally necessary for it to reach fruition. But if we can engage in that translation—if we can improve the fit of radical vulnerability with HCI's growing interest in designing for trust in AI—the result will be, we hope, a more honest and reflective field, one where the broader register of perspectives produces new possibilities for changing technology, and the self.

5 INCONCLUSION

In her introduction to *Hungry Translations*, AnaLouise Keating writes that "Partiality is inevitable. We are partial; our work is partial; our insights are partial. But this recognition of partiality contains the promise of progressive change, the assurance that when we acknowledge and embrace partiality, we open ourselves to innovative possibilities and fresh perspectives. But it is not easy to embrace partiality. Before we can do so, we must risk "radical vulnerability." We must step out of our self-enclosed boundaries and recognize our inevitable interconnectedness with others"[18]. This attitude—one of humility, acknowledged partiality and radical vulnerability—is something we attempt to adapt with this paper, and intend to bring to the workshop. Our position is not that we have the answers and others must adopt them; rather, it is that we are struggling with the questions and hope to struggle together. Correspondingly, we conclude not with a restatement of our argument, but with a set of open questions that we wish to discuss at the workshop:

(1) *How might we apply radical vulnerability?* As discussed, there are clear resonances between radical vulnerability as an orientation with feminist methods, along with forms of action research, that are familiar to many researchers in HCI[11]. But many methodological approaches—particularly the more positivistic frameworks often deployed around AI trust—are further afield. How might we interplay radical vulnerability and such methods? How would methods change, in continent or prominence; how might our understanding of radical vulnerability change?
(2) *What does radical vulnerability depend on?* Radical vulnerability, as an orientation, brings with it certain pragmatic needs and preconditions. One example that comes to our minds is time; the time to explore, to be frustrated with each other, to be willing to tolerate disagreement and approach moments of closure as, at best, contingent. It requires a willingness to accept the uncontrollability of the world [26], and that includes the uncontrollability of timelines and expectations. These questions are ones that HCI and design theory have already begun to confront in other domains[21, 29], and we are excited to entangle radical vulnerability with this work.
(3) What other conclusions—aside from radical vulnerability—could be drawn from considering vulnerability in trust research?

ACKNOWLEDGEMENTS

The authors would like to thank Claire and Margaret Hopkins and Rida Khan for their support, feedback and presence, along with Margret Wander for her timely feedback. This publication was supported in part by the Microsoft Ada Lovelace Fellowship.

REFERENCES

[1] Louise Amoore. 2020. *Cloud ethics.* Duke University Press.

[2] Theo Araujo, Natali Helberger, Sanne Kruikemeier, and Claes H de Vreese. 2020. In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & SOCIETY* 35, 3 (2020), 611–623.

[3] Annette Baier. 1994. Trust and Antitrust. *Moral prejudices* (1994), 95–129.

[4] Annette Baier. 1994. Trust and its vulnerabilities. *Moral prejudices* (1994), 130–151.

[5] Eric PS Baumer and M Six Silberman. 2011. When the implication is not to design (technology). In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2271–2274.

[6] Jeff Buechner. 2013. Trust and Ecological Rationality in a Computing Context. *SIGCAS Comput. Soc.* 43, 1 (may 2013), 47–68. https://doi.org/10.1145/2505414.2505419

[7] Judith Butler, Zeynep Gambetti, and Leticia Sabsay. 2016. *Vulnerability in resistance.* Duke University Press.

[8] Alyson Cole. 2016. All of us are vulnerable, but some are more vulnerable than others: The political ambiguity of vulnerability studies, an ambivalent critique. *Critical Horizons* 17, 2 (2016), 260–277.

[9] Mary FE Ebeling. 2016. *Healthcare and big data.* Springer.

[10] Batya Friedman, Peter H. Khan, and Daniel C. Howe. 2000. Trust Online. *Commun. ACM* 43, 12 (dec 2000), 34–40. https://doi.org/10.1145/355112.355120

[11] Sucheta Ghoshal. 2020. *A Grassroots Praxis of Technology: View from The South.* Ph.D. Dissertation. Georgia Institute of Technology.

[12] Omri Gillath, Ting Ai, Michael S Branicky, Shawn Keshmiri, Robert B Davison, and Ryan Spaulding. 2021. Attachment and trust in artificial intelligence. *Computers in Human Behavior* 115 (2021), 106607.

[13] Erinn Gilson. 2011. Vulnerability, ignorance, and oppression. *Hypatia* 26, 2 (2011), 308–332.

[14] Erinn Gilson. 2013. *The ethics of vulnerability: A feminist analysis of social life and practice.* Routledge.

[15] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 624–635.

[16] Karen Jones. 1996. Trust as an affective attitude. *Ethics* 107, 1 (1996), 4–25.

[17] Karen Jones. 2019. Trust, distrust, and affective looping. *Philosophical studies* 176, 4 (2019), 955–968.

[18] AnaLouise Keating. 2019. Foreword. In *Hungry translations: Relearning the world through radical vulnerability.* University of Illinois Press.

[19] Amanda Lazar, Norman Makoto Su, Jeffrey Bardzell, and Shaowen Bardzell. 2019. Parting the Red Sea: sociotechnical systems and lived experiences of menopause. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–16.

[20] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, et al. 2019. WeBuildAI: Participatory framework for algorithmic governance. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–35.

[21] Ann Light, Alison Powell, and Irina Shklovski. 2017. Design for existential crisis in the anthropocene age. In *Proceedings of the 8th International Conference on Communities and Technologies*. 270–279.

[22] Catriona Mackenzie. 2020. Vulnerability, Insecurity and the Pathologies of Trust and Distrust. *International Journal of Philosophical Studies* 28, 5 (2020), 624–643.

[23] Richa Nagar. 2019. *Hungry Translations: Relearning the World Through Radical Vulnerability.* University of Illinois Press.

[24] Richa Nagar and Roozbeh Shirazi. 2019. Radical vulnerability. *Keywords in Radical Geography: Antipode at 50* (2019), 236–242.

[25] Oxford Living Dictionaries. 2022. Vulnerability [Definition].

[26] Hartmut Rosa. 2020. *The uncontrollability of the world.* John Wiley & Sons.

[27] Piet Tutenel and Ann Heylighen. 2021. Interweaving vulnerability and everyday design: Encounters around an aquarium in a paediatric oncology ward. *Design Studies* 73 (2021), 101004.

[28] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L Arendt. 2020. How do visual explanations foster end users' appropriate trust in machine learning?. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 189–201.

[29] Daisy Yoo, Katie Derthick, Shaghayegh Ghassemian, Jean Hakizimana, Brian Gill, and Batya Friedman. 2016. Multilifespan design thinking: two methods and a case study with the Rwandan diaspora. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 4423–4434.