Destabilizing Auditing: Auditing artificial intelligence as care-ful socio-analogue/digital relation

Cheshta Arora[1], Debarun Sarkar[2]

[1]Independent Researcher, Norway

[2]Independent Researcher, India

DOI 10.3217/978-3-85125-976-6-03

Abstract. The paper aims to highlight the emerging figure of the 'expert auditor' in the field of AI ethics that seeks to legitimize artificial intelligence through a technocratic solution. The paper builds on de la Bellacasa's nonnormative notion of 'care as a provocation' to speculate what careful AI auditing could look like and in what ways it can allow us to destabilise normative AI auditing practices. By going back to the etymological root of 'to audit' to the Latin audio, i.e., to listen to, to pay attention to, we argue for a notion of auditing that's a narrow checklist solutionist notion of auditing artificial intelligence.

1 Introduction

An open letter published in early 2023 following the release of GPT-4 which demanded a pause to giant AI experiments quoted Asilomar AI Principles (*Future of Life Institute*, 2017) to suggest that "advanced AI could represent a profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources"(Russell et al., 2023). The authors of the open letter went on to point out that "[u]nfortunately, this level of *planning and management* is not happening, even though recent months have seen AI labs locked in an out-of-control race to develop and deploy ever more powerful digital minds that no one – not even their creators – can understand, predict, or reliably control" (Russell et al., 2023) (emphasis added). With this open letter, the threat of AI (embodied in GPT-4) and its management blew up in the popular discourse. The authors called "on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4" (Russell et al., 2023).

Such concerns vis-à-vis AI auditing and governance have been proliferating over the last decade. Given the series of high-profile cases that popularised AI harms and injustices and the advent of audit culture post-70s (Strathern, 2003), one can say that there is an established liberal consensus around the need to build robust auditing processes and

practices to govern AI. What is debatable, nonetheless, is the nature of these processes i.e., what is to be audited, by whom, how, and when. The open letter thus emerges within this context. What is crucial however is the phrase “commensurate care” and the discursive function that it plays.

The phrase follows the conditional phrase: "could represent" which evokes the uncertainty of AI’s future on earth, a bane or a boon, depending on how we proceed. In that sense, the phrase “commensurate care” could imply a warning, a caution, or an appeal to slow down in the face of imminent danger.

To care, however, is anything but straightforward. It raises a series of questions that can reformulate the very nature of the world and our relationship with it. However, before we move on and away from the letter, it is important to ask what are we being asked to care for? The following part of the letter offers a hint:

> “*Should* we develop nonhuman minds that might eventually outnumber, outsmart, obsolete and replace us? *Should we risk loss of control of our civilization? Such decisions must not be delegated to unelected tech leaders*. Powerful AI systems should be developed only *once we are confident that their effects will be positive and their risks will be manageable.”* (Russell et al., 2023) (emphasis added)

The above quote is rife with meaning. We, however, focus on three key temporal moments that serve the rhetorical structure of the above paragraph:

> *An imminent future*: “*Should* we risk loss of control of our civilization?”
>
> *The unfortunate present*: “Such decisions must not be delegated to unelected tech leaders”.
>
> *A future that can be*: “*AI systems should be developed only once we are confident that their effects will be positive and their risks will be manageable.***”**

Or to put it differently, the paragraph begins with a rhetorical strategy that pivots on blatant anthropocentricism (“our civilization”), a ‘we’ premised upon an empty ideal of ‘democracy’ strengthened by a sly reference to big tech organisation leaders (“unelected tech leaders”), and hope for restored status-quo, a happy ending implicit in the projected vision of the world where “effects will be positive” and “their risks will be manageable.” Thus, the letter is merely asking us to care for the existing status quo: the man at the centre of the world, worlding a world of manageable risks, and its subdued others—the animal, the machine, and other abject affiliates.

It can be said that concerns vis-à-vis auditing have remained at the level of governance and management, of mitigating risks and harms. Auditing culture, or the culture of management and accountability, that pervaded the workplace, public institutions and academic from the 1980s onwards had come under anthropological scrutiny only in the

2000s (Strathern, 2003). The debate pivoted around creativity vs. responsibility, drawing attention “to the inability of the audit culture to recognize experimentations in creativity in writing, research and teaching” (Ananta Giri, 2003: 179) and examining the political, social and cultural consequences of auditing practices. It brought to light the valuation of skills, bodies, objects, and behaviour inherent in the quantifying gestures of the culture of auditing.

A similar kind of anthropological scrutiny which asks if we should audit AI or what would a democratic, and transparent auditing of AI look like, however, is still lacking. The reasons for this, from a critical point of view, are *at least* twofold and depend on how AI is being perceived: 1) For those who see AI as just another tool deployed to intensify quantification of the world and existing social relations, auditing AI offers one way to resist this quantification. 2) For those who perceive AI as an emergent threat that can either “eventually outnumber, outsmart, obsolete and replace us,” auditing AI becomes a necessary recourse to tame this threat. Thus, both these positions vis-à-vis AI, i.e., the one that is critical of its quantifying gestures and ensuing harms and injustice, as well as the one that perceives AI as a threat, are in favour of establishing robust AI auditing mechanisms. A third position, that raises an important question of response-ability and perceives AI as an actor that allows us to reiterate the ethico-political question “who is and what is not considered to be a subject of rights and obligations” (Gunkel 2022), allows certain critical distance vis-à-vis AI auditing and clears the theoretical space for two questions:

- In the spirit of pursuing an egalitarian, democratic ideal, can auditing be more than merely managing risks?
- In the name of AI auditing and ethics, what are we being asked to care for?

Building a chain of interdependent intellectual debts, Henry & Oliver (2022) learn from de Bellacasa (who borrows from and reinterprets Tronto’s political theory on the ethics of care) to deploy care as a provocation rather than a moral stance or invocation for motherly love. Asking “how distribution of power, privilege and resources lead to inadequate care in society”, Henry et al via Bellacasa ask not only questions of ‘“For whom?” but also “Who cares?” “What for?” “Why do ‘we’ care?” and mostly “How to care?”’(Puig de la Bellacasa (2017) in Henry & Oliver (2022). We use Bellacasa’s nonnormative notion of ‘care as a provocation’, to speculate what careful AI auditing could look like and in what ways it can allow us to destabilise normative AI auditing practices.

The introduction is followed by four sections. The second and third sections tease out the normative assumptions underpinning the discourse of AI auditing and ethics. The fourth section develops a notion of auditing as ‘careful-analog/digital-relation’. The paper

concludes with a destabilised notion of AI auditing that disassociates it from the concerns of governance and management towards an ethico-political relation with the world.

2 AI Auditing and AI ethics: The problem of toolkits, guidelines, and checklists

The “gold rush” (Ayling and Chapman, 2022) in the field of artificial intelligence has been accompanied by a similar “gold rush” in the field of AI ethics. The unintended consequences, potential misuse of predictive systems, and the representative and socio-cultural biases reflected in the design of data-driven innovations fuelled the debate on AI ethics in the last few years. Today, the field has been divided into three phases: the first phase (2016-2019) operating in the applied ethics mode, resulted in a series of high-level principles, operationalized in the form of ethical statements and checklists, to ensure ethical, trustworthy, and responsible AI (Jobin et al., 2019). These principles and guidelines were issued by a range of actors but the research institutions played a primary role[10] with some involvement from the private and public sectors (Jobin et al., 2019) resulting in the emergence of an industry around AI ethics (Ayling and Chapman, 2022). The AI Ethics Guidelines Global Inventory maintained by AlgorithmWatch lists 167 guidelines published till 2020. They note, that only a handful of guidelines advocated for an oversight or enforcement mechanism (*AI Ethics Guidelines Global Inventory*, 2020). Via an analysis of 87 AI guidelines, Jobin (2019) identifies 11 ethical principles of transparency, justice and fairness, non-maleficence responsibility, privacy, beneficence, freedom and autonomy, trust, sustainability, dignity, and solidarity while noting a “substantive divergences among all 11 ethical principles in relation to four major factors: (1) how ethical principles are interpreted; (2) why they are deemed important; (3) what issue, domain or actors they pertain to; and (4) how they should be implemented” (Jobin et al., 2019).

The second phase witnessed “a more technical” shift with greater involvement from the computer science community “focussing on fairness, accountability, and transparency as an engineering ‘ethical-by-design’ problem-solving exercise” (Ayling and Chapman, 2022). This phase led to the development of computational definitions of ethical principles such as fairness and bias which contributed to a series of toolkits for detecting and mitigating algorithmic bias[11]. These toolkits, however, were quickly criticized for being

[10] See Algorithm Watch’s project AI Ethics Guidelines Global Inventory (https://inventory.algorithmwatch.org/)

[11] See AI Fairness 360, Fairlearn etc.

reductive and pivoting upon a certain fantasy of ‘technical fixes’ to social problems (Lee et al., 2021) which brought to light the need for interdisciplinary approaches to AI ethics.

From the dataset of 169 guidelines, Ayling and Chapman (2022) identified 39 AI auditing or impact assessment tools. One of the notable findings reported by their analysis points towards the limited group of stakeholders using the tool. The tools/toolkits developed during this phase were “clustered around the product development phase of AI (developers, delivery, quality assurance), with the output from the tools (reporting) being used by management Decision-Makers” (Ayling & Chapman, 2022) as well as “little participation in the assessment or audit process by certain stakeholder groups (Voiceless, Vested Interests and Users) who are not included in the process of applying the tools or interacting with the outputs as tools for transparency or decision-making” (Ayling & Chapman, 2022). A thematic shift was already reflected in their analysis as they posed the rhetorical question in the title of the paper, “Putting AI ethics to work: are the tools fit for purpose”? to argue for a range of auditing techniques that can incorporate more long-range and democratic principles of “participation process, baseline study, life-cycle assessment, change measurement or expert committees” (Ayling & Chapman, 2022).

Not surprisingly, the third and current phase has moved to the ‘how’ question focusing on “governance, mechanisms, regulation, impact assessment, auditing tools, and standards leading to the ability to assure and ultimately, insure AI systems” (Ayling & Chapman, 2022). Where the second phase emphasized self-regulation, on “organisation or profession marking its own homework” (Raab, 2020) there is an emerging consensus developing in the third phase vis-à-vis third, party expert auditing that is premised upon the conceptualizing of “ethics as a service” (Morley et al., 2021). This consensus is premised upon two assumptions: 1) an ethical evaluation requires “a lot of relevant and reliable information and quite a good management” of choices, responsibilities and moral evaluations (Floridi, 2017), 2) Auditing as facilitation wherein publicly available documentation can facilitate participation process.

3 AI as a social-technical system, auditing as a socio-technical practice

The three phases identified in the literature complement each other in producing a notion of AI as a socio-technical system and of auditing as a socio-technical practice. Where the first phase asked the ‘what’ question, the second phase, in asking the ‘how’ question, produced and legitimized auditable artefacts that can explain how decisions were arrived at, or systems and processes were implemented (Ayling and Chapman, 2022; Jobin et al., 2019; Morley et al., 2021). The third phase is working towards consolidating the first

two phases via a figure of a (human) expert that can comprehend AI as a socio-technical system and engage AI ethics as a socio-technical practice (von Eschenbach, 2021).

A notion of AI as a socio-technical system emerged in response to the reductive and technical fixes predominating the first and second phases. Opposing these technical fixes, the relevance of social scientific approaches and theories to understand algorithmic harms and injustice quickly brought to the centre the sociotechnical dimensions of AI. This third phase has worked towards shifting the focus away from the siloed visions of the technical to a more relational approach that insists on the co-constitutive nature of objects, bodies, processes, tools and relations (Birhane, 2021). A socio-technical approach decentred the anthropocentrism of a property-based normative ethical approach to insist upon the ontological primacy of relations. In a relational view, properties of an object—such as human, animal, or technological—emerge in relation rather than a priori. The increased stress on AI as a sociotechnical system has however not been able to bypass, remove or radically question the role of the expert in the sociotechnical system particularly in the discourse of auditing AI (Arora and Sarkar, 2023).

The relational approach opened the field of ethics to the ethico-political question of response-ability in the face of the other—the animal, the machine, the bot. The field of robot ethics that concerns itself with the problem of robots as moral agents served as a fertile ground to interrogate the limits of normative ethics and foreground the leaky nature of these categories—human, machine, animal—and the subsequent boundary work that goes into stabilising these categories.

As the debate shifted away from the “principled and the technical, to practical mechanisms for rectifying power imbalances” (Ayling and Chapman, 2022), AI as a sociotechnical system became a catchphrase for both relational as well as normative ethical approaches albeit with very different connotations. Where the relational ethicists insist on the socio-technical aspect of AI to foreground emergent relations that are deeply contextual, the normative ethical approaches describe AI ethics as the sociotechnical to foreground the part-whole problematic to argue that, “AI is *only part* of the decision-making process” (von Eschenbach, 2021). This allows von Eschenbach (2021) to argue not only that “If we are justified in trusting the socio-technical system, of which AI is a part, then we can still use AI for high-stake decisions because AI is not the sole decision-maker” but also foreground the figure of an expert as an inevitable part of the socio-technical system since for “most of us, our trust in AI will be mediated through our trust in the experts and their testimony about the trustworthiness of these technologies”.

These two divergent use cases of the catchphrase ‘AI as socio-technical systems’ lead to two different ethico-political visions of auditing. Whereas for normative ethicists, the introduction of an expert auditor signals the return of the ‘human-in-the-loop’, for

relational approaches the expert auditor signals the foreclosure of ethico-political possibilities that were emergent in this encounter between the human and the machine. Moreover, for the normative ethicists, the political hope remains in the assumption that an auditor will play the role of a (trustworthy) facilitator who will act as a rational, political conscience of the society and uphold the democratic principle, whereas, for the relational ethicist, the figure of an expert auditor betrays the democratic principle and signals the return of governmentality and strict ordering of bodies and relations.

4 Auditing as 'careful-socio-analog/digital-relation'

As mentioned previously, the third phase of AI ethics is coalescing around a notion of auditing that is premised upon a notion of AI as a socio-technical system. While there's still confusion vis-à-vis what different actors mean by auditing wherein processes such as impact assessment, risk management, and audit are all used interchangeably under a broader category of auditing (Carrier and Brown, 2021). However, it is clear that the meaning of AI auditing is being stabilised by envisioning an entire ecosystem of processes that will be implemented at multiple scales – of team, organization, industry and state and will include technical practices (of maintaining audit trails, SE workflows, verification and bias testing, explainable UI), management strategies (leadership commitment, training, internal reviews, industry standards), and external reviews (independent oversight, government regulation, auditing firms, insurance companies, NGOs & civil society, professional organizations) (Shneiderman, 2020). The ecosystem will give way to new process, best practices, legal and regulatory regimes, and experts who mediate between different scales, processes, and practices. As has been the experience with auditing practices in other industries such as finance, environment etc, it is quite possible that the ecosystem will be operationalized through an investment in, what Power (1997) had called, "shallow rituals of verification", "a form of learned ignorance", which is promoted "at the expense of other forms of organizational intelligence" (1997: 123). The open letter and how it imagines risk, uncertainty and control underscores "the programmatic faith in auditing" which is a symptom of "wider social anxieties and a need to create images of control in the face of risk" (1997: 121). At the same time, given the need for stability, the ethical quality of 'trustworthiness' is paramount in this ecosystem whose success depends on interdependent trust between teams, management, organization, government, citizens and auditing experts (von Eschenbach, 2021).

However, it is worthwhile to destabilise this vision of auditing by foregrounding other ethico-political relations emergent in this encounter with the machine, "a kind of external

threat proceeding from the future" that foregrounds the limits of the present (Gunkel, 2022). We believe the field of AI ethics can gain through such endeavours wherein auditing visions are critically destabilized (and disassociated from the normative vision of governance and regulations) when refracted through other ethico-political relations. In this paper, we focus on de Bellcasa's provocation of 'care' to foreground auditing as a 'careful socio-digital/analog' relation.

To suggest that the signifier and the notion of 'care' has been abused in the past few years as a commonsensical liberal (feminist) ethics would not be an exaggeration. To that extent, even the advocates of the predominant 'auditing as governance' approach will not be averse to a liberal notion of 'care' that is premised upon a "liberal vision of a subject as a moral agent" (Braidotti, 2006: 119) that is involved in "the static contemplation of the perpetuation of the regime of the Same" (Braidotti, 2006: 123). This vision of a subject is destabilized when the self-identifying moral agent is face to face with the provocative questions of care—"Who cares?" "What for?" "Why do 'we' care?" and mostly "How to care?"'(Puig de la Bellacasa, 2017). While today we more or less agree that we care for the socio-technical systems, to ask further what for, why do 'we' care and how is to point towards the limits of the top-down liberal vision that cannot answer "why should people care? How can one make them care? And what do we do with those who do not care at all?" (Braidotti, 2006: 119)—questions that point towards the vertical ordering of bodies and relations. To ask, 'how to care?' is first and foremost a political commitment. The reworking of the notion of 'care' into an analytics or provocation – where "'how to care?' is insistent but not easily answerable" (Puig de la Bellacasa, 2017: 7) – points to how the "ethics" in an ethics of care cannot be about a realm of normative moral obligations but rather about thick, impure, involvement in a world. It is not invested in maintaining the word as it is but asking what it could be.

Thus, to think of auditing as a 'careful relation' then would be to brace 'ethical vertigo', (Braidotti, 2006: 123) where we confront complex ethical dilemmas. It is in this light that it remains exigent to highlight the tension of analogue/digital wherein a processual fuzzy world is translated into a binary logic of good and bad. The discourse of auditing AI, if it takes the premise of socio-technical systems seriously, would need to acknowledge the always-shifting terrain of auditing and what is considered morally good and bad.

5. Conclusion: Destabilising AI auditing

While thinking AI often invokes both tech optimism and tech pessimism, we want to remain open to the novelty of this particular reconfiguration of social relations. If we agree with the basic premise that there is something qualitatively new that the current (and

future) generation of AI does to our relationships with the world around us then it remains exigent to retain a possibility of an encounter with the unknown.

To destabilise auditing necessitates interrogating its current preoccupation with notions such as truth, mediating through an expert and commonsensical evocation of principles like transparency, fairness, bias etc. Etymologically 'to audit' traces itself back to Latin *audio* i.e., to hear, to pay attention to, to attend to. This etymological notion of auditing is lost today amidst a checklist notion of auditing wherein auditing is seen as a mere process for accountability. While problems of stakeholder participation[12] have been raised by certain authors, auditing AI is slowly but surely emerging as a new profession (Phan et al., 2022) which to echo Preitl (2021) seeks to "preserve power". By introducing the figure of the expert auditor[13] normative AI ethics aims to make us care for AI at the expense of egalitarianism. It is crucial to highlight this foreclosure of the political in AI ethics. By inculcating a regime of regulation and governance normative AI ethics forecloses AI as a site of the political.

If we hold on to the etymological meaning of 'to audit', auditing wouldn't resemble current notions, images and practices of auditing as a straightforward, to-the-book normative process. To hear, pay attention to, to attend to, necessitates an open-ended approach to not just AI assemblages but also auditing assemblages. Auditing in such a case would evoke allied notions of research, investigation, being a detective, being curious and more than anything, exploring. Anyone could do that, fail, mess it up, or provide contradictory analyses. Such a notion of auditing would be open to AI assemblages and their cultures of care i.e., it would be open to the range of questions that de la Bellacasa raises. It very well might not resemble auditing anymore in its contemporary sense but would certainly attend to, pay attention to and hear AI assemblages.

Funding Disclosure

This material is based upon work supported in whole or in part by The Notre Dame-IBM Tech Ethics Lab. Such support does not constitute endorsement by the sponsor of the views expressed in this publication.

[12] This is not a place to rehearse this argument but the notion of stakeholder inclusion risks repeating consensus-based models of governance where there is no place for dissensus (Hillier, 2003).

[13] As Arora and Sarkar (2023) note, there remains something to be said about the tendency to foreclose discussions on propriety AI systems which remains at the heart of discussion on AI ethics.

References

AI Ethics Guidelines Global Inventory (2020) About. Available at: https://inventory.algorithmwatch.org/about (accessed 3 May 2023).

Ananta Giri (ed.) (2003) Audited accountability and the imperative of responsibility Beyond the primacy of the political. In: Audit Cultures: Anthropological Studies in Accountability, Ethics and the Academy. Routledge.

Arora C and Sarkar D (2023) Auditing Artificial Intelligence as a New Layer of Mediation: Introduction of a new black box to address another black box. Hipertext.net (26). 26: 65–68.

Ayling J and Chapman A (2022) Putting AI ethics to work: are the tools fit for purpose? AI and Ethics 2.

Birhane A (2021) Algorithmic injustice: a relational ethics approach. Patterns 2(2): 100205.

Braidotti R (2006) Transpositions: On Nomadic Ethics. Cambridge, UK ; Malden, MA: Polity Press.

Carrier R and Brown S (2021) Taxonomy: AI Audit, Assurance, and Assessment. For Humanity. Available at: https://forhumanity.center/web/wp-content/uploads/2021/09/ForHumanity.center_Taxonomy_AI_Audit_Assurance_Assessment.pdf.

Floridi L (2017) Why Information Matters. The New Atlantis (51): 7–16.

Future of Life Institute (2017) AI Principles. Available at: https://futureoflife.org/open-letter/ai-principles/ (accessed 3 May 2023).

Gunkel D (2022) The Symptom of Ethics; Rethinking Ethics in the Face of the Machine. Human-Machine Communication 4: 67–83.

Henry JV and Oliver M (2022) Who Will Watch the Watchmen? The Ethico-political Arrangements of Algorithmic Proctoring for Academic Integrity. Postdigital Science and Education 4(2): 330–353.

Hillier J (2003) `Agon'izing Over Consensus: Why Habermasian Ideals cannot be `Real'. Planning Theory 2(1): 37–59.

Jobin A, Ienca M and Vayena E (2019) The global landscape of AI ethics guidelines. Nature Machine Intelligence 1(9). 9. Nature Publishing Group: 389–399.

Lee M, Floridi L and Singh J (2021) Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics. AI and Ethics 1.

Morley J, Elhalal A, Garcia F, et al. (2021) Ethics as a Service: A Pragmatic Operationalisation of AI Ethics. Minds and Machines 31(2): 239–256.

Phan T, Goldfein J, Kuch D, et al. (2022) Introduction: Economies of Virtue. In: Phan T, Goldfein J, Kuch D, et al. (eds) Economics of Virtue – The Circulation of 'Ethics" in AI. Amsterdam: Institute of Network Cultures, pp. 6–22. Available at: https://networkcultures.org/wp-content/uploads/2022/12/EconomiesofVirtueINC2022TOD46-2.pdf.

Power M (1997) The Audit Society: Rituals of Verification. Oxford University Press.

Prietl B (2021) Why Ethics Norms are Not Enough, or: How Current Critique of Digital Data Technologies Preserves Power. In: Conference Proceedings of the STS Conference Graz 2021 (eds Cole, Nicki Lisa, Jahrbacher, Michaela, and Getzinger, Günter), 2021. Verlag der Technischen Universität Graz. Available at: https://diglib.tugraz.at/download.php?id=6188f5d25fb1b&location=datacite (accessed 4 May 2023).

Puig de la Bellacasa M (2017) Matters of Care: Speculative Ethics in More than Human Worlds. Posthumanities 41. Minneapolis: University of Minnesota Press.

Raab CD (2020) Information privacy, impact assessment, and the place of ethics∗. Computer Law & Security Review 37: 105404.

Russell S, Bengio Y, Marcus G, et al. (2023) Pause Giant AI Experiments: An Open Letter. In: Future of Life Institute. Available at: https://futureoflife.org/open-letter/pause-giant-ai-experiments/ (accessed 3 May 2023).

Shneiderman B (2020) Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems. ACM Transactions on Interactive Intelligent Systems 10(4): 1–31.

Strathern M (2003) Audit Cultures: Anthropological Studies in Accountability, Ethics and the Academy. Routledge.

von Eschenbach WJ (2021) Transparency and the Black Box Problem: Why We Do Not Trust AI. Philosophy & Technology 34(4): 1607–1622.