From Explainable AI to Explaining to AI (X2AI): Representational Practices in AI at Work

Ella Hafermalz, Marleen Huysman & Jana Retkowsky[3]

Abstract

This paper introduces the concept of "Explaining to AI" (X2AI) in the context of organizational and work environments, contrasting it with traditional "Explainable AI" (XAI). While XAI focuses on making AI systems transparent to human users, X2AI emphasizes the interactions where humans explain themselves to AI, specifically through representational practices of training, prompting, and feeding AI models. This shift highlights the political dimensions of representation and recognition within AI systems, stressing the need for AI to understand human contexts and identities. We discuss the implications of these representational practices for work and organizational studies, proposing future research avenues to address the sociotechnical dynamics of AI integration in workplaces in a way that goes beyond traditional emphases on transparency as an antidote to opacity.

Keywords: Explainable AI, XAI, recognition, transparency, future of work, Artificial Intelligence, Generative AI

Over the past decade, a key response to ethical concerns over the opacity of AI systems has been the need for "Explainable AI" (XAI). Explainable AI refers to the demand for AI models to be interpretable by human users, either by simplifying how AI systems work or by adding additional techniques that make the processes of such models inspectable and intelligible to developers, auditors, end users, and/or decision subjects. We have previously reflected on how XAI initiatives are often divided and talk past each other (Hafermalz & Huysman, 2022). Demands from policy makers and ethics commentators such as the European Commission have tended to diverge from the concerns and capabilities of technical developments emerging from, for example, DARPA's Explainable AI project.

A further concern that we raised at the time was how XAI conversations missed, and could benefit from, an added organizational perspective. Our point was that AI systems are often deployed in the context of work and organizing, yet both ethical and technical XAI initiatives tend to imagine a consumer context when developing solutions and policies. Therefore, they tend to assume that consumers will only interact with these systems in simplistic or indirect ways, for example taking advice from a probabilistic recommendation (such as a recommendation to watch a film or buy a product), or be a 'decision subject' of for example a positive or negative recommendation from a loan calculation. In this way it is often assumed that a consumer will simply 'accept' or 'reject' an AI system, which overlooks the socio-technical process of interacting with technology, particularly in the context of work, a perspective that we term "AI at Work".

Two years later, we maintain this position that efforts to make AI more transparent and explainable are important, and that this conversation deserves attention and contribution from the work and organizational studies research community. Yet we as a community and as a public are now *also* confronted with a new suite of Generative AI technologies that forces us to reconsider key assumptions, agendas, and recommendations for advancing research on AI at Work. If we look to policy makers concerned with the ethical implications of Generative AI technologies, such as the European Parliament's 2024 Artificial Intelligence Act, we again see an emphasis on transparency as a way to hold systems and the companies that run them to account (European Parliament, 2024). In addition to such concerns, and also in response to the unique qualities of Generative AI and its rapidly spreading role in work and organizations, we take this opportunity to outline a new concept that builds on previous Explainable AI conversations from an alternative ethical basis: *Explaining to AI* (X2AI).

Previous 'discriminatory' AI models provoked ethical concerns around transparency, which were met with a need for the model to be explained. The term X2AI however is grounded in our observation that new 'generative' AI models provoke a different type of interaction, that involves people explaining themselves *to* the model, in the form of training, prompting, and feeding these models with information. In this latter scenario, explaining 'oneself' *to* AI is, we argue, also an ethical act. Rather than being driven by a moral desire for transparency, Explaining to AI is driven by a need to be recognised, seen and understood - a politics of *recognition* (Butler & Athanasiou, 2013; Hafermalz, 2021; Suchman, 1995). Key differences between an ethics of transparency and a politics of recognition, including how these relate to Explanations and AI, are summarised in Table 1.

Politics of recognition concern being known, respected, and heard within a system (Baygi et al., in press; Fraser, 2008). This is tied up in identity politics because it involves making visible a particular identity within social and political discourse, usually with the aim of attracting rights such as access, assistance, or protection from discrimination. Because being visible is needed in order to be 'counted' in this way, Butler and Athanasiou (2013, p. 75) point out that being recognised via visibility and

[3] Kin Research Group - Vrije Universiteit Amsterdam

representation is something that we “cannot not want”. Being visible, known, and ‘accurately' represented within a system or data set, is an important part of being catered to. Such visibility is however a double-edged sword, because it can lead to intrusion on privacy, stereotyping, and constant demands to articulate who one is.

	Ethics of Transparency	*Politics of Recognition*
Ethical Concerns	Concerned with honesty, openness, accountability, and the integrity of processes that impact people's lives.	Concerned with justice, inclusivity, representation, recognition, and the impact of portrayals on marginalized groups.
Relationship to Explanations	Explanations are requested, to "give an account" of actions e.g., in an audit aiming to detect wrongdoing and/ or ways to make a system fairer	Explanations are offered, to define "who one is" and “what one wants” e.g., so that unique qualities and needs are recognised and catered to
Challenges in relation to AI	Ensuring that openness does not lead to information overload or the violation of privacy and confidentiality. Storing computations for possible future inspection and reporting is costly. Predictive capability may be reduced in efforts to make models more explainable. Some machine learning processes are not intelligible to humans.	Balancing the need of diverse populations to be seen and represented accurately and sensitively without perpetuating stereotypes, or encroaching on privacy. Ensuring that the full diversity of needs of different populations are recognised and included in AI systems is costly and political. Cultural contexts of development are likely to differ from contexts of use.

Table 1. Comparing an Ethics of Transparency to a Politics of Recognition in Relation to AI and Explanations

In the following, we conceptualize Explaining to AI and the politics of recognition by drawing on three representational practices that involve explaining to AI ‘who we are’ and ‘what we want’: training, prompting and feeding the AI. After introducing each practice, we provide ideas for future research avenues from an organisational perspective on how to further study X2AI.

Explaining to AI in Work and Organizing: Representational Practices of Training, Prompting, and Feeding

1. ***Training*** all forms of AI requires work. Apart from developing algorithms and models, critical research has shown the often undervalued manual labour of tagging, labelling, cleaning, and supervising the data flows that sustain AI systems (Justesen & Plesner, 2024). Usually attention is brought to these work practices to highlight the poor conditions under which some repetitive tasks are performed (Gray & Suri, 2019; Wood et al., 2019), and the lasting psychological damage that can be caused by exposure to extreme content that needs to be labelled so that others can be protected from it. An Explainable AI perspective would take an interest in such training work because how data is organised and named gives important clues to the source of biases, for example.

Yet considering the work of training AI from an X2AI perspective highlights that all forms of AI training (which remains largely hidden from the end user) are also means by which AI is taught to understand diverse human contexts, populations, needs, desires, and values (Tubaro, Casilli, & Coville, 2020). In Generative AI, unsupervised learning is the norm. However, data inputs are still overseen by humans, and fine-tuning is needed to ensure that outputs are in line with both practical and societal expectations. Training can therefore be seen as a representational practice of explaining to AI ‘who we are’, so that it can operate acceptably within the sociotechnical context in which it is deployed. Apart from further understanding this work of Explaining to AI in AI training processes, we also urge future research on the meta question of *how different forms of work are explained to AI*: what are we teaching AI systems about work and organizing?

Adding such a work and organizational perspective here draws attention to the following illustrative lines of inquiry: *how is work and organising being identified, captured, labelled, and organised in the training stages of (Generative) AI model development? What 'images' of work and organizing are being constructed through AI training processes? What are the (potential) implications of these constructions for the way that AI systems are deployed and used in work and organizing?* These questions emphasise that the work of training AI systems is important, also because of the manner in which such training 'teaches' AI to make sense of work (Barley & Kunda, 2001; Morgan, 1997; Suchman, 1995) and to act, at times autonomously, in organizational contexts during its deployment phase.

2. ***Prompting*** is the name that has been given to the conversational act of instructing Generative AI systems such as ChatGPT. 'Prompt engineering' has even been hailed as a new commercial skill that attracts consulting fees, microcredential certificates, and even saleable prompts that are created for purchase and use by others. The consequences of this relatively sudden appearance of prompting as a way of interfacing with AI are yet to be fully explored. From an ethics perspective, the act of 'conversing' with an artificial agent/chatbot has been viewed with suspicion, on the basis that these often sycophantic tools masquerade as if they 'know', or 'understand' what we ask them for, while in fact operating mainly on a probabilistic level by putting one probable word in front of another without any deeper capacity for comprehension or empathy (Roberts et al., 2024).

An X2AI perspective here highlights the iterative, conversational, and creative process (Pangaro, 2008, 2010) by which interactants try to make themselves and their goals clear and comprehensible to AI. We note however that the phrase 'prompt engineering' implies a strongly instrumental and largely one-way interaction, whereas our research on and experience with Generative AI tools thus far (Retkowsky et al., 2024) reveals a far more 'intertwined' relationship that is at play when for example ChatGPT is called upon for help, inspiration, advice, and feedback.

Rather than being a one-way act of instructing or ordering AI to carry out a task, it is often through chains of iterative prompts that we learn what it is we want in the first place. Through a repeated process of being misunderstood, clarifying, receiving erroneous or surprising outputs and providing feedback in response, a 'conversation' emerges that can lead the human instigator to places they did not expect. Cybernetics theorists have characterised such experiences as being fundamental features of good conversations-as-systems, where "We certainly want to know more or to understand more than when we started—if we are in the same place at the end of the journey, then what was the point?" (Pangaro, 2008, p. 37).

Appreciating the emergent and relational nature of explaining to AI means treating this representational practice as formative. Rather than merely 'telling' AI who we are or what we want, the process of interacting with AI shapes who we are and what we want. X2AI is in this way a political issue - because the act of representing oneself to a system means, at least to some extent, understanding oneself in relation to that system. People, and workers in particular, are therefore not merely prompting AI with instructions to receive a useful output. Rather, the act of telling AI about our tasks and requirements is also shaping work, as well as the worker. In Foucauldian terms, representational acts of explaining to AI constitute a process of subjectification that shapes the subject (Foucault, 1977). In anticipating what will 'make sense' to AI workers are, whether inadvertently or intentionally, thinking about their work in terms of that system. Their worker-self is in this way performatively and iteratively shaped in and through interacting with AI.

In sum, we contend that conversational and iterative acts of explaining *to* AI systems (such as ChatGPT, MidJourney, or Github CoPilot) what we want and need is increasingly becoming a part of daily working life and that these interactions are shaping how workers understand and practice their work, and themselves, in significant ways.

Questions that might be asked in future research on such a topic include: *Where do workers start versus where do they 'end up' when turning to Generative AI for assistance with a task? How does the repeated act of conversing with AI shape other collaborative interactions, processes, and subjectivities in an organization? How do system level prompts shape the 'interactional frame' of interacting with customised Generative AI systems? What are the (unintended) consequences of striving to make one's work explicable so that it is comprehensible by artificial agents? What kinds of worker-AI relations are evolving from these daily and at times frustrating collaborations?*

3 ***Feeding*** AI is the term we give to the act of end-users uploading files, documents, images, and other artefacts to Generative AI systems, in efforts to get things done. For example, a set of PDF files of academic articles may be uploaded with a request to compile a list of their similarities and differences. A profile photograph might be uploaded to a system such as Dall-E with a request for it to create a digital avatar likeness of the image. In some emerging artistic practices, images and descriptions of local scenes, people, accents, and artefacts are uploaded or fed to already trained systems to fine tune what the model comes out with in terms of the users' preferences, local context, and specialised requirements. We treat 'feeding AI' as analytically distinct from training in the sense that it involves end-users, and occurs after the initial model is trained, with the goal of tailoring a system to a particular use context.

This act of feeding AI with content that is important or relevant to oneself or one's community is a representational practice aimed at asking AI to "know me" or "know us". When understood in terms of X2AI, we can highlight how such feeding is tied to some of the downsides of recognition (Butler & Athanasiou, 2013), in particular how the need to be visible within a system in order to be catered to could come at the cost of privacy and ownership over personal data. Companies such as OpenAI are famously vague about how data that is fed to ChatGPT by users is handled, and employees who have fed company data into free personal GenAI accounts have been reprimanded for 'leaking' information (Krietzberg, 2024; Ray, 2023). Yet in framing such acts of feeding as representational practices we offer an additional, alternative perspective to such a focus on the ethics of privacy and security, which helps to make sense of why workers continue to offer information to AI even given these risks.

We are currently studying an organisational implementation of Microsoft Co-Pilot in a media organisation - following along at training sessions and conducting qualitative interviews with those who have early access to this tool. Several participants that we have spoken to in these early stages of our study have been disappointed that Co-Pilot (so far) does not seem to have 'read' their stored files and emails to the extent that it can mimic their tone and style of writing emails. Apart from the convenience of having an AI agent that can convincingly write an email that passes as personal correspondence, we identify here a more fundamental interest in having AI systems, at least on some level, 'understand' us. Consider what it was like trying to interact with the original Siri or Alexa voice assistants, particularly with an accent or language other than American English. Such experiences of non-recognition are jarring to one's sense of identity and belonging. Now that AI agents are suddenly far more capable, we are witnessing amongst users a willingness and even eagerness to explain to AI everything it needs to compute, in order to better fulfil our requests, even and perhaps particularly in the workplace.

A final set of illustrative questions that relate to practice of feeding AI in a work and organizational context includes: *What information and artefacts are employees willingly sharing with AI? How/is the feeding of artefacts used to (try to) shape AI's 'local knowledge' of organizational and national/regional culture, for example by uploading local lexicons or onboarding manuals? When do misunderstandings and conflicts occur in relation to fed artefacts, and how are these breakdowns dealt with?*

Conclusion

Explaining to AI (X2AI) is fast becoming a skilled and significant kind of work. Workers are now training, prompting, and feeding a variety of Generative AI systems in efforts to make human contexts, interests, and aims intelligible to machines. This work reflects a politics of recognition that is impactful, because how AI sees and understands us is becoming increasingly important for how work gets done. On an individual level, AI is now often 'speaking for us' as generated content is posted and sent to colleagues and clients. Workers therefore have an added task of taking care of how AI systems represent them in systems of communication. Models that have been trained in one context, with a particular notion of for example what it means to work, collaborate and interact with others, need to be taught and tailored for local organizational and cultural contexts. Will local quirks, accents, mannerisms, and signs of personal attentiveness and care be lost, in favour of generic corporate speak and smooth AI imagery? The answer depends largely on how ongoing efforts to explain to AI proceed.

References

Barley, S. R., & Kunda, G. (2001). Bringing work back in. *Organization Science, 12*(1), 76-95.

Baygi, R. M., Introna, L. D., & Ostovar, M. (in press). Beyond categories: A flow-oriented approach to social justice on online labour platforms. *MIS Quarterly*.

Butler, J., & Athanasiou, A. (2013). Dispossession: The performative in the political: John Wiley & Sons.

European Parliament. (2024, March 13). *Artificial Intelligence Act: MEPs adopt landmark law* [Press release]. https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law

Foucault, M. (1977). Discipline and punish: The birthof the prison. *Trans. Alan Sheridan. New York: Vintage-Random*.

Fraser, N. (2008). Social justice in the age of identity politics: Redistribution, recognition, and participation. In *Geographic Thought* (pp. 72-89). Routledge.

Gray, M. L., & Suri, S. (2019). Ghost work: How to stop Silicon Valley from building a new global underclass. Eamon Dolan Books.

Hafermalz, E. (2021). Out of the Panopticon and into Exile: Visibility and control in distributed new culture organizations. *Organization Studies, 42*(5), 697-717.

Hafermalz, E., & Huysman, M. (2022). Please explain: Key questions for explainable AI research from an organizational perspective. *Morals & Machines, 1*(2), 10-23.

Justesen, L., & Plesner, U. (2024). Invisible Digi-Work: Compensating, connecting, and cleaning in digitalized organizations. *Organization Theory*, 5(1), 26317877241235938.

Krietzberg, I. (2024, January 30). ChatGPT is leaking users' passwords, report finds. *TheStreet*. https://www.thestreet.com/technology/chatgpt-sam-altman-artificial-intelligence-privacy-ethics-passwords

Morgan, G. (1997). *Images of Organization*. California: SAGE Thousand Oaks.

Pangaro, P. (2008). Instructions for design and designs for conversation. In *Handbook of conversation design for instructional applications* (pp. 35-48): IGI Global.

Pangaro, P. (2010). How can I put that? Applying cybernetics to "Conversational Media". *Cybernetics & Human Knowing, 17*(1-2), 59-75.

Ray, S. (2023, May 2). Samsung bans ChatGPT and other chatbots for employees after sensitive code leak. *Forbes*. https://www.forbes.com/sites/siladityaray/2023/05/02/samsung-bans-chatgpt-and-other-chatbots-for-employees-after-sensitive-code-leak/

Retkowsky, J., Hafermalz, E., & Huysman, M. (2024). Managing a ChatGPT-empowered workforce: Understanding its affordances and side effects. *Business Horizons*. Advance online publication. https://doi.org/10.1016/j.bushor.2024.04.009

Roberts, J., Baker, M., & Andrew, J. (2024). Artificial intelligence and qualitative research: The promise and perils of large language model (LLM)'assistance'. *Critical Perspectives on Accounting, 99*, 102722.

Suchman, L. (1995). Making work visible. *Communications of the ACM, 38*(9), 56-64.

Tubaro, P., Casilli, A. A., & Coville, M. (2020). The trainer, the verifier, the imitator: Three ways in which human platform workers support artificial intelligence. *Big Data & Society, 7*(1), 2053951720919776.

Wood, A. J., Graham, M., Lehdonvirta, V., & Hjorth, I. (2019). Good gig, bad gig: autonomy and algorithmic control in the global gig economy. *Work, Employment and Society*, *33*(1), 56-75.