DRS Biennial Conference Series

DRS2022: Bilbao

Jun 25th, 9:00 AM

From explanations to shared understandings of AI

Iohanna Nicenboim
Delft University of Technology, The Netherlands

Elisa Giaccardi
Delft University of Technology, The Netherlands

Johan Redström
Umeå Institute of Design, Sweden

Citation

Nicenboim, I., Giaccardi, E., and Redström, J. (2022) From explanations to shared understandings of AI, in Lockton, D., Lenzi, S., Hekkert, P., Oak, A., Sádaba, J., Lloyd, P. (eds.), *DRS2022: Bilbao*, 25 June - 3 July, Bilbao, Spain. https://doi.org/10.21606/drs.2022.773

DRS 2022
Bilbao 25th June — 3rd July

From explanations to shared understandings of AI

Iohanna Nicenboim[a,*],Elisa Giaccardi[a], Johan Redström[b]

[a]Delft University of Technology, The Netherlands
[b]Umeå University, Sweden

*corresponding e-mail: i.nicenboim@tudelft.nl, e.giaccardi@tudelft.nl, johan.redstrom@umu.se

doi.org/10.21606/drs.2022.773

Abstract: A key challenge in the design of AI systems is how to support people in understanding them. We address this challenge by positioning explanations in everyday life, within ongoing relations between people and artificial agents. By reorienting explainability through more-than-human design, we call for a new approach that considers both people and artificial agents as active participants in constructing understandings. To articulate such an approach, we first review the assumptions underpinning the premise of explaining AI. We then conceptualize a shift from explanations to *shared understandings,* which we characterize as situated, dynamic, and performative. We conclude by proposing two design strategies to support shared understandings, i.e. looking across AI and exposing AI failures. We argue that these strategies can help designers reveal the hidden complexity of AI (e.g., positionality and infrastructures), and thus support people in understanding agents' capabilities and limitations in the context of their own lives.

Keywords: explainability; artificial intelligence; everyday life; more-than-human design

1. Introduction

From everyday services to critical domains, Artificial Intelligence (AI) is used to make decisions that profoundly affect people’s lives. However, since many machine learning techniques are not interpretable, one of the biggest challenges in designing AI systems is to be able to explain those decisions. To tackle this problem, researchers, governments, activists, and companies have called for making AI more explainable (Barredo Arrieta et al., 2020; Whittaker et al., 2018; Amershi et al., 2019). While researchers have found ways to make AI models somehow interpretable for developers, how to explain AI decisions to a broader range of people, including everyday users and people affected by AI decisions who are not direct users, remains a significant challenge (Bhatt et al., 2019; Vera Liao et al., 2020).

Technical explanations can be useful for developers, but they are not a good fit for a broader range of people (Mittelstadt et al., 2018). One simple reason is that people affected by the decisions of an AI system might be less concerned with how that model works in a technical sense, and instead, might want to know if they could trust the decisions the model produced. Thus, designers need to find alternative ways to support people in making sense of AI that go beyond technical explanations. Indeed, the challenge of supporting people in making sense of technologies, in contrast to showing how technologies simply work, has been a concern of designers for quite some time. In the case of the graphical user-interfaces, for example, there are considerable differences between how files, folders, and functions are presented and how a computer stores data. But given the complexity of infrastructures behind AI systems, how these systems 'model the world' and change over time, situated understandings of AI might have to differ considerably from strictly technical explanations.

In this paper, we argue that in order for designers to address the challenge of making AI decisions understandable to a broad range of people, we first need to position explanations in everyday contexts and within interactions of people and intelligent agents. This requires designers to address how understandings may originate from people's everyday experiences when interacting with AI, and to give an active role to users and artificial agents in the process of understanding AI. By adopting a more-than-human design orientation, we propose to consider explanations within ongoing relations of humans and artificial agents, in which both people and artificial agents are active participants in building understandings.

1.1 Reorienting explanations through more-than-human design

In the last few years, Explainable AI (XAI) has become a growing area of interdisciplinary research, one concerned with enabling human users to understand, appropriately trust, and effectively manage AI decisions (Turek, 2020). A key obstacle in this domain is the tension between the ambition of XAI to provide explanations to all users, and the current state-of-the-art which focuses on explanations mostly for experts (Abdul at.al., 2018; Barredo Arrieta et al., 2020; Mittelstadt et al., 2018). To address this gap, there are calls for expanding XAI by integrating and developing methods together with research from other fields, such as social and cognitive sciences (Miller, 2019; Mittelstadt et al., 2018), and interaction design (Abdul et al., 2018).

Design approaches to XAI include a range of human-centered design frameworks and tactics. Among them, scenario-based XAI (Andres et al., 2020), user's trust (Weitz et al., 2019), usability guidelines for explainable interfaces (Amershi et al., 2019), and tangible and embodied interactions (Ghajargar et al., 2021). These approaches are now organized around the community of human-centered XAI (HCXAI) (Wang et al., 2019, Ehsan and Riedl, 2020). Scholars in HCXAI have called for a shift from algorithm-centered XAI to socially-situated XAI (Ehsan et al., 2021). Their research has shown that existing work on "opening" the black-box of AI does not properly address how to design explainable AI for different user groups and how to make explanations actionable. The more inclusive and empowering approach of HCXAI has highlighted the importance of investigating explainability in relation to user's background (Ehsan and Riedl, 2020) and user's experiences (Vera Liao et al., 2020).

To complement these efforts and to respond to the call of situating XAI, we adopt a more-than-human design approach to technology (Giaccardi and Redström, 2020; Wakkary, 2021; Forlano 2017; Coulton and Lindley 2019, DiSalvo et al., 2011). A more-than-human design orientation calls for including nonhuman perspectives (such as the ones of everyday things) to problematize the design space, to unsettle assumptions, expose human-originating biases, and "demonstrate the problem to be more uncertain, more nuanced or more complex than originally assumed or regarded" (Giaccardi, 2020, p. 104).

More-than-human design can help to re-orient explanations, by positioning them within ongoing relations between human and artificial agencies (Giaccardi and Redström, 2020; Kuijer and Giaccardi, 2018); and by acknowledging the active role of *both* humans and agents in ordering, resisting, enabling and mediating everyday life (Frauenberger, 2019; Giaccardi and Redström, 2020). A more-than-human approach is also useful to grapple with questions of agency in explainability (Lindley et al., 2020). Given that artificial agents are not just tools, but entities capable of self-initiated action, explaining AI as if it was a mere tool would not be effective (Redström and Wiltse, 2018; Kuijer and Giaccardi, 2018). For example, when explaining why a digital assistant (such as Amazon Echo) responds better to certain voices than others, we might fall short if we overlook the possible biases that are amplified by the training data set, and the gendered stereotypes that these devices perpetuate (Strengers and Kennedy, 2020). Explaining interactions with AI agents as simple tools might lead to unethical designs, because it masks the sort of delegations that are in place when interacting with smart systems (De Mul, 2010), and the implications that that might have in people's lives (Pierce and Di Salvo, 2018, Giaccardi, Kuijer, and Neven, 2016).

1.2 From explanations to understandings

In XAI research, authors have highlighted the central importance of the notion of understanding, and how that differs from the notion of explanation (for a summary see Langer, 2021). Among these scholars, Páez (2019) investigates the relationship between explanation and understanding in the context of opaque machine learning models. He proposes to differentiate explanation and understanding as a way to open up new avenues of research that can lead to better grasping the workings and decisions of opaque AI models. The reason for this call is that, unlike explanation, understanding can include other paths (e.g., models, simulations, or experiments) which do not need to be accurate representations of a phenomenon, as long as they are useful for organizing human experience.

To unpack the shift from explanations to understandings for the design of AI, we argue that it is important to review the epistemological commitments underpinning the very premise that AI should be explained. This is crucial to make AI understandable for a wide range of people because the premise of explaining AI might fundamentally assume a passive role for people (and AI agents) by privileging the perspective of the ones building the system over the ones affected by it. Thus, in the next section, we unpack some seemingly hidden assumptions in relation to 'what' is explained and to 'whom.'

2. What is assumed in the premise of explaining AI?

In order to consider a different approach, we first need to unpack the logic on which the call for explaining AI rests. When articulating the limitations of transparency in AI, Ananny and Crawford (2018) argue that no model of accountability can avoid the questions of "accountable for what?" and "accountable to whom?". Following this argument, we review some of the epistemological commitments that currently underpin the idea of explaining AI, first in relation to what needs to be explained and then in relation to who are the explanations for, and who creates them.

2.1. What is explained?

Let us begin by asking to what extent it is reasonable to assume that explanations need to be explicitly causal, factual, and simple.

In the principle of explaining AI there is a basic assumption that explanations are strictly the ones that can state the cause of a phenomenon, i.e., that understanding AI systems is to know what decisions were made and why. This idea is based on a historical connection between explanations and causes, beginning with Aristotle and continuing with the defenders of causal explanations, which have argued that understanding cause implies knowledge (Páez, 2019). However, causal knowledge does not come exclusively from explanation. On the other hand, there are other methods, such as experimentation, which are not explicitly causal but can be useful to understand AI systems. For example, direct manipulation, such as adjusting a lever and observing its effects on other parts of a system, is a way of understanding how a system works. Further, manipulating a system into new desired states it is a sign of understanding which requires the ability to think counterfactually (Páez, 2019).

Another assumption is related to the factual character of explanations, i.e., to think that the more facts revealed about a phenomenon, the more it will be understood. This is because factivity is an essential feature of explanations in science. However, finding a complete technical explanation (in the traditional sense) of AI models is impossible, as most of them are designed as black-boxes (Páez, 2019). Furthermore, even if we could 'look inside' deep-learning models, we would not be able to understand them beyond a temporal constrain, because AI systems are constantly changing and evolving.

Interaction design has developed approaches to support understanding computational systems in ways that go beyond, or circumvent, technical types of explanations. One prominent example is the desktop metaphor as used in a typical graphical user interface (GUI). But since AI systems constantly change and evolve by learning through interaction with people, we cannot precisely define, nor completely determine, what that understanding needs to be like, and then build an interface on that definition. Even if we could explain the source code, training data set, and testing data that helped build those agents, this would describe only some particular aspects of it. Such snapshots of a system tell us little about its logic, like how it will respond in the future, and how it will change in relation to new data.

Finally, there is an assumption that the simpler the explanations the better. The vast majority of work in XAI is based on creating simplified approximations of complex decision-making functions (Mittelstadt et al., 2018). These are useful for developers, both for pedagogical purposes and for making reliable predictions of how the system might behave over a restricted domain. However, they can be misleading when presented as an explanation of how the model functions to everyday users and people affected by AI decisions. Mittelstadt and colleagues argue that the simplified approximations resemble more scientific models than 'everyday explanations,' which are contrastive, selective, and social. Thus, along with other scholars, they point to the importance of supporting debate and contestation of AI decisions as productive strategies to achieve understandings (Mittelstadt et al., Vaccaro et. al, 2019; Lyons et al., 2021).

2.2. Explainable to whom?

Given the issues raised above, we also need to question for whom, and by whom, such explanations are made.

The depiction of a neutral user in various XAI diagrams is based on the logic that users can be generalized and understood as single and neutral. This model assumes that the AI system, or the developers who built it, are the ones that deliver knowledge, while users are just recipients of it. It also assumes that general explanations can fit all users independently of their identity and position in the world. This tends to obscure the fact that not everyone can, or will, interpret information in the same way. Furthermore, it simply ignores that many people use AI systems indirectly (Aizenberg and van den Hoven, 2020), and therefore have no access to such information.

What is implicitly assumed in the aim of explaining AI is that the ways in which end-users understand AI are less accurate than the experts who build it (Adam, 1993). While this may be the case for technical explanations, it is in contrast to the 'standpoint' type of knowledge of much research in design, where the implicated subject is considered as the expert of their own domain (Suchman, 2006). Privileging the perspective of those who design and build the systems over alternative views poses the risk of reinforcing implicit biases and preserving socially legitimated knowledge, and offers limited scope to consider alternative understandings (Adam, 2020).

3. Towards shared understandings of AI

As discussed above, traditional explanations might not be the only or most effective way to support people in understanding artificial agents in everyday life. Moreover, traditional explanations might not be inclusive enough, because they cannot account for the multiple people implicated by AI, and the fact that not everyone can understand the workings of a system or its significance in the same way. So, what could be a more inclusive alternative?

In order to explore this alternative, we conceptualize a shift from explanations to *shared understandings*. When looking at 'who are the explanations for' and 'how is that knowledge produced,' there seems to be a need for addressing understandings in plural. It seems that there is no single explanation that could fit all, but that there is a need to design for the possibility of multiple understanding(s) to be produced within ongoing relations between situated users and artificial agents. Indeed, with respect to these systems fulfilling their intended roles, the multiple ways people implicated by AI may come to understand what such systems are and what they do as part of everyday life is no less critical or legitimate than supporting the experts that have built, or are in charge of, the systems.

To unpack the idea of *shared understandings* and find new spaces for designing more understandable AI systems, we draw upon work done at intersections between information technology and other areas such as science and technology studies, philosophy, and posthumanist and feminist theory.

In her studies of situated action, ethnomethodologist Lucy Suchman (1987) argues that the understanding that arises from interaction with technology should always be regarded as a practice of knowledge production: "The coherence of situated action is tied in essential ways [...] to local interactions contingent on the actor's particular circumstances. A consequence of action's situated nature is that communication must incorporate both a sensitivity to local circumstances and resources for the remedy of troubles in understanding that inevitably arise" (Suchman, 1987, p.27-28).

As argued by political philosopher Hannah Arendt (1954/1994), understanding something is a dynamic process. "Understanding, as distinguished from having correct information and scientific knowledge, is a complicated process which never produces unequivocal results. It is an unending activity by which, in constant change and variation, we come to terms with and reconcile ourselves to reality, that is, try to be at home in the world" (Arendt, 1994, p.307).

With respect to articulating situated understandings, Donna Haraway's work is central. Haraway (1988) describes knowledge as always situated--that is, produced by positioned actors working up/on/through all kinds of relation(ships). What is known, and how it can be known, are both subject to the position of the knower, i.e., their situation and perspective. This points to a more-than-human epistemology, according to which multiple understandings need to be situated within contexts and shared by agents (humans and nonhumans) that are always differently positioned and in different relation(ships) to each other.

When thinking about shared understandings of AI from a more-than-human design orientation, we can start considering the agential role that *both* people and artificial agents play in situated practices of knowledge production. Thus, it seems important not only to account for the positionality of the people implicated by AI, but also the positionality of the agents. This trajectory suggests that instead of technical explanations, users could benefit from understanding who owns the infrastructures behind the devices they use, and what are their limitations. Thus, we see situated understandings as the ones that can expose the different dimensions of artificial agents, from their identities and worldviews, to the larger infrastructures in which they are embedded.

4. Design strategies to support shared understandings of AI

In this section, we discuss some implications that the shift from explanations to shared understandings might have for designing more-than-human interactions with AI. While in the first sections we have described a conceptual shift, in this section we focus on two design strategies, as possible ways to expand the scope of explanations. The unpacking of these strategies is supported by examples of speculative work in art and design aimed at creating tension and calling out the hidden complexity of AI (e.g., the infrastructures, positionality, and limitations of artificial agents), which the premise of explaining AI is currently not accounting for.

4.1. Looking across AI

As discussed in the previews sections, positioning understandings as part of broader socio-technical systems brings to the front that we need to move past the idea of a neutral user, and instead try to paint a more diverse and inclusive picture of the different actors implicated by AI. Thus, to make AI more understandable, designers could try to account for the multiple agencies (humans and nonhumans alike) that are part of making a particular interaction with AI.

Ananny and Crawford describe this socio-technical approach as “looking across the AI system” instead of looking inside (2018). They ask: *“What is being looked at, what good comes from seeing it, and what are we not able to see?”* (Ananny and Crawford, 2018, p. 13). Those questions are visualized in the “Anatomy of an AI system” (Crawford and Joler, 2018). The map visualizes what we are not able to see when we interact with an Amazon Echo, namely the extractive processes of material resources, human labor, and data that are required to build and operate it. In a similar map called “Architectures of choice”, Marenko and Benque (2019) use diagrams to trace recommendations on YouTube as experiments to explore what new understandings are created. They foreground relations and paths to build a mode of knowledge-making that is situated, incomplete, and speculative.

These maps bring to light new questions for designing AI agents. How can we design explanations that situate not only users but also agents? For example, how might interacting with Amazon Echo reveal its ecosystem, biases, beliefs, and worldviews? Can we image different interactions with conversational agents if instead of ‘closed’ products we allowed users to intervene in different parts of this map (e.g., training the algorithm)? What if instead of ad-hoc explanations, agents could rely on local contexts to develop shared and situated understandings together with their owners?

4.2 Exposing AI failures

Mapping the entangled relations of AI systems means also making visible its boundaries. In the Nooscope map, Joler and Pasquinelli (2020) created a diagram of Machine Learning errors, biases, and limitations. They describe their project as a cartography of the limits of AI, to illustrate not only how AI works but also how it fails. In this section, we try to articulate some benefits that exposing AI's limitations might reveal for users and designers. We ask: *What kind of understandings can be developed by exposing what is hidden? What could be the value of such friction for users?*

Knowing the limitations of an AI system seems crucial for understanding it, since users can adjust expectations and calibrate their trust (Luger and Sellen 2016). In fact, making clear what a system can do (and how well) are among the first principles identified by Amershi and colleagues in the Guidelines for human-AI interaction (2019). Technologies running into their limits can be found also in science fiction, for example, in Isaac Asimov's series 'I Robot' or the popular series Black Mirror. But failures can do more than just motivate users to seek explanations. Beyond exposing the frictions of AI, breakdowns in the interaction with agents could be a way to move closer to shared understandings, because when artificial agents fail in everyday interactions, both humans and agents need to be actively involved in repairing practices.

Empirical studies with conversational agents have shown that people actively try to repair breakdowns by using different strategies, such as modifying the words and the tone of voice they use (Sciuto et al., 2018; Luger and Sellen, 2016). People take an even more active role in relation to artificial agents when testing their limits to understand their capabilities, and probing how 'intelligent' the agent is (Druga et al., 2017; Porcheron et al., 2018; Pelikan and Broth, 2016). For example, common strategies to test conversational agents involve using convoluted sentences (such as when asking about the weather using unusual expressions) to see if the agent can handle them. How could the design of artificial agents support this kind of experimentation?

4.3 Illuminating strategies through design examples

As a design strategy, 'looking across AI' is an invitation for designers to see AI as a socio-technical system. This shifts the role of design from masking the complexity of AI systems in seamless interactions, to exposing that complexity by revealing the infrastructures and tensions that are part of it. For example, Desjardins and colleagues (2021) used interdisciplinary performance to critically examine conversational agents, from what is physically hidden inside the speakers, to the hidden labor and the surveillance practices that are behind them. In the design exploration "AYA" and "U", Juul Søndergaard and Hansen (2018) explored different ways a voice assistant may push back on sexual harassment using design fictions tactics to reveal issues of trust, gender and algorithmic bias. In the performance called "Lauren", McCarthy (2017) installed a series of networked smart devices in people's homes and remotely watched over them, to reveal issues of surveillance, decision making, and privacy.

In “The sound of speech as it echoes in the cloud,” the collective Tropozone (2021) exposed the frictions of the ecological emergency through a network of voice assistants that are geographically distributed, inviting people to reroute their attention to the overlooked, the unfamiliar, and the forgotten.

From a more-than-human design orientation, ‘looking across AI’ is a strategy that invites designers to think that AI systems do not only contain complexity but enact complexity, by “connecting to and intertwining with assemblages of humans and non-humans” (Ananny and Crawford, 2018, p. 2). Thus, looking across AI involves accounting for the situated encounters that different humans and nonhumans agencies have when they relate to each other. An example of an inquiry into the agencies involved in human-AI interactions was explored in a series of design workshops held at the Research Through Design (RTD) conference in 2019 and the Designing Interactive Systems (DIS) conference in 2020. In the RTD workshop “Encountering ethics through design”, Reddy and colleagues (2020) have invited participants to co-speculate with intelligent things, by enacting things in different scenarios, giving them an active role in their interactions. In doing so, the workshop helped to consider autonomous behavior not as a simplistic exercise of anthropomorphization, but within the more significant ecosystems of relations, practices and values in which intelligent things are involved and through which they are encountered. In the DIS workshop “More-than-human design and AI”, Nicenboim and colleagues (2020) used speculative interviews to interrogate conversational agents. Their inquiry addressed issues of biases, ownership, and responsibility within conversations, along three dimensions, i.e., how agents present themselves to humans; what relations and ecologies they create within the contexts in which humans use them; and what infrastructures they need. By looking across AI these workshops have highlighted that the decisions AI agents make are both part of complex infrastructures, and yet situated in intimate encounters in people’s homes.

As a strategy, ‘Exposing AI failures’ offers a tangible way to provoke frictions in the everyday encounters between people and AI and to probe how people could have an active role in trying to understand AI. In other words, exposing failures provokes situations where understandings can be co-constructed within ongoing relations between people and AI. For example, in “Project alias,” Karmann and Knudsen (2018) designed a parasite that feeds smart speakers with white noise while allowing users to train custom wake-up names. In “Autonomous trap 001,” James Bridle (2017) challenged self-driving cars which rely on machine vision, by drawing a circle to trap the vehicle inside. In the projects “Objects that withdraw” and “Unpredictable things,” Nicenboim (2017) has explored the limits of object recognition algorithms by creating occlusions using different materials and shapes to modify everyday things until they are not recognizable by machines. In contrast to spontaneously testing the limits of a product for understanding what can it do, these design practices intentionally challenge and obfuscate AI (Brunton et al., 2017) to question conventional approaches to social, ethical, or political issues (Di Salvo, 2012).

The design projects presented in this section reveal new directions for supporting shared understandings of AI. We have unpacked the strategies 'looking across AI' and 'exposing AI failures' through design examples, and discussed how they could help people to critically and actively engage with AI's complexity, grappling with questions of ethics and values, social perspectives, and politics in specific contexts. One might argue that the proposed design examples remain tentative in going further with the material of AI – by watching, interviewing, traversing how does the material of AI become part of design practice. In other words, one might argue that they circumvent a system and take a position of critical distance rather than deep engagement. For future work, it will be important to consider how we might support shared understandings between people and AI in everyday life not simply to make their criticalities tangible but rather to account for the effects of the system.

5. Conclusions

While explaining how AI systems work is important, we have argued in this paper that technical explanations might not be enough to support people in understanding AI. In the first part of the paper, we use a more-than-human design lens to highlight the importance of reorienting explainability in the context of everyday relations between people and AI, and considering both people and artificial agents as active participants in creating a shared understandings. Based on that call, we reviewed the assumptions behind the idea of 'explaining AI' which assumes a passive role for people and privileges the perspective of those who design and build the AI system over people affected by it. By considering 'what' is explained and to 'whom,' we have argued that the current approach to explainability cannot account for all the people implicated by an AI system, because it ignores that not everyone can understand the system in the same way. Contrary to these assumptions, we have pointed to work from AI research arguing that there are many people implicated by AI decisions, and all of these people have ways of engaging with and understanding AI systems that are no less legitimate than those of the experts that have developed the model.

To help designers bypass these assumptions, we have proposed a shift from traditional explanations (which are typically factual, causal, and technical) to shared understandings, ones that are situated, relational, and dynamic. Lastly, we have drafted a series of alternative strategies for designers to support shared understandings between people and AI in everyday life. These strategies are based on exposing the different agencies and infrastructures involved in the making of AI, as well as the positionality of artificial agents, including their biases. We believe this will open a novel design space for supporting shared understandings by showing not only how AI works and fails, but also accounting for the effects of the system and giving handles to people for navigating them in the context of their own lives.

Acknowledgements: This work is supported by a Microsoft Research PhD Scholarship.

6. References

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18, 1–18.

Adam, A. (1993). Gendered knowledge: Epistemology and artificial intelligence. *AI & Society*, *7*(4), 311–322.

Adam, A. (2000). Deleting the Subject: A Feminist Reading of Epistemology in Artificial Intelligence. *Minds and Machines*, *10*(2), 231–253.

Aizenberg, E., & van den Hoven, J. (2020). Designing for Human Rights in AI. In *arXiv [cs.CY]*. arXiv. http://arxiv.org/abs/2005.04949

Amershi, S., Inkpen, K., Teevan, J., Kikin-Gil, R., Horvitz, E., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., & Bennett, P. N. (2019). Guidelines for Human-AI Interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, 1–13.

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, *20*(3), 973–989.

Anderson, E. (2020). Feminist Epistemology and Philosophy of Science. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2020). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2020/entries/feminism-epistemology/

Andres, J., Wolf, C. T., Cabrero Barros, S., Oduor, E., Nair, R., Kjærum, A., Tharsgaard, A. B., & Madsen, B. S. (2020). Scenario-based XAI for Humanitarian Aid Forecasting. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–8.

Arendt, H. (1994). Understanding and Politics (The Difficulties of Understanding). In J. Kohn (Ed.), *Essays in Understanding: 1930-1954* (1st ed.). Harcourt.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *An International Journal on Information Fusion*, *58*, 82–115.

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M. F., & Eckersley, P. (2019). Explainable Machine Learning in Deployment. *arXiv [cs.LG]*. arXiv. https://arxiv.org/abs/1909.06342

Bridle J. (2017) / Autonomous Trap 001. Retrieved March 13, 2020, from http://jamesbridle.com/works/autonomous-trap-001

Brunton, F., Nissenbaum, H., Echizen, I., Houmansadr, A., Cote, N., Hammond, R., & Friedler, S. (2017). *Obfuscation Workshop Report*. https://par.nsf.gov/servlets/purl/10046384

Coulton, P., & Lindley, J. G. (2019). More-Than Human Centred Design: Considering Other Things. *The Design Journal*, *22*(4), 463–481.

Crawford & Joler (2018) *Anatomy of an AI System*. Retrieved September 4, 2020, from https://anatomyof.ai/

De Mul, J. (2010). Moral Machines: ICTs as Mediators of Human Agencies. *Techné: Research in Philosophy and Technology*, *14*(3), 226–236.

Desjardins, A., Psarra, A., & A. Whiting, B. (2021). Voices and Voids: Subverting Voice Assistant Systems through Performative Experiments. In *Creativity and Cognition* (pp. 1–10). Association for Computing Machinery.

Di Salvo, C. (2012). *Adversarial Design*. The MIT Press.

DiSalvo, L. (2012). Nonanthropocentrism and the nonhuman in design: possibilities for designing new forms of engagement with and through technology. In Marcus Foth, Laura Forlano, Christine Satchell, and Martin Gibbs (Ed.), *From social butterfly to engaged citizen : urban informatics, social media, ubiquitous computing, and mobile technology to support citizen engagement* (pp. 421–435). MIT Press.

Druga, S., Williams, R., Breazeal, C., & Resnick, M. (2017). Hey Google is it OK if I eat you?: Initial Explorations in Child-Agent Interaction. *Proceedings of the 2017 Conference on Interaction Design and Children*, 595–600.

Ehsan, U., & Riedl, M.O. (2020). Human-centered Explainable AI: Towards a Reflective Sociotechnical Approach. HCI.

Ehsan, U., Liao, Q.V., Muller, M.J., Riedl, M.O., & Weisz, J.D. (2021). Expanding Explainability: Towards Social Transparency in AI systems. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.

Forlano, L. (2017). Posthumanism and Design. *She Ji: The Journal of Design, Economics, and Innovation*, *3*(1), 16–29.

Frauenberger, C. (2019). Entanglement HCI The Next Wave? *ACM Trans. Comput.-Hum. Interact.*, *27*(1), 1–27.

Ghajargar, M., Bardzell, J., Renner, A. S., Krogh, P. G., Höök, K., Cuartielles, D., Boer, L., & Wiberg, M. (2021). From ”Explainable AI” to ”Graspable AI”. *Proceedings of the Fifteenth International Conference on Tangible, Embedded, and Embodied Interaction*, 1–4.

Giaccardi, E. (2020). Casting things as partners in design: Towards a more-than-human design practice. In H. Wiltse (Ed.), *Relating to Things: Design, Technology and the Artificial*. Bloomsbury.

Giaccardi, E., Kuijer, L., & Neven, L. (2016, June 27). Design for Resourceful Ageing: Intervening in the Ethics of Gerontechnology. *DRS 2016 Design Research Society 50th Anniversary Conference*. https://doi.org/10.21606/drs.2016.258

Giaccardi, E., & Redström, J. (n.d.). Technology and more-than-human design. *Design Issues*, *36*(4).

Haraway, D. (1988). Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies: FS*, *14*(3), 575–599.

Joler & Pasquinelli (2020) *The Nooscope Manifested*. Retrieved September 4, 2020, from https://nooscope.ai/

Kuijer, L., & Giaccardi, E. (2018). Co-performance: Conceptualizing the Role of Artificial Agency in the Design of Everyday Life. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 125:1–125:13.

Karmann & Knudsen *Project Alias*. (2018). Retrieved March 13, 2020, from http://bjoernkarmann.dk/project_alias

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. (2021). What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, *296*, 103473.

McCarthy. L (2017). *LAUREN* —Retrieved December 1, 2021, from https://lauren-mccarthy.com/LAUREN

Lindley, J., Akmal, H. A., Pilling, F., & Coulton, P. (2020). Researching AI Legibility through Design. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13.

Luger, E., & Sellen, A. (2016). “Like Having a Really Bad PA”: The Gulf between User Expectation and Experience of Conversational Agents. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 5286–5297.

Lyons, H., Velloso, E., & Miller, T. (2021). Conceptualising Contestability: Perspectives on Contesting Algorithmic Decisions. In *arXiv [cs.CY]*. arXiv. http://arxiv.org/abs/2103.01774

Marenko, B., & Benqué, D. (n.d.). *Speculative Diagrams: experiments in mapping Youtube*. Retrieved December 1, 2021, from https://davidbenque.com/projects/architectures-of-choice/Marenko_David%20Benqu%C3%A9_2019_Speculative%20diagrams.pdf

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38.

Mittelstadt, B., Russell, C., & Wachter, S. (2018). Explaining Explanations in AI. In *arXiv [cs.AI]*. arXiv. https://doi.org/10.1145/3287560.3287574

Nicenboim, I., Giaccardi, E., Søndergaard, M. L. J., Reddy, A. V., Strengers, Y., Pierce, J., & Redström, J. (2020). More-Than-Human Design and AI: In Conversation with Agents. *Companion Publication of the 2020 ACM Designing Interactive Systems Conference*, 397–400.

Nicenboim, I. (2017). Objects that Withdraw – Retrieved March 13, 2020, from https://iohanna.com/Objects-that-Withdraw

Páez, A. (2019). The Pragmatic Turn in Explainable Artificial Intelligence (XAI). *Minds & Machines*, *29* , 441–459.

Pelikan, H. R. M., & Broth, M. (2016). Why That Nao? How Humans Adapt to a Conventional Humanoid Robot in Taking Turns-at-Talk. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 4921–4932.

Porcheron, M., Fischer, J. E., Reeves, S., & Sharples, S. (2018). Voice Interfaces in Everyday Life. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 640:1–640:12.

Reddy, A., Nicenboim, I., Pierce, J., & Giaccardi, E. (2020). Encountering ethics through design: a workshop with nonhuman participants. *AI & Society*. https://doi.org/10.1007/s00146-020-01088-7

Redström, J., & Wiltse, H. (2018). *Changing Things: The Future of Objects in a Digital World*. Bloomsbury Visual Arts.

Sciuto, A., Saini, A., Forlizzi, J., & Hong, J. I. (2018). “Hey Alexa, What’s Up?”: A Mixed-Methods Studies of In-Home Conversational Agent Usage. *Proceedings of the 2018 Designing Interactive Systems Conference*, 857–868.

Søndergaard, M. L. J., & Hansen, L. K. (2018). Intimate Futures: Staying with the Trouble of Digital Personal Assistants through Design Fiction. *DIS ’18 Proceedings of the 2018 Designing Interactive Systems Conference*, 869–880.

Strengers, Y., & Kennedy, J. (2020). The Smart Wife: Why Siri, Alexa, and Other Smart Home Devices Need a Feminist Reboot. The MIT Press.

Suchman, L. (2006). Human-Machine Reconfigurations: Plans and Situated Actions. Cambridge University Press.

Suchman, L. A. (1987). Plans and Situated Actions: The Problem of Human-Machine Communication. Cambridge University Press.

Tropozone. (2021). *The Sound of Speech As It Echoes in the Cloud*. Retrieved December 1, 2021, from http://tropozone.com/

Turek, M. (n.d.). *Explainable Artificial Intelligence*. Defense Advanced Research Projects Agency, Program Information. Retrieved June 7, 2020, from https://www.darpa.mil/program/explainable-artificial-intelligence

Vaccaro, K., Karahalios, K., Mulligan, D. K., Kluttz, D., & Hirsch, T. (2019). Contestability in Algorithmic Systems. *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*, 523–527.

Vera Liao, Q., Gruen, D., & Miller, S. (2020). Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *arXiv [cs.HC]*. arXiv. https://doi.org/10.1145/3313831.3376590

Wakkary, R. (2021). Things We Could Design: For More Than Human-Centered Worlds (Design Thinking, Design Theory). The MIT Press.

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing Theory-Driven User-Centric Explainable AI. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–15.

Weitz, K., Schiller, D., Schlagowski, R., Huber, T., & André, E. (2019). “Do you trust me?”: Increasing User-Trust by Integrating Virtual Agents in Explainable AI Interaction Design. *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, 7–9.

Whittaker, Crawford, Dobbe, Fried, Kaziunas, Mathur, West, M., Richardson, Schultz, & Schwartz. (2018). *AI Now Report*. AI Now. https://ainowinstitute.org/AI_Now_2018_Report.pdf

About the Authors:

Iohanna Nicenboim is a Microsoft Research PhD fellow at TU Delft, investigating human-AI interactions through more-than-human design. For the past ten years, she has been practicing speculative design to create fictions that highligh the ethics of living with autonomous technologies in everyday life.

Elisa Giaccardi is Professor in Post-Industrial Design at TU Delft. Her work on matters of digital transformation has contributed to the development of post-industrial and post-humanist approaches in design and HCI. She is the Scientific Coordinator of DCODE (https://dcode-network.eu).

Johan Redström is Professor in Design at Umeå Institute of Design, Sweden. He is part of DCODE (https://dcode-network.eu), a member of the Committee for Artistic Research at the Swedish Research Council, and the International Advisory board of the Design Research Society.