FROM MENTAL MODELS TO ALGORITHMIC IMAGINARIES TO CO-CONSTRUCTIVE MENTAL MODELS

CHRISTIAN SCHULZ

As technology in the fields of machine learning and artificial neural networks has advanced in recent years, triggering widespread public debate on the regulation and transparency of so-called artificial intelligence (as demonstrated by the current debates around generative AIs such as ChatGPT or Dall-E), there have been increasing demands for explainable AI.

The term "explainable artificial intelligence" (XAI) was first mentioned in 2004 in a paper by Michael van Lent, William Fisher, and Michael Mancuso, to highlight the ability of a system to explain the behavior of AI-controlled entities in simulation games (van Lent et al. 2004), but it is only since 2016 that we can speak of a renewed boom and a systematically diversifying research field of XAI, after the term "explainability" was mentioned since the 1980s in the context of explaining expert systems (e.g., Moore and Swartout 1988; Swartout and Moore 1993). However, as Tim Miller pointed out in a seminal paper in 2019, almost all work in the field of XAI is aimed solely at the perspective of researchers and developers and their intuitions of what constitutes a good explanation (Miller 2019).

The Transregional Collaborative Research Center 318, "Constructing Explainability," established in 2021 at the Universities of Paderborn and Bielefeld and funded by the German Research Foundation (DFG), has therefore set itself the task of expanding this perspective beyond the disciplines of computer science, cognitive science, and human-computer interaction, which are the obvious fields for AI development. This research network will work on the explainability of AI systems that primarily function co-constructively (Rohlfing et al. 2021). It will explicitly include not only the perspective of experts, but also that of different groups of everyday users who are not experts and differ substantially in their prior technical knowledge, gender, education, and socio-economic status (Finke et al. 2022). In the course of this program, it is important to locate a co-constructive explainability of AI systems on the concrete level of implementation.

However, this also requires a revision or even reconceptualization of certain theoretical concepts that play a central role in AI development, such as "mental models." Such mental models represent a form of user modeling and, to put it simply, are supposed to indicate the user's understanding of a technology. In computer science and human-computer interaction, however, they are almost always based on two quite problematic fundamental texts. Accordingly, this paper is divided into two parts.

Starting with a very brief sketch of the genealogy of the concept of mental models in the AI context, the first part will show that the problem of asymmetry,

in which the developer or researcher perspective ultimately always triumphs over that of the (everyday) user, has been inscribed in AI development from the very beginning. This is problematic not only because it provides a potential gateway for all forms of discrimination (Chun 2021), but also because quite pragmatically, with such an emphasis on the developer position at the conceptual level, no co-constructive explainability of AI systems can be implemented. Based on this finding, the second part of the article will then briefly outline how the concept of "algorithmic imaginaries" (Bucher 2018, 113-116) could make it possible to develop a symmetrical approach to such mental models. This would ultimately represent a decisive factor for the implementation of co-constructive explainability in AI systems. Admittedly, this cannot be done in the present paper, but it provides a useful starting point and framework for further research, which seems necessary for a reconceptualization of mental models in the context of AI development and, in particular, for the explainability of AI systems.

1. A BRIEF HISTORY OF MENTAL MODELS IN THE CONTEXT OF AI

Mental models are an essential component of explanations. Although the antecedents of concepts of mental models can be traced back to the nineteenth century and can be found, for example, in the work of Charles Sanders Peirce (Johnson-Laird 2004), the first work to use the term explicitly was Kenneth Craik's book *The Nature of Explanation* (1943). In the key fifth chapter of his book, Craik describes the anticipation and prediction of events as central properties of human thought. However, in his three-level model with the levels "translation, reasoning, and retranslation" (Craik 1943, 50), which in a way represent a kind of cybernetic feedback loop (Ashby 1956, Wiener 1948), he also assumes a dichotomy between inside and outside or subject and object, whereby the mental model can imitate reality or at least establish a similarity between world and image. Craik postulates a "similar relation-structure" between the image and the process it imitates (Craik 1943, 51), and thus ultimately a symbolism that is constitutive of human thought (ibid., 58). In this respect, it is hardly surprising that in his conceptualization of mental models there is not necessarily a direct connection between this inside and outside, or subject and object, but rather a focus on symbolic meaning.

This is quite in contrast to more recent variants and appropriations of the notion of mental models, such as Donald Norman's conceptualization, which has been influential for computer science and human-computer interaction. Relevant publications and papers in these two disciplines almost always refer to Norman's text, first published in 1983. Norman's conceptualization of a mental model distinguishes between a "conceptual model" on the part of the developers, who develop a certain system ("target system") for a certain purpose, and the mental models that users construct to explain this developed system – often involving different modes of use than those envisaged by the developers (Norman 1983, 7). These mental models of users are influenced by their previous technical

knowledge, their experiences with similar systems, and their perceptual ability. Norman therefore also speaks of a second-order conceptualization, thus the model of a model (Norman 1983, 8). Implicitly, this already addresses a co-constructive perspective in addition to the aspect of processuality (Jones et al. 2011). However, such a conceptualization clearly emphasizes the developer's perspective, since both models are constructed on the developer's side. This also subjects the mental model to an asymmetry that obstructs the explainability of AI systems as a social practice. Thus, there is no real co-construction here, since ultimately the developers decide on the models of the users.

Nevertheless, it is interesting to note that these two foundational texts, belonging to cognitive science, are in some ways also symptomatic of the two paradigms in the history of AI development. Craik's conceptualization of mental models can still be clearly associated with the paradigm of symbolic AI that prevailed until the 1980s (Haugeland 1985). This paradigm is based on logical-mathematical methods and semantics and states that both the mind and the computer are physical symbol systems. The hypothesis here is that the human brain and the digital computer share a common functional description at some level of abstraction (Mitchell 2019, 9-12). In contrast, starting in 1987 and based on the influential work of John McClelland, David Rumelhart, and Geoffrey Hinton, there was a paradigm shift toward a connectionist approach, also called subsymbolic AI (McClelland et al. 1987). This approach states that AI should be created by modeling the brain rather than by symbolically representing the world through the mind, and the roots here are not so much in philosophy as in neuroscience. It was not the conscious rational action of humans but the dynamic coupling of the organism or machine with the world that advanced here to model intelligent behavior (Mitchell 2019, 12-17).

Norman's conceptualization of mental models preceded this paradigm shift toward subsymbolic AI by only a short time. While highly influential, his concept inevitably lags behind the paradigm shift toward connectionist approaches that became increasingly widespread in AI development from the late 1980s onward, since he ultimately remains caught in a symbolic paradigm.[1] Subsequently, approaches and efforts towards a "human-centered design" (Norman 2004) or, more recently, a "more than human-centered design" (Wakkary 2021) influenced by posthumanist theory do not change anything, since they remain bound to a concept of mental models that can be traced back to symbolism. In this respect, and to summarize very briefly, either Craik's antiquated mental models concept

1 It is also interesting in this context that Norman has a paper in the influential book by McClelland and Rumelhart, in which he discusses cognition and parallel distributed processing and at one point also indirectly mentions mental models. He writes: "Don't I need to have mental variables, symbols that I manipulate? My answer is 'yes'. I think this lack is a major deficiency in the PDP approach" (Norman 1987, 541). This shows that Norman remains more or less attached to a symbolic paradigm, which clearly reveals the desideratum of user modeling via co-constructive mental models with regard to a subsymbolic AI.

or Norman's asymmetric model is still referenced here, but both remain tied to the symbolic AI paradigm, which must remain unsatisfactory in terms of a co-constructive AI development. A preliminary sampling of the different lines of reception of Craik's and Norman's concepts of mental models in computer science and HCI seems to confirm the urgent need for a reformulation of the concept. On the one hand, Craik's and Norman's positions are taken up by approaches from, for example, organizational learning or educational research, which perpetuate the problem of an inside/outside dichotomy described above (e.g. Rook 2013), or reproduce the emphasis on the position of developers (e.g. Greca and Moreira 2000).

In contrast, Volkamer and Renaud, in the field of computer science, have pointed out the asymmetry arising from this emphasis on the developer perspective (Volkamer and Renaud 2013, 256). It is therefore understandable that the disciplines of psycholinguistics and cognitive science, which are also heavily involved in AI development, are increasingly focusing on the dialogic interaction situation (e.g., Brennan and Hanna 2009, Brennan et al. 2010, Brown-Schmidt 2012).

This is particularly evident in the concept of "partner models," which, while building on earlier research on mental models, co-constructively foregrounds the "communicative capacities" (Doyle et al. 2021, 5) of human and non-human interlocutors. This "partner models" approach is important for a micro perspective and focuses on the communicative abilities and mutual understanding of the actors involved in the situation. However, it is not able to say anything about social aspects and contexts, whose relevance is not confined to concrete situations of human-human dialogues (HHD) or human-machine dialogues (HMD). For example, the social situatedness of people (e.g., race, class, and gender), as well as affective-emotional states associated with interface designs (Drucker 2014), play an essential role that must always be taken into account in explanations as social practices. However, the social situatedness of people is only very vaguely addressed in the conceptualization of partner models, with the notion of "global partner models" and the "broad stereotypes" behind them (Doyle et al. 2021, 2). The aspect of the social beyond the dialogic situation is not really considered. In a sense, then, the problem that emerges here corresponds to the relationship between micro and macro perspectives in the social sciences, but does not resolve the asymmetry of the influential mental models concepts.

In the meantime, Norman also seems to have become aware of this problem, at least to some extent, because in his most recent book he explicitly emphasizes that a "human-centered design" approach (and thus also his conceptualization of mental models) is no longer sufficient and instead argues for a "humanity-centered design." He writes:

> The phrase "human centered" fails to emphasize the larger concerns and the need for increased sensitivity to biases and prejudices against certain societal groups. The phrase "humanity centered" emphasizes

> designs that take into account the sociotechnical system in which people reside. (Norman 2023, 181f)

Nevertheless, he still seems to regard the principles of "human-centered design" (with which he explicitly associates his definition of mental models, Norman 2004, 75) as central for "humanity-centered design": he writes that of course the principles of "human-centered design" must be adhered to, but within a broader scope (Norman 2023, 182). It is therefore not surprising that in his sporadic explanations of what "humanity-centered design" means for AI development, he only briefly touches on the concept of "human-centered AI" developed by Ben Shneiderman (Norman 2023, 269). Yet Shneiderman's "human-centered AI" actually still means designer-centered AI, thus reproducing exactly those asymmetries that were highlighted above with regard to the history of the concept of mental models (Shneiderman 2022, 79–81). Norman nevertheless seems to feel a certain discomfort, for which, however, he cannot offer a concrete solution, but exhausts himself in vague proclamations. The paragraph on AI in his book concludes with the following sentence: "My personal bias is that we must combine the good parts of old-fashioned symbolic reasoning with neural networks to allow the power of each to overcome the other's deficiencies" (Norman 2023, 269).

It becomes clear, then, that no solution to co-constructive user modeling is to be expected in either Craik's or Norman's models. The following section will therefore turn to research on algorithmic imaginaries in media and cultural studies. This concept of algorithmic imaginaries represents, in a way, the user-centered equivalent to the mental models of computer science and HCI. With a theoretical reorientation, it can be used to develop a new, symmetrical approach to mental models, as the second part of the text will now show.

2. THE IMPORTANCE OF ALGORITHMIC IMAGINARIES FOR A SYMMETRICAL CONCEPT OF MENTAL MODELS

The concept of the "algorithmic imaginary" (Bucher 2018) has received attention in discourses around media studies and in particular social media research, as it brings into focus for the first time users' appropriations of algorithmic processes operating in opacity and their imaginaries of these operations. Bucher describes this concept as

> [. . .] ways of thinking about what algorithms are, what they should be, how they function, and what these imaginations, in turn, make possible. While there is no way of knowing for sure how algorithms work, the personal algorithm stories illuminate how knowing algorithms might involve other kinds of registers than code. […] In other words, the algorithmic imaginary emerges in the public's beliefs, experiences, and expectations of what an algorithm is and should be. (Bucher 2018, 113f)

Thus, this approach is primarily concerned with the users' perspective. This is important for a comprehensive theoretical understanding of AI systems, because without a precise description of users' practices and imaginaries of how different AI systems work, one would inevitably fall into some form of reductionism and end up with an asymmetrical view. This would once again highlight the developers' perspective, like the concepts of mental models discussed above. Bucher's concept, however, cannot offer a symmetric conceptualization of mental models on a micro level because it omits the processes on the side of the developers. It is therefore important that concepts of the algorithmic imaginary should include the algorithmic processes, which are usually opaque from the user's perspective and are located in the so-called "backend," i.e. the invisible part of the interface. This has been demonstrated elsewhere for the context of social media platforms (Schulz 2023). Here, too, the algorithm "imagines" the future behavior of the users via so-called "predictor modules" of machine learning, which are supposed to predict the future behavior of the users from all their actions. Although it must be noted that the algorithm (or more generally an AI system) does not "imagine" in the human sense, and the metaphorical part must always be taken into account, it is worth extending the concept of the imaginary to non-human entities. The algorithm not only computes according to predefined parameters, but also constantly plays through transforming and supposedly fitting models in the backend depending on the user's behavior. Fisher and Mehozay (2019) suggest that, on the one hand, the algorithm observes the behavior of users and derives imaginaries from this, but that, on the other hand, the designers and developers also rely on "imaginary interlocutors" (Fisher and Mehozay 2019, 1179). This approach must be extended to include the user perspective, just as, conversely, the perspective of the developers and algorithms must be integrated into Bucher's concept of an algorithmic imaginary.

Indeed, the users' imaginaries are a concrete part of the infrastructure of AI, and precisely because of this, the algorithmic imaginaries also affect and change the behavior of these very users via the backend. This is a constant interplay (Schulz and Matzner 2020), which is coupled to the imaginary and converges in the interfaces of AI systems. This makes it necessary to design a theoretically more comprehensive algorithmic imaginary, which includes algorithms and the developer's perspective, as well as the perspectives of different users and, not least, the level of the interface. Such a reconceptualization of the algorithmic imaginary enables us to move from a more theoretical and cultural science perspective towards an interdisciplinary and co-constructive approach to mental models, which encompasses AI systems, developers, and users alike, and does not place one entity above another.

This is all the more important because recent scientific appropriations of an imaginary in the field of AI research are based almost exclusively on science and technology studies (Jasanoff and Kim 2009; Jasanoff and Kim 2015), where they reflect the so-called "practice turn" (Schüttpelz et al. 2021). Curiously, however,

they only deal with the big social imaginaries (e.g., about ethical concerns or the dangers of a general artificial intelligence), or reproduce on a micro level the asymmetric perspective of designers. Arguing along these lines, Lucy Suchman, following Keith Grint and Steve Woolgar, has already pointed out that "design imaginaries" (Suchman 2007, 187-205) on the part of developers play a central role in "user configuration" (Grint and Woolgar 1997, 92) in the course of technology developments.

However, the role of everyday users is not included here, not least because of the ethnographic focus on developers and expert users. For this very reason, imaginaries are necessary – both on the part of the developers, about possible ways of using the technology by the users (especially with regard to the implementation of new functions), and on the algorithmic level (with regard to the explainability of algorithmic decisions). Of course, a newly introduced (algorithmic) technology or function and a corresponding explanation by no means always leads to the usage behavior anticipated by the developers, although certain incentives can be set up in the interface design to influence behavior (e.g., pop-up windows that display explanations when needed). Conversely, however, it is also not the case that users determine the behavior of AI systems, although they can try to influence them by giving good or bad ratings for the explanations they are offered. Such practices stem from users' ideas about how the AI works and correspond quite closely to what Bucher meant by her suggestion of an "algorithmic imaginary." Nonetheless, it is essential to consider the other side of the "design imaginaries" and algorithms for a comprehensive understanding of how AI systems work, without playing one side off against the other.

In this respect, Bucher's concept of the algorithmic imaginary, which places users at the center, provides a useful counterpoint to the asymmetric concepts of mental models presented in the first part of this paper, with their emphasis on the position of developers. Beyond that, however, an algorithmic imaginary that includes users as well as developers and algorithms (Schulz 2023), and also the level of the interface, potentially offers the possibility of a fundamentally new and symmetrical approach to user modeling in AI development via mental models. And this is precisely where interdisciplinary AI development can concretely benefit from media studies research.

3. OUTLOOK

This leads me to the question what such an endeavor might look like and how it might succeed. First of all, it is crucial to compile a compendium of such models, based on an exact review and reading of the different lines of reception of mental models in the disciplines of cognitive science, computer science and HCI. The primary goal is to gain a detailed overview of the heterogeneous ramifications and transformations of mental models.

At the same time, the concept of a multi-perspectival algorithmic imaginary, which encompasses both the developer and the user perspective and also the AI systems themselves and their interfaces in their respective sociotechnical situatedness (Haraway 1988), makes it necessary to explicitly understand this negotiation process as a social practice, without privileging one perspective over the other. This also makes it necessary not to lose sight of the differences between the two theoretical concepts of mental models and imaginaries. While imaginaries, in our conceptualization, address both the imaginaries of users in relation to AI technologies and the imaginaries of developers about certain behaviors of users in dealing with them, the notion of mental models is traditionally narrower and targets already formalized and specified application contexts. This is to a significant extent related to the appropriation of the concept by the disciplines of computer science and HCI.

In this respect, mental models also have a latent tendency to address more expert knowledge. Or at least, this is still the case at present, but this is exactly what a reconceptualization and the development of a symmetric concept of mental models aims to change. In addition, however, it is also important to look at current developments in the relevant fields or disciplines. For example, in the field of HCI, there are interesting developments in the area of participatory design (Bødker and Kyng 2018, Bødker et al. 2021), where we find very similar calls for greater inclusion of users right from the development level. The approach of a co-constructive XAI makes similar demands. In any case, "Designing is entangling – the simple act of encouraging interdependence" (Easterling 2021, 13). There could be worse perspectives for media studies research than to participate in the concrete development of better AI.

REFERENCES

Ashby, William Ross. 1956. *An Introduction to Cybernetics*. London: Chapman & Hall.

Bødker, Susanne, and Morten Kyng. 2018. "Participatory Design that Matters – Facing the Big Issues." *ACM Transactions on Computer-Human Interaction* 25 (1) (February): 1–13.

Bødker, Susanne, Christian Dindler, Ole Sejer Iversen, and Rachel Charlotte Smith. 2021. "Participatory Design." *Synthesis Lectures on Human-Centered Informatics* 14 (5): i–143.

Brennan, Susan E., and Joy E. Hanna. 2009. "Partner-Specific Adaptation in Dialog." *Topics in Cognitive Science* 1 (2): 274–291.

Brennan, Susan E., Alexia Galati, and Anna Katharina Kuhlen. 2010. "Two Minds, One Dialog: Coordinating Speaking and Understanding." In *The Psychology of Learning and Motivation: Advances in Research and Theory*, edited by Brian H. Ross, 301– 344. Burlington, MA: Academic Press.

Brown-Schmidt, Sarah. 2012. "Beyond Common and Privileged: Gradient Representations of Common Ground in Real-Time Language Use." *Language and Cognitive Processes* 27: 62–89.

Bucher, Taina. 2018. *If... Then: Algorithmic Power and Politics.* Oxford: University Press.

Chun, Wendy Hui Kyong. 2021. *Discriminating Data: Correlation, Neighborhoods, and the New Politics of Recognition*. Cambridge, MA: MIT Press.

Craik, Kenneth. 1943. *The Nature of Explanation*. Cambridge: Cambridge University Press.

Doyle, Philip R., Leigh Clark, and Benjamin R. Cowan. 2021. "What Do We See in Them? Identifying Dimensions of Partner Models for Speech Interfaces Using a Psycholexical Approach." In *CHI 2021: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, May 2021. 1–14.

Drucker, Johanna. 2014. *Graphesis: Visual Forms of Knowledge Production*. Cambridge: Harvard University Press.

Easterling, Keller. 2021. *Medium Design: Knowing How to Build the World*. London, New York: Verso.

Finke, Josefine, Ilona Horwath, Tobias Matzner, and Christian Schulz. 2022. "(De)Coding Social Practice in the Field of XAI: Towards a Co-Constructive Framework of Explanations and Understanding Between Lay Users and Algorithmic Systems." In *Artificial Intelligence in HCI*, edited by Helmut Degen and Stavroula Ntoa, Lecture Notes in Computer Science. Cham: Springer.

Fisher, Eran, and Yoav Mehozay. 2019. "How Algorithms See Their Audience: Media Epistemes and the Changing Conception of the Individual." *Media, Culture & Society* 41 (8): 1176–1191.

Greca, Ileana Maria, and Marco Antonio Moreira. 2000. "Mental Models, Conceptual Models, and Modelling." *International Journal of Science Education* 22 (1): 1–11.

Grint, Keith, and Steve Woolgar. 1997. *The Machine at Work: Technology, Work and Organization*. Cambridge: Polity Press.

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

Haugeland, John. 1985. *Artificial Intelligence: The Very Idea*. Cambridge, London: MIT Press.

Jasanoff, Sheila, and Sang-Hyun Kim, eds. 2015. *Dreamscapes of Modernity: Sociotechnical Imaginaries and the Fabrication of Power.* Chicago: University of Chicago Press.

Jasanoff, Sheila, and Sang-Hyun Kim. 2009. "Containing the Atom: Sociotechnical Imaginaries and Nuclear Power in the United States and South Korea." *Minerva* 47: 119–146.

Johnson-Laird, Philip N. 2004. "The History of Mental Models." In *Psychology of Reasoning: Theoretical and Historical Perspectives*, edited by Ken Manktelow and Man Cheung Chung, 179–212. London: Psychology Press.

Jones, Natalie A., Helen Ross, Timothy Lynam, Pascal Perez, and Anne Leitch. 2011. "Mental Models: An Interdisciplinary Synthesis of Theory and Methods." *Ecology & Society* 16 (1): 46.

McClelland, James L., David E. Rumelhart, and Geoffrey Hinton. 1987. "The Appeal of Parallel Distributed Processing." In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*, edited by David E. Rumelhart, James L. McClelland, and PDP Research Group, 3–40. Cambridge, MA: MIT Press.

Miller, Tim. 2019. "Explanation in Artificial Intelligence: Insights from the Social Sciences." *Artificial Intelligence* 267: 1–38.

Mitchell, Melanie. 2019. *Artificial Intelligence: A Guide for Thinking Humans*. London: Penguin Random House.

Moore, Johanna D., and William R. Swartout. 1988. *Explanation in Expert Systems: A Survey*. Research Report, Defense Technical Information Center, https://apps.dtic.mil/sti/citations/ADA206283.

Norman, Don. 2023. *Design for a Better World: Meaningful, Sustainable, Humanity Centered*. Cambridge: MIT Press.

Norman, Don. 2004. *Emotional Design: Why We Love (or Hate) Everyday Things*. New York: Basic Books.

Norman, Don. 1987. "Reflections on Cognition and Parallel Distributed Processing." In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*, edited by David E. Rumelhart, James L. McClelland, and PDP Research Group, 531–546. Cambridge, MA: MIT Press.

Norman, Don. 1983. "Some Observations on Mental Models." In *Mental Models*, edited by Dedre Gentner and Albert L. Stevens, 7–14. New York: Psychology Press.

Rohlfing, Katharina, Philipp Cimiano, Ingrid Scharlau, Tobias Matzner, Heike M. Buhl, Hendrik Buschmeier, Elena Esposito et al. 2021. "Explanation as a Social Practice: Toward a Conceptual Framework for the Social Design of AI Systems." *IEEE Transactions on Cognitive and Developmental Systems* 13 (3): 717–728.

Rook, Laura. 2013. "Mental Models: A Robust Definition." *International Journal of Knowledge and Organizational Learning Management* 20 (1): 38–47.

Schulz, Christian. 2023. "A New Algorithmic Imaginary." *Media, Culture & Society* 45 (3): 646–655.

Schulz, Christian, and Tobias Matzner. 2020. "Feed the Interface: Social-Media-Feeds als Schwellen." *Navigationen* 2: 147–162.

Schüttpelz, Erhard, Ulrike Bergermann, Monika Dommann, Jeremy Stolow, and Nadine Taha, eds. 2021. *Connect and Divide: The Practice Turn in Media Studies*. Zürich: Diaphanes.

Shneiderman, Ben. 2022. *Human-Centered AI*. Oxford: Oxford University Press.

Suchman, Lucy. 2007. *Human-Machine Reconfigurations: Plans and Situated Actions*. 2nd edition. Cambridge: Cambridge University Press.

Swartout, William R., and Johanna D. Moore. 1993. "Explanation in Second Generation Expert Systems." In *Second Generation Expert Systems*, edited by Jean-Marc David, Jean-Paul Krivine, and Reid Simmons. Berlin: Springer.

van Lent, Michael, William Fisher, and Michael Mancuso. 2004. "An Explainable Artificial Intelligence System for Small-Unit Tactical Behavior." In *Proceedings of the Nineteenth National Conference on Artificial Intelligence, Sixteenth Conference on Innovative Applications of Artificial Intelligence*, July 25-29, 2004, San Jose, California, USA.

Volkamer, Melanie, and Karen Renaud. 2013. "Mental Models: General Introduction and Review of Their Application to Human-Centered Security." In *Number Theory and Cryptography: Papers in Honor of Johannes Buchmann on the Occasion of His 60th Birthday*, edited by Marc Fischlin and Stefan Katzenbeisser, 255–280. Karlsruhe: Springer.

Wakkary, Roy. 2021. *Things We Could Design: For More Than Human-Centered Worlds*. Cambridge, MA: MIT Press.

Wiener, Norbert. 1948. *Cybernetics or Control and Communication in the Animal and the Machine*. Cambridge, MA: MIT Press.