Sociologica. V.16 N.3 (2022)
ISSN 1971-8853

Does Explainability Require Transparency?

Elena Esposito*

Department of Political and Social Sciences, University of Bologna (Italy)
Faculty of Sociology, University of Bielefeld (Germany)

Submitted: November 15, 2022 – Revised version: February 6, 2023
Accepted: February 6, 2023 – Published: March 15, 2023

Abstract

Dealing with opaque algorithms, the frequent overlap between transparency and explainability produces seemingly unsolvable dilemmas, as the much-discussed trade-off between model performance and model transparency. Referring to Niklas Luhmann's notion of communication, the paper argues that explainability does not necessarily require transparency and proposes an alternative approach. Explanations as communicative processes do not imply any disclosure of thoughts or neural processes, but only reformulations that provide the partners with additional elements and enable them to understand (from their perspective) what has been done and why. Recent computational approaches aiming at *post-hoc explainability* reproduce what happens in communication, producing explanations of the working of algorithms that can be different from the processes of the algorithms.

Keywords: Explainable AI; transparency; explanation; communication; sociological systems theory.

Acknowledgements

This work was supported by the European Research Council (ERC) under Advanced Research Project PREDICT no. 833749, and by the German Research Foundation (DFG) within the framework of The Collaborative Research Centre *Constructing Explainability*. My thanks to the participants in the conference *Explaining Machines* at Bielefeld University on June 23-24, 2022, where this paper was presented. For their helpful comments, criticisms, and suggestions I am grateful to Fabian Beer, Dominik Hofmann, David Stark and the anonymous reviewers of *Sociologica*.

* elena.esposito9@unibo.it

1 The Barrier of Explainability

The more and more pervasive, powerful and effective role of digital technologies in our social life is making a non-technological topic increasingly central: the issue of explanation. The latest algorithms that employ advanced machine learning techniques and use big data produce results that are increasingly difficult for their users to understand, creating serious problems of trustworthiness, transferability and control — as well as concerns about the fairness and appropriateness of the use of their results. One would then want to be able to explain their processes, and this is what the recent and very active branch of Explainable Artificial Intelligence (XAI) is about. The research is mainly carried out by programmers and computer scientists, but the underlying problem is not only technological. It is primarily a matter of understanding what explanations are and how they work: what is meant by explanation, for whom, and of what[1]?

These questions are primarily addressed by philosophy, psychology, and sociology, which have been dealing with them for a long time (e.g., Miller, 2019). Philosophical research has investigated the logical form and the assumptions of explanations (Hempel, 1966; von Wright, 1971; Pearl & Mackenzie, 2018), while psychology has analyzed the cognitive processes involved in explanations (Heider, 1958; Malle, 1999). When it comes to the practical use and the impact of algorithms in our society, however, the key contribution should come from sociology, whose perspective is singularly absent from the XAI debate.[2] This paper aims to contribute to filling this gap by addressing one of the key issues concerning the obscurity of algorithms: the difference between transparency and explainability. As elaborated in more detail in the following pages, the frequent overlap between transparency and explainability produces seemingly unsolvable dilemmas, which the sociological perspective can help to clarify and manage more productively.

The advantage lies first of all in the possibility of turning a seemingly insoluble problem into an opportunity — for programming techniques, but also for the social use of algorithms. Reading the literature on XAI, one has the impression that it faces an insurmountable difficulty: the "barrier of explainability" (Arrieta et al., 2020, p. 82) of the latest self-learning algorithms, primarily Deep Neural Networks. They have become extremely effective and powerful, but also so sophisticated, complex and autonomous that almost no human intervention is required for their design and deployment. These systems depend on the use of distributed representations (LeCun et al., 2015) on many distinct hidden layers, which work independently of each other. The role of any particular feature of the input variables can be very different at any level, and it becomes impossible to identify general cause and effect relationships. It is therefore advisable to abandon the idea of trying to understand how a deep neural network works as a whole by understanding what the individual layers do (Frosst & Hinton, 2017). Machines work by themselves, and their output is often impossible for human observers to reconstruct (Goodfellow et al., 2016; Burrel, 2016; Gilpin, 2018). Such *black-box models* are intrinsically impenetrable, or *opaque*, to their users, and the very programmers who designed them can be unable to understand how the algorithms get at their results. According to this approach, human observers can at most propose remedies or suggestions to deal with systems that remain basically unexplained.

The "barrier" to observation has considerable practical consequences, since these opaque techniques are the ones underlying the recent "spring" of AI by enabling machines to achieve amazing performances and give the impression that they have finally become intelligent (Es-

1. A difficulty often complained of is that there is no agreement among scholars and across disciplines on what an explanation precisely is (e.g., Vilone & Longo, 2021; Coeckelbergh, 2020).

2. Besides occasional references to Grice's (1975) conversational maxims.

posito, 2022, Ch. 1). As David Weinberger (2018) argues, requiring an explanation from such incomprehensible intelligent agents could amounts to "forcing the AI to be artificially stupid enough for us to understand how it arrives at its conclusions." What then can be the task of XAI? Does the barrier of explainability of intelligent automated systems imply then that XAI should aim at artificial stupidity? This is certainly not an encouraging perspective. The dilemma is often presented as a *trade-off between model performance and model transparency* (Arrieta et al., 2020, p. 83; Doshi-Velez & Kim, 2017; Montavon et al., 2018; Rudin, 2019; Busuioc, 2020): if you want to take full advantage of the intelligence of the algorithms, you have to accept their unexplainability — or find a compromise. If you focus on performance, opacity will increase — if you want some level of explainability you can maybe better control negative consequences, but you will have to give up some intelligence.[3]

This is where the sociological perspective can make a contribution and change the terms of the question, showing that this somewhat depressing approach to XAI is not the only possible one. Explanations can be observed as a specific form of communication, and their conditions of success can be investigated. This properly sociological point of view leads to question an assumption that is often taken for granted, the *overlap between transparency and explainability*: the idea that if there is no transparency (that is, if the system is opaque), it cannot be explained — and if an explanation is produced, the system becomes transparent. From the point of view of a sociological theory of communication, the relationship between the concepts of transparency and explainability can be seen in a different way: explainability does not necessarily require transparency, and the approach to incomprehensible machines can change radically.

2 Explaining Black Boxes

To illustrate this change in perspective, it is first necessary to define the two notions of transparency and explainability in the context of an interaction between the user and the opaque machine. What can be done to manage the barrier, and who should do what?

Transparency or interpretability (the two terms are often used interchangeably: Lipton, 2018; Bibal et al., 2021) refers to an inherent "passive" (Arrieta et al., 2020, p. 84) characteristic of some systems, that do nothing and perform no additional operations. Such systems are inherently made in a way that is accessible to their users. If humans have enough information and skills, they can know what the machine does and why, and act accordingly. In computer science, people talk of models interpretable-by-design,[4] and sometimes invite to privilege trans-

3. If one looks at the tradition of computer science, however, this opposition between opacity and transparency is really curious. For a long time, in fact, transparency was pursued not to enable users to look inside systems (against opacity) but to make them invisible to their users. A computer program was said to be transparent if the user was unaware of it (e.g. https://www.techtarget.com/whatis/definition/transparent) — that is, if it worked so well that one was oblivious to its intervention. In sociological perspective as well, Luhmann (1997) spoke of computers as "invisible machines" (p. 117). According to Latour (1999) precisely the most successful technologies tend to become increasingly obscure: for which of us is understandable how our cars or washing machines work? Obscurity and transparency, in this sense, were not opposed: the more transparent machines are, the more they can remain obscure. Today, however, recent developments in algorithmic technologies make the issue of transparency more challenging. If the machine remains obscure to everyone (if it is opaque), being oblivious of its working might mean giving up control — and is obviously risky.

4. Via algorithmic transparency (when the users are able to follow the processes performed by the algorithm to produce the output from the input), via decomposability (when they understand input, parameters, and calculations separately), or via simulatability (when human beings are able to reproduce the operations).

parency in systems programming (Robbins, 2019; Rudin, 2019). The opposite are black boxes systems, which are not accessible to users (Buhrmester et al., 2019; Pasquale, 2015).

Explainability, on the other hand, does not refer to an intrinsic characteristic of the systems but to the "active" behavior of some of them that have procedures to provide users with explanations of their workings (Arrieta et al., 2020, p. 84). Explainable is a system that is made (or makes itself) understandable[5] — which does not imply that it is or becomes transparent. The goal of explanation is not to give full access to the operations, criteria and elements by which the machine operates, which can remain obscure. The goal is *understandability*, defined in a way that does not imply transparency (Ananny & Crawford, 2018): it "denotes the characteristic of a model to make a human understand its function — how the model works — without any need for explaining its internal structure or the algorithmic means by which the model processes data" (Arrieta et al., 2020, p. 84). In this meaning, a model is understandable to someone if that user can make sense of it, in her way and according to her capacities and interests. It is sufficient that the human observer understands what the algorithm does well enough to be able to elaborate, control and possibly contest its results.

Explainability in this sense does not exist "passively" per se — the system must "actively" do something to be understandable.[6] To this end, not only is transparency not necessary, but it can also be an obstacle. As Bucher (2018) argues, intelligibility does not necessitate opening the black box, because "too much information may blind us from seeing more clearly and, ultimately, from understanding" (p. 46). The impossibility of seeing inside the black box is not necessarily an "epistemological limit," but rather the stimulus to develop "more efficient methods of research" (Von Hilgers, 2010, p. 43). This would be the task of XAI, that according to this approach does not manage the unavoidable obscurity of algorithms but addresses an open research issue: how to develop techniques not to increase transparency but to increase control, because "we do not need to know everything in order to control" (Von Hilgers, 2010). The obscurity of algorithmic procedures, which is obviously very problematic for their acceptance by the public, can be seen from a computational point of view not as a problem but paradoxically as an advantage: by disengaging from the burden of comprehensibility, their working can be more accurate and efficient (Shmueli, 2010).

The strategy for explainability would then be different from the search for transparency, and in fact many projects on explainable AI are recently taking a different approach, compatible with the radical obscurity of algorithmic processes. Somewhat contradictory to their name, recent XAI projects are not concerned with machine intelligence. Rather, the goal is to produce a *communication* between the algorithm and the user, in which the machine allows the users to exercise a form of control by providing responses that take as input its partners' ever-changing requests for clarification (e.g., Cimiano et al., 2010; Rohlfing at al., 2021). The machine must be able to participate in a meta-communication (Bateson, 1972; Luhmann, 1997): a communication about communication, that may have as its object the processes of the machine, the data it processes or the parameters it uses. The goal is not, and cannot be, that these processes and the intelligence of the machine become transparent to the users, but that the users can make sense of what the machine conveys about its processes, data and parameters in such a way that a form of control can be applied. As Bibal et al. (2021) argue:

> it is not necessarily required to provide an interpretable representation of a mathematical model, but most importantly to provide a train of thought that can make

5. Gilpin et al. (2018) analyze "Explanation-Producing Systems".

6. O'Hara (2020) speaks of "explanation as a process".

> the decision meaningful for a user (i.e. so that the decision makes sense to him). (pp. 167–168)

Machines must be able to produce adequate explanations by responding to the requests of their interlocutors.[7] This is actually what happens in the communication with human beings as well. I refer here to Niklas Luhmann's (1995) notion of communication, which abandons the tradition that draws on Shannon and Weaver (1949) and subsequent elaborations (e.g., Eco, 1975). Whereas in this tradition communication requires the sharing of a thought or part of a thought among the participants, Luhmann argues that there is no unit of information that moves from one end of communication to the other one and can be shared — each person's thoughts remain only their own. Everyone constructs their information and thoughts in their own way, differently from each other, but in communication does so starting from the stimuli provided by the partners and their intention to communicate. Each of us, when we understand a communication, understand in our own way what the others are saying or communicating, and do not need access to their thoughts. In this view, the mutual intransparency of the partners is not a liability but the normal condition of communication, allowing each participant to develop their own specific thoughts from the thoughts of others, in a coordinated but autonomous way.

Social structures such as language, semantics, and communication forms normally provide for sufficient coordination (Luhmann, 1997, Ch. 2), but perplexities may arise, or additional information may be needed. In these cases, we may be asked to give explanations — the purpose of which is that the recipient understands our communication and is able to make sense of it. In Charles Tilly's words, it is a matter of answering the question "Why?" giving reasons for what people "have done, for what others have done to them, or more generally or what goes on in the world" (Tilly, 2006, p. IX). One reacts to a communication problem with another communication, which takes the form of an explanation and must be understandable and credible. One must give reasons, and reasons giving is a social activity, that varies from one social situation to another and provides the appropriate information for each of them (Tilly, 2006; Bibal et al., 2021). But what information do we get when we are given an explanation? We continue to know nothing about our partner's neurophysiological or psychic processes — which (fortunately) can remain obscure, or private. To give a good explanation we do not have to disclose our thoughts, even less the connections of our neurons (Tilly, 2006). We can talk about our thoughts, but our partners only know of them what we communicate, or what they can derive from it. We simply need to provide our partners with additional elements, which enable them to understand (from their perspective) what we have done and why.

Explanations offer "alternative descriptions" of the ongoing communicative situation (Wright Mills, 1971, p. 905) or, in Luhmann's terms (1990), "reformulations with the added benefit of better connectivity" (p. 410). The sender produces a new communication that provides additional elements related to the partner's specific request and needs. Reasons are given, which can take different forms and whose appropriateness depends on the social setting: Tilly (2006) lists conventions, stories, codes, and technical accounts. The process is entirely communicative: we do not need transparency as access to the brains or the minds of our interlocutors — we only need to get clues that allow a specific communication to move forward in a controlled, non-arbitrary way. An explanation is successful if the partner understands enough to respond, object, elaborate.

7. Not necessarily in verbal form. As we will see later, communication can make use of written or oral texts, visualizations, or other means.

3 Programming Explanations

One difficulty with the common approach to explanation is that it frequently reproduces the *double standard* often adopted in the interaction with machines vis-à-vis the interaction with human partners: algorithms are required to illuminate the cognitive processes that lead to their conclusions (Zerilli et al., 2018; Roscher et al., 2020), while in the case of humans one is satisfied with descriptions and rationalizations that leave the black box of the partner's mind and brain obscure. In the legal field, for example, explainability requirements "are stronger when the decision-making process is completely automated" (Bibal et al., 2021, p. 150). The sociological approach abandons this double standard and adopts a more realistic attitude toward machines: also interacting with algorithms it is not necessary to know the *causes*, which are notoriously a problem even dealing with the processes of the human mind (Kahneman et al., 1982; Pearl & Mackenzie, 2018), and all the more so when dealing with a radically different mode of functioning (Shmueli, 2010). It is sufficient to obtain *reasons* that are understandable to the user. The goal is to make machines, opaque or not, produce "reformulations" of their processes that match the demands of their interlocutors and allow them to exercise the form of control appropriate to the context. Different users in different contexts get different explanations, that enable them to explore, verify, contest what the machine has done.

This seems to be also the purpose of the European Commission's recent proposal (28.9.2022)[8] to manage the social consequences of the opacity of algorithms.[9] The proposal intends to update AI liability to include cases involving black box AI systems that are so complex, autonomous and opaque that it becomes difficult for victims to identify in detail how the damage was caused. As the GDPR (European Union, 2016) declares, however, recipients of automated decisions must "be able to express their point of view and to contest the decision" (Art. 22.3). The goal of the directive is to eliminate the legal uncertainty associated with black box algorithms[10] by abandoning the requirement of transparency and opting for a "presumption of causality" in cases of opacity (Art. 4) — in practice, requiring convincing explanations. Even when harm is produced by an intransparent algorithm, the company using it and the company producing it must respond to requests and explain that they have done everything necessary to avoid the problems — enabling the recipients to challenge their decisions. As already stated earlier, though, the companies using the algorithms have to deliver motivations, not "a complex explanation of the algorithms used or the disclosure of the full algorithm" (European Data Protection Board, 2017, p. 25).

Also in the technical field, several recent reflections on XAI seem to go more or less explicitly in this direction, moving away from the goal (or illusion) of transparency. According to Gunning (2017), for example, "XAI will create a suite of machine learning techniques that enables human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners."[11] Arrieta et al. (2020) observe that human be-

8. https://ec.europa.eu/info/business-economy-euro/doing-business-eu/contract-rules/digital-contracts/liability-rules-artificial-intelligence_en

9. According to Bibal et al. (2021, p. 156), legal obligations on explainability pursue two main objectives. The first one is to allow the recipients of a decision to understand its rationale and to act accordingly. The second one is to allow the public authority to exercise a meaningful effective control on the legality of decisions (European Commission, 2020, p. 14).

10. On the responsibility gaps caused by AI-systems see Beckers & Teubner (2021).

11. See also Gilpin (2018): explainable models "are able to summarize the reasons for neural network behavior, gain the trust of users, or produce insights about the causes of their decisions."

ings are not all the same, and what is understandable for some users may remain obscure for others. One must take into account the specific *audience*, or what Langer et al. (2021) call the stakeholders: various groups of people in different situations, with different competences and different interests in the results of the artificial processing of data. The definition of Explainable AI is then integrated as follows: "Given an audience, an explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand" (Arrieta et al., 2020, p. 85) — where clarity does not necessarily imply transparency.

What does this mean concretely? Once the idea (or the possibility) of transparency-by-design is abandoned and the inevitable obscurity of models is accepted, several approaches aim at *post-hoc explainability*, having the algorithms reproduce what happens in communication (Lipton, 2018). The processes by which we explain our actions and decisions are separate from those by which we make them. First we act and then, if required, we produce explanations fitting the needs of our partners. In Wright Mills' words (1940), we produce "additional (or 'ex post facto') lingualizations" (p. 907) reacting to the requests by others — or other ex post explanations using non-verbal tools: gestures, images, various kinds of evidence. Similarly, in the field of XAI, designers are training programs to produce explanations that reformulate a posteriori how algorithms work. As effective explanations do not require that our partner gets to know the operations of our neurons in human communication, in the same way the processes that produce explanations of the work of the algorithms can be different from the processes of the algorithms. Summarizing a growing trend in research (Guidotti et al., 2018; Mittelstadt et al. 2019), Bibal et al. (2021) define explainability as "the capacity of a model to be explainable by using methods that are external to the black-box model" (p. 150). Explanations may use, for example, machine-produced verbal arguments, visualizations, local tools such as salience maps, examples, simplifications or feature relevance. All these techniques have the purpose to allow the user to autonomously make sense of the results of the machine, producing information that may, and in most cases will, be different from the one used by the machine (Esposito, 2022, Ch. 3).

Obviously, the explanations provided, being formulated in response to the requests of the communication partner, will be different for each interaction, i.e., automatically tailored to the specific audience — who will understand, or not understand, what they can, according to their skills and expertise. What the user understands of the explanations of the machine, in any case, need not be the processes of the machine.

4 Conclusions

By integrating the sociological perspective into the XAI debate, the issue can be reframed in such a way that the painful trade-off between model explainability and model performance need not be taken for granted, and a more positive approach becomes possible. Explanations do not need to affect the process of the machines and do not have to reduce their performance. The need and the ability to provide explanations arises at a second stage, after the systems have performed their operations and achieved their results — post-hoc. To provide an explanation for its processes, AI systems do not have to become more stupid — they must learn to communicate (Esposito, 2022).

As a consequence, however, the scope and claims of XAI also change. The goal of its research is only to make possible an effective meta-communication between users and algorithms, which allows for control even under opaque conditions. That is, the goal is to provide explanations, which are only communications. That the explanation provided by the algorithms is

successful, however, in no way implies that the use of the results of the work of the algorithm will be successful. It only means that users can make sense of what the machine communicates to them, not that they use it in the right way. Explainable AI gives no guarantee for *Responsible AI* inquiring whether the results of the explanations and the related understanding lead to beneficial effects or deleterious social consequences (e.g. Theodorou & Dignum, 2020; Mikalef et al., 2022). This important issue does not concern Explainable AI and is of course completely open (Keenan & Sokol, 2023).

References

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. https://doi.org/10.1177/1461444816676645

Arrieta, A.B., 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. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bateson, G. (1972). *Steps to an Ecology of Mind*. Chicago, IL: University of Chicago Press.

Beckers, A., & Teubner, G. (2021). *Three Liability Regimes for Artificial Intelligence: Algorithmic Actants, Hybrids, Crowds*. Oxford: Hart. https://doi.org/10.5040/9781509949366

Bibal, A., Lognoul, M., de Streel, A. et al. (2021). Legal Requirements on Explainability in Machine Learning. *Artificial Intelligence and Law*, *29*(2), 149–169. https://doi.org/10.1007/s10506-020-09270-4

Bucher, T. (2018). *If... Then: Algorithmic Power and Politics*. Oxford: Oxford University Press.

Burrell, J. (2016). How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms, *Big Data & Society*, *3*(1). https://doi.org/10.1177/2053951715622512

Busuioc, M. (2020). Accountable Artificial Intelligence: Holding Algorithms to Account. *Public Administration Review*, *81*(5), 825–836. https://doi.org/10.1111/puar.13293

Buhrmester, V., Münch, D. & Arens, M. (2019). Analysis of Explainers of Black Box Deep Neural Networks for Computer Vision: A Survey. *arXiv*, 1911.12116. https://arxiv.org/pdf/1911.12116.pdf

Cimiano, P., Rudolph, S. & Hartfiel, H. (2010). Computing Intensional Answers to Questions – An Inductive Logic Programming Approach. *Data & Knowledge Engineering*, *69*(3), 261–278. https://doi.org/10.1016/j.datak.2009.10.008

Coeckelbergh, M. (2020). *AI Ethics*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/12549.001.0001

Doshi-Velez, F. & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. *arXiv:1702.08608v2*. https://arxiv.org/abs/1702.08608v2

Eco, U. (1975). *Trattato di semiotica generale*. Milano: Bompiani.

Esposito, E. (2022). *Artificial Communication. How Algorithms Produce Social Intelligence*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/14189.001.0001

European Commission (2020). White Paper on Artificial Intelligence – A European approach to Excellence and Trust. European Commission. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020DC0065&from=FR

European Data Protection Board (2017). Guidelines of the European Data Protection Board on Automated Individual Decision-making and Profiling. European Data Protection Board. https://ec.europa.eu/newsroom/article29/items/612053

European Union (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

Frosst, N. & Hinton, G. (2017). Distilling a Neural Network Into a Soft Decision Tree. *arXiv*, 1711.09784. https://arxiv.org/abs/1711.09784

Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining Explanations: An Overview of Interpretability of Machine Learning. *arXiv*, 1806.00069. https://arxiv.org/pdf/1806.00069.pdf

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning. Adaptive Computation and Machine Learning*, Cambridge, MA: MIT Press.

Grice, H.P. (1975). Logic and Conversation. In P. Cole & J.L. Morgan, *Speech Acts* (pp. 41–58). New York, NY: Academic Press. https://doi.org/10.1163/9789004368811_003

Guidotti, R., Monreale, A., Ruggieri , S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, *51*(5), 1–42. https://doi.org/10.1145/3236009

Gunning, D. (2017). Explainable Artificial Intelligence (XAI) (Technical Report). Defense Advanced Research Projects Agency.

Heider, F. (1958). *The Psychology of Interpersonal Relations*. New York, NY: Wiley. https://doi.org/10.1037/10628-000

Hempel, C.G. (1966). *Philosophy of Natural Science*. Englewood Cliffs, NJ: Prentice-Hall.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511809477

Keenan, B., & Sokol, K. (2023). Mind the Gap! Bridging Explainable Artificial Intelligence and Human Understanding with Luhmann's Functional Theory of Communication. *arXiv*, 2302.03460. https://doi.org/10.48550/arXiv.2302.03460

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 Stake-

holder Perspective on XAI and a Conceptual Model Guiding Interdisciplinary XAI Research. *arXiv*, 2102.07817v1. https://doi.org/10.48550/arXiv.2102.07817

Latour, B. (1999). *Pandora's Hope: Essays on the Reality of Science Studies*. Cambridge, MA: Harvard University Press.

LeCun, Y., Bengio, Y., & Hinton G. (2015). Deep Learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lipton, Z.C. (2018). The Mythos of Interpretability. *ACM*, *16*(3), 31–57. https://doi.org/10.1145/3236386.3241340

Luhmann N. (1995). Was ist Kommunikation?. In *Soziologische Aufklärung*, Vol. 6 (pp. 109–120). Opladen: Westdeutscher.

Luhmann N. (1997). Die Gesellschaft der Gesellschaft. Frankfurt am Main: Suhrkamp.

Malle, B.F. (1999). How People Explain Behavior: A New Theoretical Framework. *Personality and Social Psychology Review*, *3*(1), 23–48. https://doi.org/10.1207/s15327957pspr0301_2

Mikalef, P., Conboy, K., Eriksson Lundström J., & Popovič, A. (2022). Thinking Responsibly about Responsible AI and The Dark Side' of AI. *European Journal of Information Systems*, *31*(3), 257–268. https://doi.org/10.1080/0960085X.2022.2026621

Miller, T. (2019). Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence*, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining Explanations in AI. In d. boyd & J. Morgenstern (Eds.), *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 279–288). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3287560.3287574

Montavon, G., Samek, W., & Müller, K. (2018). Methods for Interpreting and Understanding Deep Neural Networks. *Digital Signal Processing*, *73*, 1–15. https://doi.org/10.1016/j.dsp.2017.10.011

O'Hara, K. (2020). Explainable AI and the Philosophy and Practice of Explanation. *Computer Law & Security Review*, *39*. https://doi.org/10.1016/j.clsr.2020.105474

Pasquale, F. (2015). *The Black Box Society. The Secret Algorithms that Control Money and Information*. Cambridge, MA: Harvard University Press. https://doi.org/10.4159/harvard.9780674736061

Pearl, J., & Mackenzie D. (2018). *The Book of Why: The New Science of Cause and Effect*. New York, NY: Basic Books.

Robbins, S. (2019). A Misdirected Principle with a Catch: Explicability for AI. *Minds and Machines*, *29*, 495–514. https://doi.org/10.1007/s11023-019-09509-3

Rohlfing, K., Cimiano, P., Scharlau, I., Matzner, T., Buhl, H.M., Buschmeier, H., Esposito, E., Grimminger, A., Hammer, B., Häb-Umbach, R., Horwath, I., Hüllermeier, E., Kern, F., Kopp, S., Thommes, K., Ngonga Ngomo, A.-C., Schulte, C., Wachsmuth, H., Wagner, P., Wrede, B. (2021). Explanations 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. https://doi.org/10.1109/TCDS.2020.3044366

Roscher, R., Bohn, B., Duarte, M.F., & Garcke, J. (2020). Explainable Machine Learning for Scientific Insights and Discoveries. *IEEE Transactions on Cognitive and Developmental Systems*, *8*, 42200–42216.

Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stake Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*, *1*, 206–215. https://doi.org/10.1038/s42256-019-0048-x

Shannon, C.E., & Weaver, W. (1949). *The Mathematical Theory of Communication*. Urbana, IL: University of Illinois Press.

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science*, *25*(3), 289–310. https://doi.org/10.1214/10-STS330

Theodorou, A., & Dignum, V. (2020). Towards Ethical and Socio-Legal Governance in AI. *Nature Machine Intelligence*, *2*(1), 10–12. https://doi.org/10.1038/s42256-019-0136-y

Tilly C. (2006). *Why?*. Princeton, NJ: Princeton University Press.

Vilone, G., & Longo, L. (2021). Notions of Explainability and Evaluation Approaches for Explainable Artificial Intelligence. *Information Fusion*, *76*, 89–106. https://doi.org/10.1016/j.inffus.2021.05.009

Von Hilgers, P. (2011). The History of the Black Box: The Clash of a Thing and Its Concept. *Cultural Politics*, *7*(1), 41–58. https://doi.org/10.2752/175174311X12861940861707

Weinberger, D. (2018). 3 Principles for Solving AI Dilemma: Optimization vs Explanation. *KDnuggets*. https://www.kdnuggets.com/2018/02/3-principles-ai-dilemma-optimization-explanation.html

von Wright, G.H. (1971). *Exploration and Understanding*. Ithaca, NY: Cornell University Press.

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2018). Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard? *Philosophy & Technology*, *32*, 661—683. https://doi.org/10.1007/s13347-018-0330-6

Elena Esposito – Department of Political and Social Sciences, University of Bologna (Italy); Faculty of Sociology, University of Bielefeld (Germany)
https://orcid.org/0000-0002-3075-292X
elena.esposito9@unibo.it; https://www.unibo.it/sitoweb/elena.esposito9/
Elena Esposito is Professor of Sociology at the University of Bielefeld (Germany) and the University of Bologna (Italy). She has published extensively on the theory of society, media theory, memory theory and the sociology of financial markets. Her current research on algorithmic prediction is supported by a five-year Advanced Grant from the European Research Council. Her latest book is *Artificial Communication: How Algorithms Produce Social Intelligence* (MIT Press, 2022).