References

Balsemão Pires E. (2013) The epistemological meaning of Luhmann's critique of classical ontology. In: Systema: Connecting matter, life, culture, and technology 1(1): 5–20. ▸ https://cepa.info/1126

Bourdieu P. (1980) Le sens pratique. Les Éditions de Minuit, Paris. English translation: Bourdieu P. (1990) The logic of practice. Translated by Richard Nice. Polity Press, Cambridge UK.

Farley B. & Wesley Clark W. (1954) Simulation of self-organizing systems by digital computer. Transaction of the IRE Professional Group on Information Theory 4(4): 76–84.

Farley B. (1960) Self-organizing models for learned perception. In: Yovits M. & Cameron S. (eds.) Self-organizing systems. Proceedings of an interdisciplinary conference 5 and 6 May 1959. Pergamon Press, Oxford, New York: 7–30.

Hebb D. (1949) The organization of behavior: A neuropsychological theory. John Wiley & Sons, New York.

Koselleck R. (1992) Kritik und Krise: Eine Studie zur Pathogenese der bürgerlichen Welt. Seventh edition. Suhrkamp, Frankfurt am Main. Originally published in 1959. English translation: Koselleck R. (1988) Critique and crisis: Enlightenment and the pathogenesis of modern society. Berg Publishers, Oxford.

Latour B. (2012) Enquête sur les modes d'existence: Une anthropologie des modernes. La Découverte, Paris. English translation: Latour B. (2013) An inquiry into modes of existence: An anthropology of the moderns. Translated by Catherine Porter. Harvard University Press, Cambridge MA.

Luhmann N. (1978) Soziologie der Moral [Sociology of morals]. In: Luhmann N. & Pfürtner S. H. (eds.) Theorietechnik und Moral. Suhrkamp, Frankfurt am Main.

Luhmann N. (1992) Beobachtung der Moderne. Westdeutscher Verlag, Opladen. Luhmann N. (1998) Observations on modernity. Translated by William Whobrey. Stanford University Press, Stanford CA.

Luhmann N. (1995) Social Systems. Stanford University Press, Stanford. German original publication in 1984.

Luhmann N. (1997) Die Gesellschaft der Gesellschaft, Volume 1. Suhrkamp, Frankfurt am Main. English translation: Luhmann N. (2012) Theory of society, Volume 1. Translated by Rhodes Barrett. Stanford University Press, Stanford CA.

McCulloch W. & Pitts W. (1943) A logical calculus of the ideas immanent in nervous activity. In Bulletin of Mathematical Biophysics 5: 115–133.

Parsons T. (1968) The structure of social action, Volume 1. The Free Press, New York.

Parsons T. & Shils E. A. (2001) Toward a general theory of action: Theoretical foundations for the social sciences. Transaction Publishers, New Brunswick.

Rosenblatt F. (1962) Principles of neurodynamics: Perceptrons and the theory of brain mechanisms. Spartan Books, Washington DC.

Edmundo Balsemão Pires is Professor of Philosophy at the University of Coimbra, Portugal. His scientific interests are focused on theory of systems and theory of society, constructivism, Hegel and Luhmann. https://constructivist.info/authors/edmundo-balsemão-pires

Funding: The author did not receive any funding while writing the manuscript.
Competing interests: The author declares that he has no competing interests.

Received: 31 May 2021
Revised: 28 June 2021
Accepted: 29 June 2021

Author's Response
Opacity and Complexity of Learning Black Boxes

Elena Esposito
Bielefeld University, Germany
elena.esposito/at/uni-bielefeld.de

> Abstract • Non-transparent machine learning algorithms can be described as non-trivial machines that do not have to be understood, but controlled as communication partners. From the perspective of sociological systems theory, the normative component of control should be addressed with a critical attitude, observing what is normal as improbable.

Handling Editor • Alexander Riegler

« 1 » The commenters' insightful remarks, in addition to productively extending the range of issues covered in our discussion, provide me with an opportunity to clarify some fundamental points – or at least to express my own views on them. For this I am very grateful. My response starts with the relationship between opacity and complexity in the operation of recent algorithms (reacting to comments by **Bernd Porr**, **Manfred Füllsack**, **Wiebke Loosen & Armin Scholl**) and then addresses other issues that highlight the specificity and productivity of the constructivist approach.

« 2 » In the widespread debate on algorithms as black boxes (e.g., Pasquale 2015), **Porr** (§6) usefully reminds us of Ross Ashby's (1951) classic reflections on the black box as a model of the interaction between observers and systems: something is a black box if it is obscure to its observer irrespective of its internal characteristics. If we now have a problem of obscurity in dealing with the latest algorithms that use advanced deep-learning techniques working with big data, a constructivist approach asks, first of all, to which observer the algorithms appear as black boxes, and for what reasons. The obscurity may be due to reasons extrinsic to the algorithms themselves, such as information limitations due to the confidentiality needs or desires of companies, or simply the lack of expertise of some observers. Here I agree with **Füllsack**'s (§1) observation that now-established statistical

techniques such as some types of Bayesian inference have certainly appeared opaque for a long time. This has always been the case with technologies (Latour 1999), and is nothing new. What then is the specificity of the much-discussed opacity of algorithms?

« 3 » Opacity may also be due to intrinsic factors, related to the specific mode of operation of algorithmic machines, which are built to learn – and to learn by themselves – from clues largely unknown to their programmers. The machines derive them from the huge amounts of heterogeneous data they access on the web. In some cases, algorithms decide for themselves what to learn and how, triggering procedures that are incomprehensible to the very programmers who designed them (Burrell 2016). Lack of expertise is not the issue: the procedures of algorithms are impenetrable to any human observer, no matter how informed or how competent. However, the constructivist perspective has something to say about this, too. Starting from Ashby's concept of the black box, Heinz von Foerster (1985: 131ff) introduced the notion of "non-trivial machines," whose behavior – like that of today's algorithms – is impenetrable for any external observer, no matter how competent. These are, as is well known (see also Foerster 2003: 311f), machines, or algorithms, whose behavior is determined not only by the inputs they receive but also by their internal state, and the internal state changes depending on the inputs – like self-learning algorithms, which use information to perform their task and also to modify themselves. Although the functions that regulate the behavior of the machines and the transitions of their internal states are fully determined, their behavior is unpredictable for external observers. At different times, depending on its history of interactions, the machine gives different responses to the same inputs, and one cannot understand why. The algorithm has learned, becoming "analytically indeterminable" (Foester 1985: 131).

« 4 » In my view, then, the opacity of self-learning algorithms does indeed have an innovative aspect, related to their unprecedented ability to learn, and to learn autonomously, which makes them inherently non-trivial machines. However, I agree with **Porr** (§5) that deep learning does not involve any deep understanding. As **Bruno Clarke** (§10) argues, many of the recent successes of digital machines are not because machines have finally learned to understand content, but because programmers have given up trying to produce machines that understand (Esposito 2017). They now accept and exploit that they work as black boxes.[1] Deep-learning algorithms work in a fundamentally different way from human intelligence, which is why they are often incomprehensible to human observers. In the opacity of algorithms, however, there is not necessarily anything mysterious. It does not imply consciousness, autonomous will or intelligence, nor the decisions of a superior entity, as in the case of divination. Algorithms, like von Foerster's non-trivial machines, are "synthetically determined," i.e., constructed by someone in a certain way, which remains the basis of their behavior. Algorithms follow the instructions of programmers, who made them so that they would learn in a complex way – and as a result they become obscure.

« 5 » However, opacity generates problems, and they are different problems from those addressed by the current approach to technology, driven by control of causality.[2] How can one control technology without claiming to control causes? This was already the question underlying the debate on "high technologies" in the 1980s (e.g., Perrow 1984; Luhmann 1991: 98–117): how can damage and accidents be avoided when the technology involves a very high complexity of different processes taking place simultaneously? Or in the case of complex algorithmic systems such as recent neural networks: how can one control the relationships between processes that occur independently at many distinct levels?

« 6 » **Loosen & Scholl** state it effectively (§7): the problem is not one of "reliability" but of validity or desirability. That is, the problem is not how to make the algorithms behave as they have been instructed to: they are determined machines and do what is requested. The result, though, however formally correct, may not be adequate. I also agree with **Loosen & Scholl**'s argument that the debate about bias (the "original sin of algorithms"[3]) is only one side of the issue (§7), but it is an aspect with an interesting ambiguity: bias is the other side of the performance of machines that appear intelligent. Without bias, this performance could not happen. Algorithms, which are structure-determined machines, in the case that machines appear intelligent "feed" on the contingency of users' behavior, expressed in their participation in Web 2.0, to implement the "double contingency" that makes the algorithms creative and effective (Esposito 2021: Ch. 1). As **Loet Leydesdorff** (§8) and **Füllsack** (§§5f) observe, in reference to Talcott Parsons, there are always too many options available, and the best way to manage them without eliminating them is to collectively constrain them, together with other communication participants.[4] Algorithms learn to select and use possibilities starting from the choices made by users on the web[5]

1| In the terms of **Leydesdorff** (§§7f), the dynamics of redundancy take the place of the dynamics of information and entropy.

2| It is not by chance that one of the most debated issues in this regard is the contrast between causality and correlations. It is on the basis of correlations that algorithms identify patterns in the available data, and use them to make predictions without indicating causes and without providing explanations. This debate was also sparked by Chris Anderson's article on the "end of theory" (*Wired*, 23 August 2008, https://www.wired.com/2008/06/pb-theory), with which, let it be said, incidentally, I do not agree – contrary to what **Porr** (§4) states. From a sociological perspective, the attitude of a second-order observer is to observe the current debate along with the conditions from which it emerges.

3| See audiobook "Sex, race, and robots: How to be human in the age of AI" by Ayanna Howard, 2019, published by Audible Originals, LLC.

4| In the terms of Nilaks Luhmann's social systems theory, cited by **Leydesdorff** (§2), the complexity of possibilities is reduced and maintained at the same time. When the stalemate of Parsons's train compartment is overcome by starting to talk about something, the complexity of possible topics is reduced, but all further possibilities for producing different contributions and learning new things are generated.

5| The first model is, of course, Google's PageRank algorithm, which already worked in this way before deep-learning techniques (Langville & Meyer 2006).

– but user behavior is inevitably biased, and so is algorithm behavior.

« 7 » The validity discussed by **Loosen & Scholl**, however, goes beyond bias, and introduces another sociological consideration: even if algorithms work correctly (they are "reliable"), their result is not necessarily socially correct, or desirable. In our research on the social impact of Predictive Policing algorithms[6] we investigate this component, referring to the distinction between predictive effectiveness and preventive effectiveness. Even if predictive algorithms were correct (which is known to be highly doubtful as, e.g., Kristian Lum and William Isaacs 2016 point out), in the social context in which they are employed, their predictions are no guarantee of correct prevention. In an influential study, Bernard Harcourt (2007) argues that the spread of algorithmic tools in criminal law risks undermining the efficacy of prevention. If profiled persons are less responsive to policy change than non-profiled persons, concentrating crime prevention on the people at risk identified by algorithms can be counterproductive. On that view, the profiled individuals do not change their behavior, because they often have no choice and commit crimes anyway, while other areas of the population where surveillance and prevention could be effective remain uncovered and overall crime increases. In cases like this, the prevention activity guided by correct predictions would not be successful – or valid, in the sense of **Loosen & Scholl**. It would also not comply with von Foerster's "Ethical Imperative," discussed by **Füllsack** (§2).

« 8 » The discourse on validity almost inevitably implies a normative component, mentioned by **Loosen & Scholl** (§8) and addressed more extensively by **Edmundo Balsemão Pires** in his **Q1** and **Q2** about the "moral core" of the idea of criticism and how it can be made valuable in a constructivist approach – also questioning the social function of critique. Referring to my claim (§12 of the interview) that systems theory could be understood, somewhat ironically, as a more critical version of the critical theory of the Frankfurt school, **Balsemão Pires** points out the deep interconnection between normative and descriptive components in the critical tradition and is skeptical about the possibility of "avoid[ing] the moral grounds of the world in the *Aufklärung*" (§5). I certainly agree with this observation, and, in my opinion, it is reminiscent of the basic paradox of any attempt to distance oneself from the critical approach: can one dissociate oneself from critique without making a critique of critique? The fascination and the problem of critique rely on a paradox: critique cannot properly be criticized without confirming critique, at the same time. However, **Balsemão Pires** also observes that the constructivist approach can offer an alternative, based on the reflexivity of second-order observation and the inevitability of the blind spot (Foerster 2003: 212f). I refer, in this regard, to Luhmann's theory, which, as **Clarke** says (§4), offers a far-reaching development of second-order cybernetics. It can also make it possible to reformulate the idea of criticism from a constructivist perspective.

« 9 » For sociological systems theory, at the level of second-order observation, the alternative to critique cannot be a refusal of the critical attitude, but rather the recognition of the blind spot of every observation perspective, including its own. The blindness of the Frankfurt school's critical theory is the blindness of the external observer, who claims to be in a position to indicate what is right and what is wrong. Since critical theory does not recognize its blindness, it can include a normative attitude, detecting crises and indicating how to overcome them. Critical theory assumes that one can refuse current society and indicate how it should be instead. However, systems theory starts from an "autological" assumption (Luhmann 1997: 16ff), recognizing that sociology is part of the society it observes and cannot take an external position that would enable it to overcome the blindness of the observed society. The critic revealing the blind spot has herself a blind spot that she cannot observe – or can only observe by moving to a different perspective with a different blind spot. One cannot overcome the blindness, but can be aware of it and take it into account in one's own observation. As **Clarke** argues (§9), acceptance of the finitude of one's own constructions takes the place of a priori foundation. Systems theory cannot adopt a normative attitude and indicate what should be done. Yet critical attitude and critical theory are separate, and one can still criticize (observe a blind spot) without "knowing better." Luhmann presents this option as having a "significant critical potential" (Luhmann 1993: 1), if critique is understood not as a "call to refusal" but as a "sharper, not self-evident ability to distinguish" (ibid: 2). In the case of sociology, a critical approach requires that one take a distance from what appears normal in society. On this understanding, critique does not refuse what is normal but observes "the other side of the normal form" (ibid) – an attitude for which systems theory is especially well equipped. Its "methodological recipe" in the analysis of social reality is namely "to look for theories that succeed in explaining what is normal as improbable" (Luhmann 1984: 161).

6| Under project ERC PREDICT.

Acknowledgements

I am grateful to Bénédicte Zimmermann and Katrin Sold for their competent, inspiring and very pleasant conduct of the interview, and to the Wissenschaftskolleg zu Berlin for their kind support.

References

Burrell J. (2016) How the machine "thinks": Understanding opacity in machine learning algorithms. Big Data & Society 1: 1–12.

Esposito E. (2017) Artificial communication? The production of contingency by algorithms. Zeitschrift für Soziologie 46(4): 249–265. ▸ https://cepa.info/7142

Esposito E. (in press) Artificial communication: How algorithms produce social intelligence. MIT Press, Cambridge MA.

Foerster H. von (1985) Cibernetica ed epistemologia: Storia e prospettive [Cybernetics and epistemology: History and perspectives]. In: Bocchi G. & Ceruti M. (eds.) La sfida della complessità. Feltrinelli, Milano: 112–140.

Foerster H. von (2003) Understanding understanding. Springer, New York.

Harcourt B. E. (2007) Against prediction: Profiling, policing, and punishing in an actuarial age. University of Chicago Press, Chicago.

Langville A. N. & Meyer C. D. (2006) Google's PageRank and beyond: The science of search engine rankings. Princeton University Press, Princeton NJ.

Latour B. (1999) Pandora's hope: Essays on the reality of science studies. Harvard University Press, Cambridge MA.

Luhmann N. (1984) Soziale systeme. Suhrkamp, Frankfurt am Main. English translation: Luhmann N. (1995) Social systems. Stanford University Press, Stanford CA.

Luhmann N. (1991) Soziologie des Risikos [Sociology of risk]. De Gruyter, Berlin.

Luhmann N. (1993) "Was ist der Fall?" und "Was steckt dahinter?" – Die zwei Soziologien und die Gesellschaftstheorie ["What is the case?" and "What's behind it?" – The two sociologies and the social theory]. Zeitschrift für Soziologie 22(4): 245–260.

Luhmann N. (1997) Die Gesellschaft der Gesellschaft [The society of society]. Suhrkamp, Frankfurt am Main.

Lum K. & Isaac W. (2016) To predict and serve? Significance 13(5): 14–19. https://rss.onlinelibrary.wiley.com/doi/epdf/10.1111/j.1740-9713.2016.00960.x

Pasquale F. (2015) The black box society: The secret algorithms that control money and information. Harvard University Press, Cambridge MA.

Perrow C. (1984) Normal accidents: Living with high-risk technologies. Basic Books, New York.

Funding: This work was supported by the European Research Council (ERC) under Advanced Research Project PREDICT no. 833749.

Competing interests: The author declares that she has no competing interests.

Received: 15 July 2021
Revised: 16 July 2021
Accepted: 16 July 2021