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

Explaining Machines: Social Management of Incomprehensible Algorithms. Introduction

Elena Esposito*

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

Submitted: January 16, 2023 – Accepted: March 8, 2023 – Published: March 15, 2023

Abstract

This short introduction presents the symposium 'Explaining Machines'. It locates the debate about Explainable AI in the history of the reflection about AI and outlines the issues discussed in the contributions.

Keywords: Explainable AI; inexplicability; transparency; explanation; opacity; contestability.

* elena.esposito9@unibo.it

The reflection about AI did not originate with computer scientists. Cybernetics, where it all began, was invented by Norbert Wiener (1948), who was first of all a philosopher, and developed in the interdisciplinary debates of the Macy Conferences, which involved anthropologists, psychologists, social scientists, neurophysiologists — as well as mathematicians and computer scientists (Pias, 2003). In the subsequent decades, the debates about strong AI, weak AI, and machine consciousness were led by philosophers and linguists (e.g., Dreyfus, 1979; Hofstadter, 1979; Searle, 1980; Haugeland, 1985; Churchland & Churchland, 1990) and driven by a strong awareness that the underlying challenges of digitization were not solely computational.

In the following period, boosted by advances in programming and by the explosive development of the web and the availability of (big) data, the contribution of the humanities and social sciences to thinking about AI receded into the background, often confined to supporting tasks such as the classificatory problems of ontologies (Mizoguchi & Borgo, 2021). Sociologists and media and communication scholars have predominantly converged in the direction of critical media studies, which has made important contributions to the identification and conceptualization of issues of bias, inequalities, and social consequences of the use of algorithms, such as filter bubbles, echo chambers, and different forms of polarization.

In the past few years, however, the latest "spring" of AI has marked a turning point: we are no longer talking about computers but algorithms, not about sampling but about big data, and especially about programs that evolve autonomously with advanced machine learning techniques (Esposito, 2022). The shift has brought different issues and new conceptual challenges to the forefront: first, all the enigmas related to the opacity of algorithms that are increasingly and more radically incomprehensible. In the emerging branch of Explainable AI (XAI), questions of sociological relevance not only concern the application of digital programs in different social domains, but also affect programming techniques. How should algorithms be designed that allow humans (users but also programmers) to exercise control over their results? What does explanation mean in the digital world, what is explained, how and to whom? In a recent article Borch and Min (2022) argue that "explainability research is too important to be left to the computer/data scientists and ML engineers populating the field of XAI research." (p. 11) The contribution of the humanities and social sciences is returning to a central position. Also among computer scientists there is a growing realization that the management of incomprehensible machines requires the contribution of non-computational skills — even and precisely to improve computational techniques.

This awareness was the background for the "Explaining Machines" conference held in Bielefeld in June 2022. The deliberate ambiguity of the title was intended to invoke an open question in the field of XAI, which requires the contribution of both computational and sociological knowledge: are machines to explain themselves or are humans to explain them? Or perhaps both at the same time? Then the needs of users, with all their diversity and contextuality, can be seen not as an additional complication, but as a contribution to the programming of machines that are not only more controlled, but possibly more efficient and more effective.

It is no coincidence that the conference was the inaugural event of a large Collaborative Research Center entitled *Constructing Explainability* at the Universities of Bielefeld and Paderborn, funded by the German Research Foundation (DFG) with more than 14 million € for the first four years (twelve planned). The conference presentation states that

> the originality of our approach lies in addressing explainable AI (XAI) as a co-constructive, social process. This is the main difference to current approaches in XAI, which mainly focus on making the inner workings of AI systems more transparent, whereas transparency usually means being understandable to computer sci-

> ence experts. In contrast, we approach explanation as a co-constructive process that can involve different addressees (experts, lay users, professionals, authorities, etc.) and approach XAI from an inherently interdisciplinary perspective: most projects are directed jointly by researchers from the social sciences and computer science.

The contributions published in this symposium result from that initiative and address the broad spectrum of questions that the management of incomprehensible machines poses to the social sciences. For example, how should the legal right to an explanation be interpreted, if the fundamental condition of contestability of decisions requires a legal justification of the outcome of black box machine learning systems that is different from a technical explanation (Hildebrandt, 2022)? What form of normativity is implied by opaque technical systems that performatively act on their environment (Rieder et al., 2022)? Is it still possible, and how, to criticize the operation of algorithms when the number of variables and the complexity of procedures make it impossible for designers to project a hypothesis space that connects the input data (observed space) and the results of the calculation (decision space) (John-Mathews & Cardon, 2022)? Dealing with highly efficient machines that are inherently incomprehensible, does the requirement of explainability make it necessary to reduce their performance, or does the sociological analysis allow for the separation of explainability from the demand for transparency (Esposito, 2022b)? And in general, should the inexplicability of algorithms be considered a failure, or can their ability to find patterns in the randomness of data rather be an opportunity to revise our notion of control and recognize the role of accidents and of the particular (Weinberger, 2022)?

All these questions require skills and approaches that do not belong to the expertise of computer scientists and data scientists, although of course they demand a thorough understanding of recent research in programming techniques and close collaboration with scholars active in that field. As the founders of cybernetics understood, advances in programming today are closely linked to advances in the reflection of the humanities and social sciences about their implications and consequences.

References

Borch, C., & Hee Min, B. (2022). Toward a Sociology of Machine Learning Explainability: Human-Machine Interaction in Deep Neural Network-based Automated Trading. *Big Data & Society*, *9*(2). https://doi.org/10.1177/20539517221111361

Churchland, P., & Churchland, P. (1990). Could a Machine Think? *Scientific American*, *262*(1), 32–39. https://doi.org/10.1038/scientificamerican0190-32

Dreyfus, H. (1979). *What Computers Still Can't Do*. New York, NY: MIT Press.

Esposito, E. (2022a). *Artificial Communication: How Algorithms Produce Social Intelligence*. Cambridge: MIT Press.

Esposito, E. (2022b). Does Explainability Require Transparency?. *Sociologica*, *16*(3), 17–27. https://doi.org/10.6092/issn.1971-8853/15804

Haugeland, J. (1985). *Artificial Intelligence: The Very Idea*. Cambridge, MA: MIT Press.

Hildebrandt, M. (2022). Qualification and Quantification in Machine Learning. From Explanation to Explication. *Sociologica*, *16*(3), 37–49. https://doi.org/10.6092/issn.1971-8853/15845

Hofstadter, D.R. (1979). *Gödel, Escher, Bach: An Eternal Golden Braid*. New York, NY: Basic Books.

John-Mathews, J.-M., & Cardon, D. (2022). The Crisis of Social Categories in the Age of AI. *Sociologica*, *16*(3), 5–15. https://doi.org/10.6092/issn.1971-8853/15931

Mizoguchi, R., & Borgo, S. (2021). Towards an Ontology of Representation. In F. Neuhaus & B. Brodaric (Eds.), *Proceedings of the 12th Int'l Conf. (FOIS), Frontiers in Artificial Intelligence and Applications* (pp. 48–63). Amsterdam: IOS Press.

Pias, C. (2003). *Cybernetics. The Macy-Conferences 1946–1953*. Zürich/Berlin: Diaphanes.

Rieder, B., Gordon, G., & Sileno, G. (2022). Mapping Value(s) in AI: Methodological Directions for Examining Normativity in Complex Technical Systems. *Sociologica*, *16*(3), 51–83. https://doi.org/10.6092/issn.1971-8853/15910

Searle, J.R. (1980). Minds, Brains and Programs. *Behavioral and Brain Sciences*, *3*(3), 417–424. https://doi.org/10.1017/S0140525X00005756

Weinberger, D. (2022). A Plea for Inexplicability. *Sociologica*, *16*(3), 29–35. https://doi.org/10.6092/issn.1971-8853/15296

Wiener, N. (1948) (1961²). *Cybernetics, or Control and Communication in the Animal and the Machine*. Cambridge, MA: MIT Press.

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).