JOURNAL OF
CROSS-DISCIPLINARY RESEARCH
IN COMPUTATIONAL LAW [CRCL]

Transparency versus explanation: The role of ambiguity in legal AI

Elena Esposito *

Abstract

Dealing with opaque machine learning techniques, the crucial question has become the interpretability of the work of algorithms and their results. The paper argues that the shift towards interpretation requires a move from artificial intelligence to an innovative form of artificial communication. In many cases the goal of explanation is not to reveal the procedures of the machines but to communicate with them and obtain relevant and controlled information. As human explanations do not require transparency of neural connections or thought processes, so algorithmic explanations do not have to disclose the operations of the machine but have to produce reformulations that make sense to their interlocutors. This move has important consequences for legal communication, where ambiguity plays a fundamental role. The problem of interpretation in legal arguments, the paper argues, is not that algorithms do not explain enough but that they must explain too much and too precisely, constraining freedom of interpretation and the contestability of legal decisions. The consequence might be a possible limitation of the autonomy of legal communication that underpins the modern rule of law.

Keywords: explanation; interpretation; ambiguity; rule of law; artificial intelligence; XAI

Replier: Federico Cabitza, University of Milan-Biccoca federico.cabitza@unimib.it

Journal of Cross-disciplinary Research in Computational Law

DOI: pending

www.journalcrcl.org

*Professor, Bielefeld University elena.esposito@uni-bielefeld.de

Introduction: from artificial intelligence to artificial communication

After repeated 'winters',[1] AI research now seems to be in a new 'spring' – in which, however, the machines, the way of working and even the problems have changed. Today we talk more about algorithms than computers. We take for granted the reference to the web (including the active participation of users) and the fact that the data to be processed is not scarce but rather overabundant. We are in the world of self-learning algorithms and big data. In this new phase the central problem is not the capacity or processing power of computers.[2] Today the crucial question is interpretation, or rather the *interpretability* of algorithms[3] and of the results of their work.

The paper argues that the shift towards interpretation requires that research on digital information processing moves from the reference to (artificial) intelligence to the reference to an innovative form of communication, which can be defined as artificial.[4] The goal is not to build intelligent machines but to be able to communicate with algorithms to obtain relevant and controlled information. What must be understood is the information generated in this communication and not the processes of the machines, which are and often must remain obscure. I make my point in the next two sections of the paper that deal with the issue of transparency and the goal of explanations.

The shift from intelligence to communication brings up problems and opportunities in many different fields, including the complex area of legal interpretation, addressed in the section 'Artificial reason and mechanical jurisprudence'. There I discuss the role of interpretation for the autonomy of the legal system, and in the following section I explore the need for ambiguity in legal argumentation and the resulting challenges for the use of algorithms. 'Mechanical jurisprudence' can affect legal practice and the principles it relies on, notably the rule of law.

The interpretation of incomprehensible machines

The recent emphasis on the problem of interpretation is a consequence of innovation in programming techniques and data management. With deep learning methods, and using big data, algorithms autonomously learn to perform their tasks in ways that were not necessarily foreseen by their programmers and that in some cases are incomprehensible to humans, including those who designed them. Even the programmers might not understand how the machine proceeds and how it achieves its results.[5] When one needs to understand the results and procedures of algorithms one has to interpret them and it is not clear how that should be achieved.

Algorithms that work with machine learning and big data are getting better and better at doing more and more things: they produce information quickly and accurately;

1 Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach* (Prentice Hall 2003) p. 29; Dominique Cardon, Jean-Philippe Cointet, and Antoine Mazieres, 'La revanche des neurons. L'invention des machines inductives et la controverse de l'intelligence artificielle' (2018) 211(5) Réseaux, p. 173.

2 Which people have tried to predict with Moore's law and its variants.

3 In recent discourse about AI and its transformations, the use of the term 'algorithm' is often inaccurate. Of course, computer programming has been using algorithms since the beginning, and the term already existed before cybernetics. In this text I follow current usage, however imperfect, and use 'algorithms' to refer to advanced programming techniques that use machine learning and big data. See Stephen F Deangelis, 'Artificial Intelligence. How Algorithms Make Systems Smart' [2014] Wired (accessed 23 June 2021)

4 Elena Esposito, 'Artificial Communication? The Production of Contingency by Algorithms' (2017) 46(4) Zeitschrift für Soziologie, p. 249; Elena Esposito, *Artificial Communication. How Algorithms Produce Social Intelligence* (MIT Press 2021).

5 Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning (Adaptive Computation and Machine Learning)* (MIT Press 2016); Jenna Burrell, 'How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms' (2016) 3(1) Big Data & Society; David Weinberger, 'Our Machines Now Have Knowledge We'll Never Understand' [2017] Wired (accessed 23 June 2021); Leilani H Gilpin and others, 'Explaining Explanations: An Overview of Interpretability of Machine Learning' [2018] ; Madalina Busuioc, 'Accountable Artificial Intelligence: Holding Algorithms to Account' (2020) 81(5) Public Administration Review 825.

they are learning to drive cars more safely and reliably than humans; they can answer our questions, make conversation, compose music, and read books; and they can even write interesting, appropriate, and – if required – funny texts. They have achieved these results, which seem to suggest that machines have finally become intelligent, since their programmers have more or less explicitly given up trying to artificially reproduce the processes of human intelligence. Algorithms work in a radically different way that can be incomprehensible to our intelligence; transparency, or the lack of it, is therefore an issue.

Machine learning algorithms are difficult to understand, first of all because they work without understanding their materials – they do something different. Recent translation programs, for example, do not try to understand the documents they translate and their designers do not rely on any theory of language.[6] Algorithms translate texts from Mandarin without knowing Mandarin; their programmers do not know it either. Examples multiply across all areas in which algorithms are most successful, for example competing with human players in chess, poker and Go,[7] producing text, recommendation programs,[8] image recognition and many others. Algorithms do not understand anything of the materials they are dealing with; they 'don't reason like people in order to write [or, one might add, to work in general] like people'.[9] Therefore, the operations of the machines and their results are often obscure to human observers.

Even if they are very effective, however, reliance on black boxes is not reassuring, especially when we know that their operations are not immune from bias and errors of various kinds.[10] In many cases we want to check the correctness of the results produced by the machines, which can be wrong or inappropriate in many different ways, and with differing consequences. In the medical field, for example, there is concern that algorithms may not adequately take into account information which, while relevant, may not be explicit.[11] For example, Caruana discusses an algorithm that predicted that asthma patients were at lower risk of death from pneumonia, disregarding the fact that the patients were already receiving intense medical assistance.[12] In other fields, such as policing,[13] the granting of consumer credit,[14] or university admission procedures,[15] there is concern that through systemic or confirmation bias they may reproduce or intensify imbalances in the data. Consequently one would like to be able to verify their results and control the way they are obtained. In the legal field, discussed in more detail later, the obscurity of algorithmic procedures can jeopardise the contestability of decisions.

The recent branch of research on 'explainable AI' (XAI) tries to respond to this concern by developing procedures to explain the operations of self-learning algorithms.[16] The results clarify various aspects of the processes of interaction with machines and are often very useful in managing such processes in specific situations. However, in the case of deep learning algorithms there is a basic obstacle: if by explanation one means a procedure that allows human observers to understand what the machine does and why, the enterprise is hopeless. The processes of recent algorithms that appear intelligent are intrinsically incomprehensible to human intelligence. As Weinberger claims, requiring an explanation in this sense would amount to 'forc[ing] AI to

6 Tom Boellstorff, 'Making Big Data, in Theory' (2013) 18(10) First Monday.

7 David Silver and Demis Hassabis, 'AlphaGo: Mastering the Ancient Game of Go with Machine Learning' [2016] (accessed 23 June 2021).

8 Robert Prey, 'Nothing Personal: Algorithmic Individuation on Music Streaming Platforms' (2018) 40(7) Media, Culture & Society 1087.

9 Kristian Hammond, *Practical Artificial Intelligence for Dummies* (Wiley 2015).

10 Frank Pasquale, *The Black Box Society. The Secret Algorithms That Control Money and Information* (Harvard University Press 2015).

11 Andreas Holzinger and others, 'Causability and explainability of artificial intelligence in medicine' (2019) 9(4) WIREs Data Mining and Knowledge Discovery e1312.

12 Rich Caruana and others, *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 'Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-day Readmission' (2015).

13 Kristian Lum and William Isaac, 'To Predict and Serve' (2016) 13(5) Significance 14.

14 Cathy O'Neil, *Weapons of Math Destruction* (Crown 2016).

15 Karen Hao, 'The UK Exam Debacle Reminds Us that Algorithms Can't Fix Broken Systems' [2020] MIT Technology Review (accessed 2 October 2020).

16 Sandra Wachter, Brent Mittelstadt, and Luciano Floridi, 'Transparent, Explainable and Accountable AI for Robotics' (2017) 2(6) Science Robotics (eaan6080); Finale Doshi-Velez and others, 'Accountability of AI Under the Law: The Role of Explanation' [2017] ; Tim Miller, 'Explanation in Artificial Intelligence: Insights from the Social Sciences' (2019) 267 Artificial Intelligence 1.

be artificially stupid enough that we can understand how it comes up with its conclusion'.[17]

The strategy must be different and indeed many projects on XAI have recently adopted another approach, compatible with the radical obscurity of algorithmic processes.[18] The key notion is *transparency*, often taken as the first element of explainable AI projects.[19] However, the debate involves many other related notions, whose relationships are not always clear,[20] as well as human-computer interactions far beyond the deep learning issues that triggered it. When and why does it become necessary to explain the operations of algorithms? Should the purpose of explanation be transparency? What is the relationship between transparency and opacity, and between explanation and interpretation? What must be explained, to whom and for what purposes? And when can one say that an explanation has actually been produced? The answer to these questions concerns the very interpretation of AI and its social relevance.

Does explanation require transparency?

In the sociological study of technology, lack of transparency has been a long-standing issue.[21] The problem becomes even more acute in the case of algorithms. Here I want to distinguish a specific kind of non-transparency, which can be called *opacity*, concerning recent machine learning methods, such as neural networks, which use 'black-box' algorithms.[22] The corresponding models can be radically incomprehensible to human observers, however experienced. Other models that are in principle understandable (not opaque), like 'white-box' algorithms based on decision trees[23] or inductive logic programming,[24] may, however, also turn out to be non-transparent, because of their size or complexity, because of restricted access to relevant information (such as sourcing and usage of training data or the development and implementation of the model), or in general, because the observer does not have the necessary competences.

In the use of algorithms, non-transparency is much broader than opacity, and even if it were mandatory for all sources of data and of all procedures to be accessible to users, most systems would still be incomprehensible to their users. In itself, however, this is neither new nor problematic: the internal working of technology has always been incomprehensible to most users.[25] The issue is rather that today algorithms do something unprecedented, different from other technological systems: they make *decisions* – on medical diagnoses, the selection of students to be admitted to universities, the moves to be made in Go, and the people to be given credit or parole. It is these decisions that must be explained, not the internal processes of the machines. The purpose of XAI is actually explanation, not transparency, and from this point of view the opacity of

17 Weinberger (n 5). On the highly debated trade-off between explainability and accuracy in the performance of algorithms cf. Finale Doshi-Velez and Been Kim, 'Towards A Rigorous Science of Interpretable Machine Learning' [2017] ; Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller, 'Methods for Interpreting and Understanding Deep Neural Networks' (2018) 73 Digital Signal Processing 1; Don Monroe, 'AI, Explain Yourself' (2018) 61(11) Communications of the ACM 11; Cynthia Rudin, 'Stop Explaining Black Box Machine Learning Models for High Stake Decisions and Use Interpretable Models Instead' (2019) 1 Nature Machine Intelligence 206; Busuioc (n 5).

18 Katharina J Rohlfing and others, 'Explanations as a Social Practice: Toward a Conceptual Framework for the Social Design of AI Systems' (2020) 13(3) IEEE Transactions on Cognitive and Developmental Systems.

19 Ribana Roscher and others, 'Explainable Machine Learning for Scientific Insights and Discoveries' (2020) 8 IEEE Access.

20 See e.g. Monroe (n 17); Mike Ananny and Kate Crawford, 'Seeing Without Knowing: Limitations of the Transparency Ideal and its Application to Algorithmic Accountability' (2018) 20(3) New Media & Society 973; Zachary C Lipton, 'The Mythos of Interpretability' (2018) 16(3) ACM Queue 1; Kieron O'Hara, 'Explainable AI and the Philosophy and Practice of Explanation' (2020) 39 Computer Law & Security Review.

21 Johannes Weyer and Ingo Schulz-Schaeffer (eds), *Management Komplexer Systeme: Konzepte Für die Bewältigung von Intransparenz, Unsicherheit und Chaos* (De Gruyter 2009); Niklas Luhmann, *Die Kontrolle von Intransparenz* (Suhrkamp 2017).

22 Vanessa Buhrmester, David Münch, and Michael Arens, 'Analysis of Explainers of Black Box Deep Neural Networks for Computer Vision: A Survey' [2019] .

23 JR Quinlan, 'Induction of Decision Trees' (1986) 1 Machine Learning 81.

24 Steven Muggleton and Luc De Raedt, 'Inductive Logic Programming: Theory and Methods' (1994) 19-20 The Journal of Logic Programming 629.

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

deep learning systems makes no difference; understanding AI is not the issue anyway.

The goal is not to disclose the procedures of the machines, but to make the machines themselves provide explanations that are informative for the user. Machines are not asked to be transparent to human observers, but to explain their decisions in a way that makes sense to their interlocutors. And since their interlocutors are always different and located in different situations and contexts, with different interest and needs, the explanations will have to be diverse and specific. The issue is to provide appropriate explanations to different users.

This is what happens when humans make decisions, for which we might also be required to offer explanations, giving clues that enable the recipient to make sense of the decision. When one obtains an explanation, one gets information about the decision without being informed about the neurophysiological or psychic processes of the explainer, which (fortunately) can remain obscure or private. Explaining our decisions does not require disclosure of our thought process, far less the connections of our neurons. Explanations, Luhmann claims, are 'reformulations with the added benefit of better connectivity'.[26] The issuer produces a new communication that provides additional elements related to the specific request of the interlocutor and his needs. In any case, this is an entirely communicative process: we do not need to access the brain or the mind of our interlocutors, nor do we need to access the external world. We only need to get clues that allow the communication to proceed in a controlled, non-arbitrary way.

The same approach can be envisioned for dealing with the dilemmas of explanation in the interaction with self-learning machines. Many have suggested that only inherently understandable models should be used in cases where explanation may be needed.[27] However, this does not solve the general problem from which the need for explainability arises.[28] Instead, machines, opaque or otherwise, should be capable of producing 'reformulations' of their processes that match the requests of their interlocutors and allow them to exercise the form of control appropriate to the context. The technical challenge in interactions with a digital partner is to reproduce the communicative situation in which explanations are requested and provided between human beings.

In fact, many recent XAI projects do not attempt to mimic the calculations made by the algorithm, but rather aim to produce 'post-hoc explanations' that reproduce what humans do in communication. Transparency cannot be the solution, because, as Lipton claims, however transparency is understood (at the level of the entire model, at the level of the individual components or at the level of the training algorithms) human explanations do not exhibit transparency.[29] The processes by which people explain their decisions are distinct from those by which they make them and are usually produced after the fact, without affecting decision-making. Similarly, in the field of XAI, designers are training programs to produce explanations that illustrate (we could say 'reformulate') after the fact the working of algorithms, without impacting their performance. Just as the linguistic processes that generate human explanations differ from the neural processes that produce the decisions to be explained, so the processes producing explanations of AI models will be different from the processes of the model.[30] They can, for example, use verbal explanations produced by the machine, visualisations and local explanations such as saliency maps.[31] The user's understanding of the explanations produced by the machine does not have to relate to the processes of the machine.

[26] Niklas Luhmann, *Die Wissenschaft der Gesellschaft* (Suhrkamp 1990) (translation of the author of this paper).

[27] See e.g. Scott Robbins, 'A Misdirected Principle With a Catch: Explicability for AI'' (2019) 29 Minds and Machines 495; Rudin (n 17).

[28] In addition to producing a difficult Catch-22. Robbins (n 27): 'If ML is being used for a decision requiring an explanation then it must be explicable AI and a human must be able to check that the considerations used are acceptable, but if we already know which considerations should be used for a decision, then we don't need ML.'

[29] Lipton (n 20) p. 15.

[30] Since successful explanations by algorithms do not require access to the working of the algorithms, the black-box nature of deep learning algorithms makes no difference for their explainability. On the contrary, complex algorithms such as deep neural networks may be more efficient in learning which representations are more effective in the communication with users (ibid).

[31] ibid p. 15 et seq.

This promising perspective implies a profound change from the approach that has guided AI projects since their beginnings in the 1950s – as the very name Artificial Intelligence indicates. Somewhat contradictorily, recent XAI projects are not focused on the intelligence of the machine. The goal is rather to produce a condition of 'dialogue' between the algorithm and the user in which the machine provides answers, taking as input the always-different requests for clarification of its interlocutors,[32] and is able to participate in a meta-communication[33] that can have as its object the processes of the machine or the data used. The purpose is not and cannot be that the interlocutors understand these processes, but that they interpret what the machine communicates about these processes in such a way that they can exercise a form of control. The debate on explanation implies a shift from intelligence to the very feature that enables algorithms to effectively contribute to the production of new information in our society: their ability to participate in communication. Machines should be able to produce adequate explanations in response to different requests from their interlocutors.

Artificial reason and mechanical jurisprudence

If XAI implies a move from the focus on intelligence to the focus on communication, the task of sociological observation would be to show how interactions with algorithms affect communication in society in general,[34] and specifically how algorithmic explanations work as communicative processes that rely on opacity. This might happen in different ways in different domains of society. In scientific research, for example in medicine, attention will be directed toward the possibility of uncovering causal structures in the data;[35] in policing it will be directed toward trust in the decisions of the algorithms; when algorithms decide on the selection of candidates or debtors the issue will be whether algorithmic decisions conform to ethical principles. This section explores the legal field: how lack of transparency and its management in the working of algorithms can affect legal practice and its presuppositions, notably the rule of law.

In the legal field, today algorithms are able to fulfil many tasks in a cheap, effective and quick way: they can automate the completion of documents, perform due diligence, gather and analyse past data, sort through legal information and carry out other activities that previously required human work. The resulting opportunities and associated risks to jobs have provoked a wide debate both in the legal field and in other sectors.[36] The issue we want to address here is more abstract and complex, involving the role of interpretation in legal arguments. Here too, computers can be usefully employed to perform many tasks. People talk about 'mechanical jurisprudence'[37] or 'computational legal science',[38] computational systems of legal reasoning that are capable of exploring legal databases,[39] identifying relevant rules, making decisions,[40] generating arguments

[32] See Philipp Cimiano, Sebastian Rudolph, and Helena Hartfiel, 'Computing Intensional Answers to Questions – An Inductive Logic Programming Approach' (2010) 69(3) Data & Knowledge Engineering 261; Rohlfing and others (n 18).

[33] Gregory Bateson, *Steps to an Ecology of Mind* (1972); N Luhmann, *Die Gesellschaft der Gesellschaft* (Suhrkamp 1997) pp. 250-251.

[34] Luhmann (n 33) p. 304; Esposito, 'Artificial Communication? The Production of Contingency by Algorithms' (n 4).

[35] The lively debate about the difference between correlation and causation in science is an influential case, triggering a deep rethinking of basic epistemological issues such as the relationship between explanations and predictions. See Judea Pearl, *Causality* (Cambridge University Press 2000); Judea Pearl and Dana Mackenzie, *The Book of Why: The New Science of Cause and Effect* (Basic Books 2018); Leo Breiman, 'Statistical Modeling: The Two Cultures' (2001) 16(3) Statistical Science 199; Galit Shmueli, 'To Explain or to Predict?' (2010) 25(3) Statistical Science 289; Elliott Sober, *Ockham's razors: a user's manual* (Cambridge University Press 2016).

[36] See Richard Susskind, *The End of Lawyers? Rethinking the Nature of Legal Services* (Oxford University Press 2008).

[37] Douglas Walton, Fabrizio Macagno, and Giovanni Sartor, *Statutory Interpretation. Pragmatics and Argumentation* (Cambridge University Press 2021).

[38] Nicola Lettieri, 'Ex Machina: Analytical Platforms, Law and the Challenges of Computational Legal Science' (2018) 10(5) Future Internet.

[39] Nikolaos Aletras and others, 'Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing Perspective' (2016) 2:e93 PeerJ Computer Science.

[40] Reuben Binns, 'Analogies and disanalogies between machine-driven and human-driven legal judgement' (2020) 1(1) Journal of Cross-disciplinary Research in Computational Law.

[41] Kevin Ashley, *Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age* (Cambridge University Press 2017).

and also explaining their chain of reasoning to the users.[41] Machines autonomously participate in legal communication: they can generate legally relevant information, make an argument and even explain it.

Here the problem is deeper and concerns not only the possible threat to the skills of human workers and to their jobs. It concerns the fundamentals of modern positive law, which involve the autonomy of law and the question of *interpretation*. As Hildebrandt has argued, our form of legal system has developed as a result of the spread of the printing press and the resulting changes in the way we produce, write and read texts.[42] The printing press produces standardised, identical and immutable texts, which are removed from the 'mouvance' of oral communication and manuscripts[43] – books that escape the practice of commentary. In previous texts, in a culture that remained predominantly oral, glosses and commentaries were added in each reading and became part of the text, which changed ('moved') continuously, each time producing a different communication.[44] The 'moving' text incorporated the interpretation.

When, with the printing press, the text became fixed and remained the same in all readings, interpretations multiplied and became variable. Writing, Luhmann argues, gives rise to the difference between text and interpretation, which the printing press generalises.[45] The fixed text must be interpreted to make sense in the specific context. However, the situations in which a text is read are all unique, different from any other; if the text remains the same, the way of considering it must change. The plurality of interpretations is inevitable and legitimate: since contexts and circumstances are always different, interpretations must vary to take them into account.[46] The interpretations of the same text, therefore, can always be different, and any interpretation can be challenged.

This happens in all fields that have to do with texts, but in legal practice it takes a more complicated form.[47] If laws are written texts and legal decisions also take this form, a lot of interpretive work is needed to take into account the variety of circumstances and legal cases. Judges interpret laws and previous cases and their observers (lawyers, litigants, the public) interpret their decisions. According to Hildebrandt, freedom of interpretation is the foundation of the modern rule of law. This freedom is the basis of the autonomy of the judiciary. It allows the judiciary to follow its own logic and criteria. These are not dictated by the sovereign and may conflict with the principles and preferences of political power. In Fried's terms, '[t]he law's rationality is a rationality apart', that does not follow the principles of general rationality but only the 'artificial reason of the law'.[48]

Autonomy of interpretation is a basic requirement for the independence of the law, but it does not mean arbitrariness or obscurity. Judges' decisions must be explained, that is, motivated (in legal terms) in accordance with the specific rationality of the law, making explicit the reasons on which they are based. According to this rationality, then, explanations are interpreted and decisions can be challenged. 'The purpose of interpretation is not to ensure that all readers understand the text in the same way, but that different people facing the same text participate in a unitary communication'.[49] This is the kind of transparency required by the controlled functioning of the legal system and the one according to which the possible transparency of algorithms must be evaluated. Does the explanation provided by artificial intelligence in mechanical jurispru-

[42] Most recently in Mireille Hildebrandt, *Law for Computer Scientists and Other Folk* (Oxford University Press 2020); see also Niklas Luhmann, *Das Recht der Gesellschaft* (Suhrkamp 1993) p. 349.

[43] Paul Zumthor, *Introduction à la poésie orale* (Seuil 1972); Elizabeth L Eisenstein, *The Printing Press as an Agent of Change. Communications and Cultural Transformations in Early-Modern Europe* (Cambridge University Press 1979).

[44] Jan Assmann and Burkhard Gladigow (eds), *Text und Kommentar. Archäologie der Literarischen Kommunikation IV* (Fink 1995).

[45] Luhmann, *Das Recht der Gesellschaft* (n 42) p. 362.

[46] Elena Esposito, *Soziales Vergessen. Formen und Medien des Gedächtnisses der Gesellschaft* (Suhrkamp 2002) p. 226-227.

[47] On the performativity of language see John L Austin, *How to Do Things with Words* (Oxford University Press 1962). In the legal field this is a basic condition: the words pronounced by a judge or a legislator are immediately fact and have concrete consequences.

[48] Charles Fried, 'Artificial Reason of the Law or: What Lawyers Know' (1981) 60(1) Texas Law Review 35 pp. 35, 39 and 58. Sociological system theory describes this condition as out-differentiation *(Ausdifferenzierung)* of the legal system in modern society. See Luhmann (n 33) p. 743 et seq.

[49] Luhmann, *Das Recht der Gesellschaft* (n 42) p. 362 (translation of the author of this paper).

dence meet the requirements of the 'artificial reason of the law'? Can a decision taken on the basis of automated procedures be justified in such a way as to allow the functioning of legal communication and possibly contestation by the people involved? Is the lack of transparency of algorithms, which as we have seen is unavoidable in their communicative use, compatible with the transparency requirements of legal decisions?

The role of ambiguity in legal arguments

On one level this would seem to be the case. That the digital processes leading to the decision are different from those of our intelligence and possibly not accessible or understandable to human observers, with regard to legal communication this does not necessarily mark a caesura with decisions taken by human agents. As Canale and Tuzet claim, '[j]urisdictional motivation does not consist in the psychological account of the process that led to the decision, but in the indication of the legal reasons that justify it',[50] or as Luhmann claims, '[t]he argument does not reflect what the reader has in mind'.[51] A correct motivation does not imply that the thoughts and steps that led to the decision are described, and therefore, it can be argued, neither should it be necessary to describe the processes followed by the algorithm to get to its result. It is not necessarily a problem that digital processes are incomprehensible to human beings, if the algorithm is able to explain its decision in a communicative sense, i.e. to indicate in a comprehensible way the legal reasons that led to it or, in the sense of Fried, the artificial reason on which it is based.

On a second level, however, things are more complicated. From a sociological perspective, the performance of the law for society as a whole is the 'absorption of uncertainty' in the management of litigation.[52] It must be possible to rely on the fact that legal rules are applied to concrete cases and in a valid way.[53] In order to absorb uncertainty, validity must be argued (motivated), i.e. the legal decision must be justified by providing grounds for it. Since the cases to be dealt with are always different, the grounds must be appropriate to the context,[54] but the very decision on what counts as context can be controversial and lead to doubts and disagreement.[55] In most cases, moreover, much of the evidence presented by both sides to support their arguments is based on conflicting rules and precedents.[56] Although all legal decisions refer to the same body of rules, the arguments (explanations) must be different from case to case and flexibly coordinated with each other.

For coordination to be possible, *ambiguity* plays a fundamental role in legal communication.[57] The arguments 'are typically vague and ambiguous',[58] that is, 'susceptible of more than one reasonable interpretation'.[59] Legal norms are characterised by multiple layers of ambiguity that hinder their organisation in a formal, fully consistent whole. In typical cases of legal argument 'inconsistency is

50 Damiano Canale and Giovanni Tuzet, *La Giustificazione della Decisione Giudiziale* (Giappichelli 2020).

51 Luhmann, *Das Recht der Gesellschaft* (n 42) p. 362 (translation of the author of this paper).

52 Niklas Luhmann, *Recht und Automation in der öffentlichen Verwaltung* (Duncker & Humblot 1966) pp. 56-57. In March and Simon's classic definition: 'Uncertainty absorption takes place when inferences are drawn from a body of evidence and the inferences, instead of the evidence itself, are then communicated', James G March and Herbart A Simon, *Organizations* (Wiley 1958) p. 165.

53 Nicola Lettieri, 'Law, Rights, and the Fallacy of Computation' (2020) XVII(2) Jura Gentium 72, p. 72.

54 Walton, Macagno, and Sartor (n 37).

55 Frank H Easterbrook, 'The Absence of Method in Statutory Interpretation' (2017) 84(1) Chicago Law Review 81, pp. 81, 83-84.

56 Donald Berman and Carole Hafner, 'Obstacles to the Development of Logic-Based Models of Legal Reasoning' in Charles Walter (ed), *Computer Power and Legal Language* (Greenwood Press 1988).

57 Lettieri, 'Law, Rights, and the Fallacy of Computation' (n 53); Mireille Hildebrandt, 'The Adaptive Nature of Text-Driven Law' (2020) 1(1) Journal of Cross-disciplinary Research in Computational Law; Wolfgang Hoffmann-Riem, 'Legal Technology/Computational Law: Preconditions, Opportunities and Risks' [2020] (1) Journal of Cross-disciplinary Research in Computational Law.

58 Walton, Macagno, and Sartor (n 37) p. 4.

59 Lawrence Solan, 'Pernicious Ambiguity in Contracts and Statutes' (2004) 79 Chicago-Kent Law Review 859, pp. 859, 862.

60 Walton, Macagno, and Sartor (n 37) p. 5; see also Bernardo Giorgio Mattarella, *La trappola delle leggi: molte, oscure, complicate* (Il Mulino 2011)

61 Luhmann, *Das Recht der Gesellschaft* (n 42) p. 356 (translation of the author of this paper).

the norm',[60] and in fact the objective of the argument can only be 'to avoid visible inconsistencies'.[61] The real goal of the argument is not to achieve abstract logical coherence, but to make the grounds of the decision appear convincing – and a legal justification is convincing not because all of its steps have been checked: 'The rationality of legal problem management lies... not in the logical correctness of its conclusions... It must suffice that it convinces everyone that it has convinced its author'.[62] The motivation (explanation) appears convincing when everyone is convinced that others find it convincing. Rhetorical effectiveness counts more than the logical consequentiality of the steps of the argument, which is not examined in detail.

Lawyers and judges, who are 'the masters of the artificial reason of the law', are by experience and professional expertise very skilled at handling ambiguity and using it for rhetorical purposes, for example by applying 'a trained, disciplined intuition where the manifold of particulars is too extensive to allow our minds to work on it deductively'.[63] The task of lawyers, Garfinkel claims, is to make ambiguous the interpretations of facts and laws.[64] It works well when interacting with human beings, because for effective communication it is sufficient to regulate 'the presentation, not the production of the decision'.[65] Lawyers and judges must provide a convincing account of the decisions they take, but their interpretation can and often must remain vague, because it is 'not concerned with how we understand or produce texts, but with how we establish the acceptability of a specific reading thereof'.[66] What the observers interpret is the usually ambiguous interpretation by the judge or the lawyer.

For algorithms, however, ambiguity is a challenge. The competent management of vagueness is notoriously a problem for machines, which has been discussed for decades in the discourses on the limits of artificial intelligence.[67] Even today it is difficult for algorithms to deal with the various levels of ambiguity always present in human communication or, in the legal field, to manage the multiplicity of possible interpretations of rules and norms.[68] Moreover, if the focus shifts from the intelligence of the machines (what they can understand and how) to their participation in communication, other problems related to ambiguity arise: not only the difficulty for machines of coping with the ambiguity of human communication, but also the difficulty of themselves generating ambiguous communication, i.e. of managing in competent ways the ambiguity required by legal arguments.

Legal explanations produced by algorithms should themselves be ambiguous, in the same way as those that result from the interpretation of legal norms by humans are. Ambiguity is not, as we tend to think, opposite to transparency,[69] but, on the contrary, is necessary to provide the multiplicity of legal interpretations required for contestability. As Hildebrandt states, '[d]ue to the ambiguity inherent in human language, text-driven ICIs[70] generate a specific type of multi-interpretability that in turn generates a specific type of contestability.'[71] To challenge a decision one needs to be able to develop a perspective on the decision that is independent of that provided by the decision maker,[72] i.e. to question their interpretation. However, to do this the motivation must appear legally ambiguous – that is, it must be, as we have seen, susceptible to more than one reasonable interpretation. The machine that does not have its own perspective, does not interpret, so its explanations lack ambiguity. The explanations it offers are reformulations of the decisions that are made following

62 Luhmann, *Recht und Automation in der öffentlichen Verwaltung* (n 52) pp. 55, 59 (translation of the author of this paper).

63 Fried (n 48) p. 57.

64 Harold Garfinkel, *Studies in Ethnomethodology* (Prentice Hall 1967) p. 111.

65 Luhmann, *Recht und Automation in der öffentlichen Verwaltung* (n 52) p. 106 (translation of the author of this paper).

66 Walton, Macagno, and Sartor (n 37) p. 9.

67 E.g. Hubert Dreyfus, *What Computers Can't Do* (The MIT Press 1972).

68 Lettieri, 'Law, Rights, and the Fallacy of Computation' (n 53).

69 Ananny and Crawford (n 20); Johan P Olsen, 'Accountability and Ambiguity' in Mark Bovens, Robert E Goodin, and Thomas Schillemans (eds), *The Oxford Handbook of Public Accountability* (Oxford University Press 2014); Maximilian Heimstädt and Leonhard Dobusch, 'Transparency and Accountability: Causal, Critical and Constructive Perspectives' (2020) 1(4) Organization Theory 1.

70 Information and communication infrastructures.

71 Hildebrandt, 'The Adaptive Nature of Text-Driven Law' (n 57) pp. 7-8.

72 O'Hara (n 20).

further rules, so it makes no sense to ask what the algorithm meant – algorithms do not mean anything.

The lack of a competent management of ambiguity is a problem that is also perceived in experiments that try to realise a form of XAI in the legal field. Even for the most recent computational models producing legal argumentation, lack of ambiguity is a constraint,[73] far beyond what is required in legal communication between human beings guided by the imperative to appear convincing and absorb uncertainty. Paradoxically, then, one could say that the problem of interpretation in legal argumentation – even and precisely when dealing with algorithms that are obscure to human intelligence – is not that the machine does not explain enough, but that it must explain too much, and too precisely. As acknowledged by scholars in this field, this level of detail may obscure rather than illuminate the practice of legal communication:

> We are well aware that in using the structured and formalistic argumentation approach there is the danger of confusing readers more than explaining to them how the courts can do a better job of grappling with the hard (so-called wicked) problems of statutory interpretation.[74]

On the one hand, therefore, there is the risk that the explanation is not convincing. On the other hand, if it *is* convincing, perhaps an even more serious problem may arise: limits may be imposed on the freedom of interpretation that underpins the autonomy of legal communication, and there may be a risk that the use of automated models may alter fundamental features of the rule of law.[75] As we saw above, the 'artificial reason of the law' does not coincide with the general rationality of society or even with the abstract coherence of a logical argument. Mechanical jurisprudence, however, when it identifies and applies the legal rules relevant to the case in question, does not work with the rhetorically effective arguments that characterise legal reasoning and interpretation,[76] which are possibly ambiguous and not fully consistent. The autonomy of legal communication, with all its implications for the structure of modern society, might take a different form as a consequence of the intervention of algorithms in communication.

What freedom remains to those who must interpret a 'mechanical' legal argument? And in particular how can the decision be contested? The arguments produced by algorithms are not interpretations, contingent and revisable, but descriptions of a series of formal steps. The observer can discover a formal error and contest the decision at this level. However, they cannot explore and challenge the interpretation, because the machine did not interpret anything. All arguments that refer to reasons and motives of interpretation, namely, 'the factors that may lead a decision maker to select one or another interpretation',[77] might be de facto disqualified, and with them a fundamental component of legal communication in modern society.

Conclusion: communicating with machines

Observing the challenge posed by opaque algorithms from the perspective of communication discloses a multiplicity of fascinating and difficult questions. Some issues dissolve, such as the one based on the Turing test: we routinely interact with digital partners without wondering if they are human beings or not. Other issues take a different form, for example, the complex problem of bias, which engages both the dimension of algorithmic bias, reflecting the values of the programmers,[78] and that of data bias, depending on the uncoordinated input of billions of participants, sensors and other digital sources.[79] Still other issues arise, connected with the practical experience accumulated in

73 Walton, Macagno, and Sartor (n 37) p. 11.

74 ibid p. 12.

75 Lettieri, 'Law, Rights, and the Fallacy of Computation' (n 53).

76 Ashley (n 41).

77 Walton, Macagno, and Sartor (n 37) p. 97 et seq.

78 Kate Crawford, 'Artificial Intelligence's White Guy Problem' *The New York Times* (25 June 2016).

79 Ninareh Mehrabi and others, 'A Survey on Bias and Fairness in Machine Learning' [2019] .

many fields. The use of algorithms for specific tasks is almost inadvertently leading to the emergence of diverse, and extremely complex, problems related to their involvement in communication. The question of interpretation in legal argumentation is a particularly significant example. The problem is not how machines work, but how they participate in legal communication.

Acknowledgements

This work was supported by the European Research Council (ERC) under Advanced Research Project PREDICT no. 833749. For their helpful comments, criticisms, and suggestions I am grateful to Giancarlo Corsi, David Stark and the anonymous *CRCL* reviewers.

References

Aletras N and others, 'Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing Perspective' (2016) 2:e93 PeerJ Computer Science.

Ananny M and Crawford K, 'Seeing Without Knowing: Limitations of the Transparency Ideal and its Application to Algorithmic Accountability' (2018) 20(3) New Media & Society 973.

Ashley K, *Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age* (Cambridge University Press 2017).

Assmann J and Gladigow B (eds), *Text und Kommentar. Archäologie der Literarischen Kommunikation IV* (Fink 1995).

Austin JL, *How to Do Things with Words* (Oxford University Press 1962).

Bateson G, *Steps to an Ecology of Mind* (1972).

Berman D and Hafner C, 'Obstacles to the Development of Logic-Based Models of Legal Reasoning' in Walter C (ed), *Computer Power and Legal Language* (Greenwood Press 1988).

Binns R, 'Analogies and disanalogies between machine-driven and human-driven legal judgement' (2020) 1(1) Journal of Cross-disciplinary Research in Computational Law.

Boellstorff T, 'Making Big Data, in Theory' (2013) 18(10) First Monday.

Breiman L, 'Statistical Modeling: The Two Cultures' (2001) 16(3) Statistical Science 199.

Buhrmester V, Münch D, and Arens M, 'Analysis of Explainers of Black Box Deep Neural Networks for Computer Vision: A Survey' [2019].

Burrell J, 'How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms' (2016) 3(1) Big Data & Society.

Busuioc M, 'Accountable Artificial Intelligence: Holding Algorithms to Account' (2020) 81(5) Public Administration Review 825.

Canale D and Tuzet G, *La Giustificazione della Decisione Giudiziale* (Giappichelli 2020).

Cardon D, Cointet J.-P, and Mazieres A, 'La revanche des neurons. L'invention des machines inductives et la controverse de l'intelligence artificielle' (2018) 211(5) Réseaux.

Caruana R and others, *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 'Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-day Readmission' (2015).

Cimiano P, Rudolph S, and Hartfiel H, 'Computing Intensional Answers to Questions – An Inductive Logic Programming Approach' (2010) 69(3) Data & Knowledge Engineering 261.

Crawford K, 'Artificial Intelligence's White Guy Problem' *The New York Times* (25 June 2016).

Deangelis SF, 'Artificial Intelligence. How Algorithms Make Systems Smart' [2014] Wired (accessed 23 June 2021).

Doshi-Velez F and Kim B, 'Towards A Rigorous Science of Interpretable Machine Learning' [2017].

Doshi-Velez F and others, 'Accountability of AI Under the Law: The Role of Explanation' [2017].

Dreyfus H, *What Computers Can't Do* (The MIT Press 1972).

Easterbrook FH, 'The Absence of Method in Statutory Interpretation' (2017) 84(1) Chicago Law Review 81.

Eisenstein EL, *The Printing Press as an Agent of Change. Communications and Cultural Transformations in Early-Modern Europe* (Cambridge University Press 1979).

Esposito E, *Soziales Vergessen. Formen und Medien des Gedächtnisses der Gesellschaft* (Suhrkamp 2002).

Esposito E, 'Artificial Communication? The Production of Contingency by Algorithms' (2017) 46(4) Zeitschrift für Soziologie.

— *Artificial Communication. How Algorithms Produce Social Intelligence* (MIT Press 2021).

Fried C, 'Artificial Reason of the Law or: What Lawyers Know' (1981) 60(1) Texas Law Review 35.

Garfinkel H, *Studies in Ethnomethodology* (Prentice Hall 1967).

Gilpin LH and others, 'Explaining Explanations: An Overview of Interpretability of Machine Learning' [2018].

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

Hammond K, *Practical Artificial Intelligence for Dummies* (Wiley 2015).

Hao K, 'The UK Exam Debacle Reminds Us that Algorithms Can't Fix Broken Systems' [2020] MIT Technology Review (accessed 2 October 2020).

Heimstädt M and Dobusch L, 'Transparency and Accountability: Causal, Critical and Constructive Perspectives' (2020) 1(4) Organization Theory 1.

Hildebrandt M, *Law for Computer Scientists and Other Folk* (Oxford University Press 2020).

— 'The Adaptive Nature of Text-Driven Law' (2020) 1(1) Journal of Cross-disciplinary Research in Computational Law.

Hoffmann-Riem W, 'Legal Technology/Computational Law: Preconditions, Opportunities and Risks' [2020] (1) Journal of Cross-disciplinary Research in Computational Law.

Holzinger A and others, 'Causability and explainability of artificial intelligence in medicine' (2019) 9(4) WIREs Data Mining and Knowledge Discovery e1312.

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

Lettieri N, 'Ex Machina: Analytical Platforms, Law and the Challenges of Computational Legal Science' (2018) 10(5) Future Internet.

— 'Law, Rights, and the Fallacy of Computation' (2020) XVII(2) Jura Gentium 72.

Lipton ZC, 'The Mythos of Interpretability' (2018) 16(3) ACM Queue 1.

Luhmann N, *Die Gesellschaft der Gesellschaft* (Suhrkamp 1997).

Luhmann N, *Recht und Automation in der öffentlichen Verwaltung* (Duncker & Humblot 1966).

— *Die Wissenschaft der Gesellschaft* (Suhrkamp 1990).

— *Das Recht der Gesellschaft* (Suhrkamp 1993).

— *Die Kontrolle von Intransparenz* (Suhrkamp 2017).

Lum K and Isaac W, 'To Predict and Serve' (2016) 13(5) Significance 14.

March JG and Simon HA, *Organizations* (Wiley 1958).

Mattarella BG, *La trappola delle leggi: molte, oscure, complicate* (Il Mulino 2011).

Mehrabi N and others, 'A Survey on Bias and Fairness in Machine Learning' [2019].

Miller T, 'Explanation in Artificial Intelligence: Insights from the Social Sciences' (2019) 267 Artificial Intelligence 1.

Monroe D, 'AI, Explain Yourself' (2018) 61(11) Communications of the ACM 11.

Montavon G, Samek W, and Müller K.-R, 'Methods for Interpreting and Understanding Deep Neural Networks' (2018) 73 Digital Signal Processing 1.

Muggleton S and De Raedt L, 'Inductive Logic Programming: Theory and Methods' (1994) 19-20 The Journal of Logic Programming 629.

O'Hara K, 'Explainable AI and the Philosophy and Practice of Explanation' (2020) 39 Computer Law & Security Review.

O'Neil C, *Weapons of Math Destruction* (Crown 2016).

Olsen JP, 'Accountability and Ambiguity' in Bovens M, Goodin RE, and Schillemans T (eds), *The Oxford Handbook of Public Accountability* (Oxford University Press 2014).

Pasquale F, *The Black Box Society. The Secret Algorithms That Control Money and Information* (Harvard University Press 2015).

Pearl J, *Causality* (Cambridge University Press 2000).

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

Prey R, 'Nothing Personal: Algorithmic Individuation on Music Streaming Platforms' (2018) 40(7) Media, Culture & Society 1087.

Quinlan JR, 'Induction of Decision Trees' (1986) 1 Machine Learning 81.

Robbins S, 'A Misdirected Principle With a Catch: Explicability for AI'' (2019) 29 Minds and Machines 495.

Rohlfing KJ and others, 'Explanations as a Social Practice: Toward a Conceptual Framework for the Social Design

of AI Systems' (2020) 13(3) IEEE Transactions on Cognitive and Developmental Systems.

Roscher R and others, 'Explainable Machine Learning for Scientific Insights and Discoveries' (2020) 8 IEEE Access.

Rudin C, 'Stop Explaining Black Box Machine Learning Models for High Stake Decisions and Use Interpretable Models Instead' (2019) 1 Nature Machine Intelligence 206.

Russell S and Norvig P, *Artificial Intelligence: A Modern Approach* (Prentice Hall 2003).

Shmueli G, 'To Explain or to Predict?' (2010) 25(3) Statistical Science 289.

Silver D and Hassabis D, 'AlphaGo: Mastering the Ancient Game of Go with Machine Learning' [2016] (accessed 23 June 2021).

Sober E, *Ockham's razors: a user's manual* (Cambridge University Press 2016).

Solan L, 'Pernicious Ambiguity in Contracts and Statutes' (2004) 79 Chicago-Kent Law Review 859.

Susskind R, *The End of Lawyers? Rethinking the Nature of Legal Services* (Oxford University Press 2008).

Wachter S, Mittelstadt B, and Floridi L, 'Transparent, Explainable and Accountable AI for Robotics' (2017) 2(6) Science Robotics (eaan6080).

Walton D, Macagno F, and Sartor G, *Statutory Interpretation. Pragmatics and Argumentation* (Cambridge University Press 2021).

Weinberger D, 'Our Machines Now Have Knowledge We'll Never Understand' [2017] Wired (accessed 23 June 2021).

Weyer J and Schulz-Schaeffer I (eds), *Management Komplexer Systeme: Konzepte Für die Bewältigung von Intransparenz, Unsicherheit und Chaos* (De Gruyter 2009).

Zumthor P, *Introduction à la poésie orale* (Seuil 1972).

A reply: Lost in communication? We need a more conscious and interactive use of AI

Federico Cabitza • University of Milan-Bicocca federico.cabitza@unimib.it

I would like to focus on the point made by Esposito when she wonders 'What freedom remains to those who must interpret a "mechanical" legal argument? And in particular, how can the decision be contested?'. Here, I see a potential misunderstanding that we should be aware of, and avoid, in this and similar discussions. Machines – especially those developed using Machine Learning (ML) techniques – can only make arguments and decisions, or even just 'speak the truth' (which cannot be contested), *to the extent that we allow them*. In fact, the output of ML systems is always[1] *probabilistic* in nature; the clear-cut categories (such as right/wrong, guilty/innocent or ill/healthy) in which their output is displayed are simply derived by comparing predefined thresholds with the probabilistic scores that these machines produce to express the intrinsic uncertainty of their classification (or prediction).

Following Sadin [11], I call this misunderstanding the *alethic stance*, which exacerbates the tendency to consider the output of computational processes as more scientific, neutral or objective than any human output [9], despite extensive evidence to the contrary [6, 4].

All computing technologies are rule-based writing technologies, performing the function of automating symbolic inscription. Just as we expect scribes and accountants not to make errors in transcriptions or calculations, we expect computations to transform input configurations into output configurations *correctly*. Most traditional computing tools give us the *right* answers faster and more efficiently than we can do by hand, by virtue of their deterministic rule-following, algorithmic nature. Conversely, data-driven decision support developed using ML techniques aims to give us *likely accurate and precise* results based on *training cases* that are most similar to the case for which we need an answer (assuming, among other things, that no relevant characteristic is ignored or neglected). In other words, these systems will always and only tell us something about the cases on which they have been 'trained', although we wish they could tell us something about the new cases we supply them.

Thus, if we adopt the *alethic stance*, we expect ML tools to relieve us from the burden of ignorance and uncertainty, not just by showing us what is true (e.g. that the fourth decimal of the square root of pi is 4, as any regular calculator can do), but rather by revealing what would otherwise be unknown and hidden, such as the future. The *alethic stance* also requires a reversal of Tertullian's well-known maxim *'veritas autem suadendo docet non docendo persuadet'*.[2]

From this stance, we also *need* ML to convince (us and) others that, by following its advice, we have made an informed decision and chosen the best possible option. This is why explanations are so important when ML machines are embedded in social practices: they are a means of persuasion. Therefore, the problem, as I see it, is how much users *want* to delegate to machines to artificially (and illusorily) dispel uncertainty from decisions, so that they can use them as a safety net (to avoid certain errors, or to prefer certain errors to others), as an *immoral* crumple zone (cf. [5]), or just as a liability shield.

This sheds light on the main difference between my take and what I understood from Esposito's paper. Esposito

[1] There are families of ML models for which this is not properly true (e.g. support vector machines, random forest, or *k-nearest neighbours*), since they do not natively process random variables and probability distributions; nevertheless, it is always possible to interpret these models as probabilistic. In general, we speak here of probability to mean 'predictive uncertainty'. See [2].

[2] 'Truth informs by persuading but does not persuade by informing'.

writes that the problem is 'not that the machine does not explain enough, but that it must explain too much, and too precisely' and also that it is 'not how machines work, but how they participate in legal communication' that matters.

As rightly pointed out, opening black boxes by having 'machines explain too much' is not without its drawbacks. In fact, making AI more *explicable* could have the unintended consequence of fostering the *alethic stance*, and hence reinforcing the persuasive power of the machine by making it a more convincing advisor – that is, an agent that more effectively makes us vulnerable to *automation bias* [8] and therefore better at misleading us when it is wrong. This is the essence of what I called the *'white box paradox'* [1], which requires further research into its effects and their possible mitigators, such as the introduction of *programmed inefficiencies*[3] that nudge users into being more vigilant or warier of the machine's advice.

However, trying to avoid the consequences of the *alethic stance* with a technical contrivance would probably end up aggravating these consequences, also by taking that stance for granted and considering it unchallengeable. Conversely, as argued in [3], we need a cultural countermove: the full realisation that ML systems are probabilistic and hence intrinsically uncertain and *incapable-by-design* of establishing (any) truth. According to this *anti-alethic* stance, we avoid viewing ML systems as (even potentially) trustworthy oracles; rather, we prefer to see them as *perceptual lenses* [10] or *peepholes* that are shaped like the case at hand and can give access to the chamber of 'past wonders' (that is, the cases used as a training set).

This countermove requires us to shift away from the idea that ML systems are *agents* [3] – that is, something with which we can establish some form of communication, in the traditional[4] sense of the term – toward the idea that these tools act as a necessarily distorting medium – one that modulates our interpretation of the present/future through what we have recorded (and interpreted) of the past. Therefore, and to conclude, I believe that the problem is not how machines *work*, but *how we want to relate to them*.

References

[1] Federico Cabitza. 'Cobra AI: Exploring some unintended consequences of Artificial Intelligence'. In: *Machines We Trust. Perspectives on Dependable AI*. Ed. by Marcello Pelillo and Teresa Scantamburlo. MIT Press, 2021.

[2] Federico Cabitza, Andrea Campagner, and Edoardo Datteri. 'To err is (only) human. Reflections on how to move from accuracy to trust for medical AI'. In: *Exploring Innovation in a Digital World. Cultural and Organizational Challenges. Lecture Notes in Information Systems and Organisation*. Ed. by Federica Ceci, Andrea Prencipe, and Paolo Spagnoletti. Vol. 51. Springer International Publishing, 2021.

[3] Federico Cabitza, Andrea Campagner, and Carla Simone. 'The need to move away from agential-AI: Empirical investigations, useful concepts and open issues'. *International Journal of Human-Computer Studies* 155 (2021).

[4] Kate Crawford. *Atlas of AI*. Yale University Press, 2021.

[5] Madeleine Clare Elish. 'Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction'. *Engaging Science, Technology, and Society* 5 (2019).

[6] Lisa Gitelman, ed. *"Raw data" is an oxymoron*. MIT press, 2013.

[7] Mireille Hildebrandt. 'Law as Information in the Era of Data-Driven Agency'. *The Modern Law Review* 79.1 (2016), pp. 1–30.

[8] David Lyell and Enrico Coiera. 'Automation bias and verification complexity: a systematic review'. *Journal of the American Medical Informatics Association* 24.2 (2017), pp. 423–431.

[9] Jerry Z. Muller. *The Tyranny of Metrics*. Princeton University Press, 2019.

[3] This idea has been expressed in several forms and sources, as I acknowledge and reconstruct in [1]. See also [7].

[4] With traditional I mean the common understanding of this term within the human-computer interaction field, where communication, unlike mere interaction, requires meaningful intention and understanding.

[10] Matteo Pasquinelli and Vladan Joler. *The Nooscope manifested: AI as instrument of knowledge extractivism*. AI & Society, 2020.

[11] Eric Sadin. *L'Intelligence artificielle ou l'enjeu du siècle. Anatomie d'un antihumanisme radical*. Éditions L'échappée, 2018.

Author's response: Relating to machines or communicating with machines?

Elena Esposito

I agree with Cabitza's insightful comments and thank him for the opportunity to clarify some misunderstandings that may arise from reading my text.

Two points seem to me central to his argument. The first is formulated in the very conclusion of his paper, when he argues that 'the problem is not how machines work, but *how we want to relate to them*'. In fact, my contribution is precisely an attempt to articulate this requirement, shifting the focus from the intelligence of machines (and the attempt to understand how they work) to the possibility of producing a condition of dialogue between users and algorithms (relating to them), that is compatible with the opacity of algorithms. This is the reason why today, dealing with machine learning techniques, the crucial issue is explanation – though not in order to create transparency but to enable users to autonomously make sense of the operations of the machines.

The second central point of Cabitza, argued at length, is related to this focus on the autonomy of the users: the move from an alethic stance to an anti-alethic stance, which questions the objectivity and neutrality of computational processes, especially those obtained with machine learning techniques. This is an inevitable shift, primarily in the social sciences, which has now been argued by many studies, including those cited by Cabitza. As he rightly observes, it is also the reason why explanations are so important.

In my opinion, regardless of the correctness of computational processes, and of their transparency, it is crucial to obtain from the explanations of algorithms clues that enable observers to autonomously use and control the results of the machines' operations – putting into practice their anti-alethic stance. Precisely this, as I argue in my paper, is the problem highlighted by the use of algorithmic procedures in legal decisions: even if the machine provides an explanation, the lack of ambiguity risks constraining the interpreters' space of freedom, and with it their autonomy with respect to the machine's processes. The problem of the 'excess' of explanation provided by machines seems to me to be exactly the one Cabitza warns us about, thus highlighting the need to distinguish between how the machines work and how we relate to them.