BIG DATA & SOCIETY

Commentary

Big Data & Society
January–March: 1–5

DOI: 10.1177/20539517241235871
journals.sagepub.com/home/bds
Sage

Prediction and explainability in AI: Striking a new balance?

Aviad Raz[1], Bert Heinrichs[2], Netta Avnoon[1], Gil Eyal[3] and Yael Inbar[4]

Abstract

The debate regarding prediction and explainability in artificial intelligence (AI) centers around the trade-off between achieving high-performance accurate models and the ability to understand and interpret the decisionmaking process of those models. In recent years, this debate has gained significant attention due to the increasing adoption of AI systems in various domains, including healthcare, finance, and criminal justice. While prediction and explainability are desirable goals in principle, the recent spread of high accuracy yet opaque machine learning (ML) algorithms has highlighted the trade-off between the two, marking this debate as an inter-disciplinary, inter-professional arena for negotiating expertise. There is no longer an agreement about what should be the "default" balance of prediction and explainability, with various positions reflecting claims for professional jurisdiction. Overall, there appears to be a growing schism between the regulatory and ethics-based call for explainability as a condition for trustworthy AI, and how it is being designed, assimilated, and negotiated. The impetus for writing this commentary comes from recent suggestions that explainability is overrated, including the argument that explainability is not guaranteed in human healthcare experts either. To shed light on this debate, its premises, and its recent twists, we provide an overview of key arguments representing different frames, focusing on AI in healthcare.

Keywords

Artificial intelligence – AI, explainable AI – XAI, healthcare, transparency, predictability, evidence-based medicine

The debate regarding prediction power and explainability in artificial intelligence (AI) centers on the trade-off between achieving high-performance models, and the ability to understand and interpret the decision-making process of those models. In recent years, this debate has gained significant attention due to the increasing adoption of AI systems in various domains, including healthcare, finance, and criminal justice (Amann et al., 2020). From a regulatory and AI ethics perspective, prediction power and explainability are mutually desirable goals, with explainability and opacity construed as opposites. Yet on the ground, for example in professional debates within computer science or between developers and expert users, the debate is framed as a trade-off between two mutually exclusive goals. This debate, we suggest, is also an arena for negotiating the relations between partially competing, partially collaborating professions, disciplines and forms of expertise. Viewing the debate from a sociological perspective highlights that the meaning of terms like explainability/opacity is not fixed, but is relative to the social positions and interests of the debating parties.

The impetus for writing this commentary comes from recent suggestions that explainability is actually overrated, including the argument that human experts, too, are unable or are not required to provide explanations (Kawamleh, 2022). This argument aims to turn the tables on regulators and ethicists who call for explainability as a condition for "trustworthy AI," but it has another, perhaps unintended, consequence. It opens up the debate to a sociological perspective. If what is required of AI should be

[1]Department of Sociology & Anthropology, Ben-Gurion University of the Negev, Beer-Sheba, Israel
[2]Institute of Neurosciences and Medicine: Brain and Behaviour (INM-7), Forschungszentrum Jülich, Germany
[3]Department of Sociology, Columbia Univerity, NY, USA
[4]Technology Management and Information Systems, Coller School of Management, Tel Aviv University, Israel

Corresponding author:
Aviad Raz, Department of Sociology & Anthropology, Ben-Gurion University of the Negev, Beer-Sheba, Israel.
Email: aviadraz@bgu.ac.il

symmetrical to what is required of human experts, then the sociological approach that understands experts as involved in struggles over jurisdiction and attribution becomes relevant and can shed light on the explainability vs. prediction debate.

We begin with a condensed presentation of the regulatory and ethical framing of the debate as positing two values that need to be balanced. We then offer an alternative sociological framing in which the umbrella of prediction power vs. explainability and the related debate on explainability vs. opacity are shown to be a space for various experts to assert tacit claims for jurisdiction and authority. We further illustrate this sociological framing with evidence from research on the integration of AI in healthcare.

Framing the debate(s)

Explainability, referred to as transparency/surveyability/interpretability/explicability, has become a major element of "algorithmic ethics" (Ananny, 2016), connected to trust/trustworthiness, accountability and liability, non-discrimination/fairness (Heinrichs, 2022), privacy, and autonomy (Goisauf and Cano Abadia, 2022). The main argument for regulating explainable AI (XAI) is that developers and market competition may prefer the modeling complexity and accuracy provided by black-box systems (Ali et al., 2023). In addition, highly predictive models may still incur errors and biases that may go unnoticed without explainability. Regulators thus argue that explainability/opacity should be addressed in the context of algorithmic auditing and accountability (Diakopoulos, 2015).

Nevertheless, XAI is not clearly regulated. The White House Blueprint for an AI Bill of Rights (2022) states that one should know how and why an outcome impacting one was determined by an automated system, emphasizing the importance of plain language and clarity. The FDA (2020) does not mention explainability in the context of AI-based medical devices, but requires transparency (clarity) of the functions of the software and its modifications. The "right to explanation" is formulated (although vaguely) in the GDPR and the proposed EU AI Act (article 13 on transparency and prevision to users), requiring that AI systems be explainable for high-risk decision making (EU, 2021). The EU's GDPR states that the data subject should have the right to obtain an explanation of the decision reached yet offers limited guidance on what constitutes a sufficient explanation. Also, in UNESCO's (2021) "Recommendation on the Ethics of AI," explainability is one of the core values.

As prediction power and explainability crystallize in the context of different professional jurisdictions, the debate takes various shapes highlighting the complexity of the social relations of expertise. The controversy may be understood by some participants to reflect the worldviews and professional interests of two distinct parties (though in reality there are a lot more than two). An example of this is the stats/ML debate. These two paradigms or "cultures" (Breiman, 2001) reflect, respectively, explanation-focused and prediction-focused approaches to science. Rather than trying to decipher the relations between variables and infer causality, as in statistical modeling, in prominent ML methods, the model learns different features of the input, compares them to a set of training outputs, with specific weights and inter-relations emerging in the process. Some ML proponents argue that the model's inner-workings opacity is a beneficial, even transhuman, feature of AI, as it allows more prediction power. Yet other developers argue for techniques of *interpretability*, such as layer-by-layer feature analysis, feature selection methods, feature maps, and prediction and bias metrics, that open a door into the black-box model and can also improve the process of debugging (Ali et al., 2023). Numerous methods for post-hoc approximation of explainability exist for black-box models.

The opacity of ML encounters new challenges when assimilated by expert users in other professional fields. Outside tech culture, such opacity can be seen as a threat to expert discretion. Arguably, human experts as well, and doctors specifically, reach their decisions drawing on tacit expertise that is often hard to articulate and explain, especially to laypeople. Nonetheless, in medicine explainability has well established conventions and "interactional expertise" (Collins, 2010) can be drawn upon to explicate doctors' reasoning. The "radical opacity" of ML thus presents a challenge for experts, undermining the long-standing rule-governed exchange of decisions within and between veteran professions, such as doctors and legal experts (Heinrichs/Knell, 2021). In contrast, in what follows we discuss examples of healthcare settings where doctors also belittle the importance of AI explainability.

The sociological frame we advocate highlights the entangled social characteristics of prediction vs. explainability in context. Their formulation in terms of an opposition may be itself part of the context (and the problem), preventing a more realistic treatment of the issues involved as dependent on real-world circumstances such as data complexity and business interests (Herm et al., 2023). This opposition sets aside the algorithm as working by itself, while in reality, algorithmic predictions are the product of a human-machine "companionship" (Borch and Hee Min, 2022) in which humans constantly "repair" (Collins, 2010) algorithmic results. We should therefore ask when, why and how, in the course of this interaction, does a lack of explainability become problematized, and repair attempts are being made?

The sociological frame also helps to realize how the argument of double standards (human experts too are not explainable, so why are we demanding algorithms to be explainable?) is actually a distraction (Maclure, 2020). It demonstrates that those who stand to benefit from the adoption of opaque ML tools may have an interest in

constructing the "radical inscrutability" of such tools as a feature rather than a bug, because such construction enables them to become the experts who speak for the unknowable machine.

Focusing on the professional dynamics

We now turn to describe representative arguments by expert users of AI and their ideas about the relevance of formal models to the "real world." The professional dynamics include the specific context/task, user requirements, and experts' weighing of ethical considerations, reflecting the enactment of the complicated relations between the experts in a given context and the specific algorithmic tool (Seaver, 2017). For example, in the field of predictive genetics called "polygenic risk score," by lowering the statistical standards for a marker as trait-associated, AI is used to trace associations by estimated effect sizes, and aggregate these associations over a large number of variants. In this manner, predictive accuracy may be increased at the expense of explainability, as any clear etiological link between specific genetic changes and the phenotype of interest is obscured. However, some genetics experts may justify this obscurity if it allows them to expand their jurisdiction in medical decision making (Raz and Minari, 2023).

Clinical practice under the contemporary paradigm of evidence-based medicine (EBM), as McCoy et al. (2022) argue, is privileged towards statistically sound evidence that an intervention does work, alongside a good enough story/theory that consistently makes sense of the output. Examples include mood stabilizers and other drugs that were found to be effective in randomized controlled trials, such as lithium or aspirin. The same argument has been made by London (2019), who emphasized that the absence of causal knowledge is already a common phenomenon in medical practices and that empirically validating the accuracy of AI in healthcare is significantly more important than explainability. Such real-world expressions that belittle explainability highlight, in our view, how a debate that purports to be general is more of a negotiation of professional jurisdictional boundaries, where certain experts claim or negotiate the legitimacy of being inscrutable. If "expertise" is, paradoxically, the "understanding of rules that cannot be expressed" (Collins and Evans, 2007: 17) then it practically entails the selective conversion of the cultural capital of non-explainability as it moves across professional fields.

Hybrid on-site expert strategies are thus emerging for negotiating partial explainability or converting explainability across professional/scientific fields. Recent use patterns in radiology, for example, demonstrate how explainability is increasingly tweaked around predictability. In image detection algorithms, usually Convolutional Neural Networks, their first layers will contain references to shading and edge detection. The human might never have to explicitly define an edge or a shadow, but because both are common among every photo, the features cluster as a single node and the algorithm ranks the node as significant to predicting the final result. The image detection model thus becomes more "explainable" to the expert radiologist. This reflects a pragmatic shift from explainability as an inherent characteristic of an algorithm, to explainability as an emergent and approximated property (Rudin, 2019).

In time-sensitive scenarios, such as emergency medical situations, prioritizing predictability can be crucial for prompt interventions. Studies on assimilating AI in stroke care show that performance is measured by reducing time-to-care while explainability is not considered (Hassan et al., 2020). However, our on-going study on assimilating AI in stroke care shows that for neurologists explainability or opacity are not the issue because they do not use AI primarily for diagnosis. Rather, they use the AI tool as a communication platform which reduces time-to-care by eliminating hierarchical workflow and enabling direct communication among all stroke team members. While the scans are uploaded into the AI system which flags high-risk cases, stroke team members continue to rely on their own expert judgment as the final call.

On the other hand, experts may decide not to engage, or partially engage, with high-performance AI that lacks explainability (Lebovitz et al., 2022), finding it challenging to accept a tool that both requires them to spend time on interpreting readings and may eventually 'kick them out'. A recent example of this, in addition to radiology, may be taking place in the remote sensing image analysis community where some experts are hesitant to engage with novel AI systems that offer great segmentation of images, even though they were developed on the basis of all kinds of scenes from other domains, not remote sensing, and have little explainability (Gevaert, 2022).

The way forward

There are several arguments for pursuing explainability in AI. While opaque yet highly accurate ML offers advantages, it can also lead to challenges related to users' trust, bias detection, and public acceptance. Medical and legal institutions as well as doctor-patient interactional relations depend on conventions of explaining decision-making, which are disrupted by opaque AI.

Although central to the perspectives of regulation and AI ethics, arguments in favor of explainable AI as a prerequisite for user autonomy and trust can be criticized on empirical and philosophical grounds (Durán and Jongsma, 2021). Sometimes we trust "the experts" even when they are not providing clear explanations. Nevertheless, the modern, secularized premise of trust requires matched expectations and potential explicability, as a contrast to "blind" faith. We trust – to a certain extent - the algorithm

because we have a degree of confidence that we can understand - in a practical sense - why it reached its predictions, or we trust the expert who developed it. A social-professional struggle is going on concerning whether we can trust ML algorithms in medical decision making. Which expertise – doctors vs. engineers– will win the public's trust depends on many factors, explainability being only one of them. And when the experts settle their debate, we should expect the rise of another debate concerning how the required explainability should be made accountable to consumers (Neyland, 2019).

If explanations and risks are socially constructed, then so is liability. The recent debate around the EU AI Act regarding AI categorization as "low," "middle" or "high" risk should take into consideration that such risk designations are quickly changing. Who is to blame if an opaque algorithm makes mistakes – the company that made the AI tool, the engineers, the doctor, or the hospital management? Does reliance on AI absolve health professionals from responsibility, or on the contrary, require them to learn new explanatory skills? Moreover, the law does not tolerate a vacuum of responsibility, and will likely attribute responsibility to some human agent, such as the developers (Martin, 2019). Making algorithms accountable also takes place within human-machine interactive companionship (Borch and Hee Min, 2022) or 'distributed agency' between humans and programs (Rammert, 2008) – where a relational justification of explainability (Coeckelbergh, 2020) is aimed at by giving the algorithms qualities that make them legible to groups of people in specific contexts.

The main lesson of our focus on the professional dynamics is that the justification for explainability should not be looked for solely in the general terms of AI regulation, in principles of AI ethics, or in computer science discussions. Rather, explainability is important because it promotes the discussion (and sometimes, resolution) of disagreements between experts. Explainability provides the context, transparency, and insights to support decision-making, enhances trust, and addresses ethical considerations. Striking a balance between predictive accuracy and explainability requires interdisciplinary efforts addressing the multiple frames of this debate, and engaging computer scientists, ethicists, sociologists, policymakers, and domain experts such as healthcare professionals. This debate should not be dominated by groups that exclusively pursue their own interests.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

Aviad Raz and Yael Inbar received funding from the Israel Science Foundation grant #1325/23 "Assimilating AI in Radiology: Expert Users' Perceptions and Activities in Socioethical and Institutional Perspective".

ORCID iDs

Aviad Raz https://orcid.org/0000-0001-6268-0409
Bert Heinrichs https://orcid.org/0000-0002-0181-0078
Netta Avnoon https://orcid.org/0000-0002-3758-2097
Gil Eyal https://orcid.org/0000-0001-7194-3864
Yael Inbar https://orcid.org/0000-0001-8319-627X

References

Ali S, Abuhmed T, El-Sappagh S, et al. (2023) Explainable Artificial Intelligence (XAI), Information Fusion. https://doi.org/10.1016/j.inffus.2023.101805.

Amann J, Blasimme A, Vayena E, et al. (2020) Explainability for artificial intelligence in healthcare. *BMC Medical Informatics and Decision Making* 20: 310.

Ananny M (2016) Toward an ethics of algorithms. *Science, Technology, and Human Values* 41(1): 93–117.

Borch C and 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): 20539517221111361.

Breiman L (2001) Statistical modeling: The two cultures. *Statistical Science* 16(3): 199–231.

Coeckelbergh M (2020) AI, responsibility attribution, and a relational justification of explainability. *Science and Engineering Ethics* 26: 2051–2068.

Collins H (2010) *Tacit and Explicit Knowledge*. Chicago: University of Chicago Press.

Collins H and Evans R (2007) *Rethinking Expertise*. Chicago: University of Chicago Press.

Diakopoulos N (2015) Algorithmic accountability: Journalistic investigation of computational power structures. *Digital Journalism* 3(3): 398–415.

Durán JM and Jongsma KR (2021) Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics* 47: 329–335.

EU (2021) ARTIFICIAL INTELLIGENCE ACT. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206.

FDA (2020) Proposed regulatory framework for modifications to AI/ML-based Software as a Medical Device (SaMD). https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf.

Gevaert CM (2022) Explainable AI for earth observation: A review including societal and regulatory perspectives. *International Journal of Applied Earth Observation and Geoinformation* 112: 102869.

Goisauf M and Cano Abadía M (2022) Ethics of AI in radiology. *Front. Big Data* 5: 850383.

Hassan A, et al. (2020) Early experience utilizing artificial intelligence shows significant reduction in transfer times and length of stay in a hub and spoke model. *Interventional Neuroradiology* 26(5): 615–622.

Heinrichs B (2022) Discrimination in the age of AI. *AI & Society* 37: 143–154.

Heinrichs B and Knell S (2021) Aliens in the space of reasons? On the interaction between humans and AI agents. *Philos. Technol* 34: 1569–1580.

Herm LV, Heinrich K, Wanner J, et al. (2023) Stop ordering machine learning algorithms by their explainability!. *International Journal of Information Management* 69: 102538.

Kawamleh S (2022) Against explainability requirements for ethical artificial intelligence in health care. *AI Ethics*. https://doi.org/10.1007/s43681-022-00212-1.

Lebovitz S, Lifshitz-Assaf H and Levina N (2022) To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. *Organization Science* 33(1): 126–148.

London AJ (2019) AI And black-box medical decisions: Accuracy versus explainability. *Hast Cent Rep* 49(1): 15–21.

McCoy L, Brenna C, Chen S, et al. (2022) Believing in black boxes: ML for healthcare does not need explainability to be evidence-based. *Journal of Clinical Epidemiology* 142: 252–257.

Maclure J (2020) The new AI spring: A deflationary view. *AI & Soc* 35: 747–750.

Martin K (2019) Ethical implications and accountability of algorithms. *Journal of Business Ethics* 160: 835–850.

Neyland D (2019) *The Everyday Life of an Algorithm*. Gewerbestrasse, Switzerland: Palgrave-MacMillan.

Rammert W (2008) Where the action is: Distributed agency between machines, humans and programs. In: Seifert, et al. (eds) *Paradoxes of Interactivity*. New Brunswick: New Jersey: Transaction Books, 63–91.

Raz A and Minari J (2023) AI-driven risk scores: Should social scoring and polygenic scores based on ethnicity be equally prohibited? *Frontiers in Genetics* 14: 1169580.

Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 1: 206–215.

Seaver N (2017) Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society* 4(2).

UNESCO (2021) Recommendation on the Ethics of Artificial Intelligence. https://unesdoc.unesco.org/ark:/48223/pf0000380455