Artificial Intelligence and Law
https://doi.org/10.1007/s10506-023-09356-9

ORIGINAL RESEARCH

The black box problem revisited. Real and imaginary challenges for automated legal decision making

Bartosz Brożek[1,2] · Michał Furman[2] · Marek Jakubiec[1] · Bartłomiej Kucharzyk[1]

Accepted: 23 March 2023
© The Author(s) 2023

Abstract

This paper addresses the black-box problem in artificial intelligence (AI), and the related problem of explainability of AI in the legal context. We argue, first, that the black box problem is, in fact, a superficial one as it results from an overlap of four different – albeit interconnected – issues: the opacity problem, the strangeness problem, the unpredictability problem, and the justification problem. Thus, we propose a framework for discussing both the black box problem and the explainability of AI. We argue further that contrary to often defended claims the opacity issue is not a genuine problem. We also dismiss the justification problem. Further, we describe the tensions involved in the strangeness and unpredictability problems and suggest some ways to alleviate them.

Keywords Black box problem · Explainable AI · AI and law · Legal decision-making · Automated decision-making

1 Introduction

One of the most pressing problems related to the use of AI in the decision-making processes is the so-called black box problem (Castelvecchi 2016; Rudin 2019). Various AI tools, especially those based on the machine learning mechanism, are designed to analyze huge sets of data, find patterns 'hidden' therein, and offer a solution (e.g., a decision to a legal case, a medical course of action, granting a loan, etc.).

The problem is that – for various reasons – we often do not know *how* or *why* the algorithm got to the proposed solution. This may obviously be problematic. Imagine

✉ Bartosz Brożek
bartosz.brozek@uj.edu.pl

[1] Faculty of Law and Administration, Jagiellonian University, Krakow, Poland

[2] Copernicus Center for Interdisciplinary Studies, Jagiellonian University, Krakow, Poland

Published online: 04 April 2023

we decided to use AI to adjudicate fairly simple legal cases in order to reduce judge's caseload and speed up legal proceedings. The question is, whether we would be content in accepting such judgment 'no matter what', or rather we would require at least the basic knowledge of how the algorithm arrives at its decisions. Should we allow AI algorithms to be 'black boxes', or should we rather have the ability to look into the boxes to understand what happens?

In the AI scholarship the black-box problem is often addressed under the label of explainability[1] (Guidotti et al. 2018; Miller 2019; Arrieta et al. 2020; Vilone and Longo 2021). In general, an AI system (e.g., a machine learning model) may be considered explainable if it can be understood by humans through an external, simplified representation (though explanatory value of internal components of the models is also discussed – see Jain and Wallace 2019; Wiegreffe and Pinter 2019). The issue is frequently tackled from two different perspectives – technical (IT) and legal – between which there seems to be a significant gap. Law itself does not define explainability; it does however provide rules, mostly requirements, which in the literature are linked to the issue of explainable automated decision making (Doshi-Velez et al. 2017, Wachter et al. 2017a, Bibal et al. 2021). Those requirements are not homogenous:

- some concern the private sector, other apply to the public administration and judiciary,
- some are specific to the automated decision making, other relate to decision making in general,
- (maybe most importantly) some pertain to the AI systems as a whole, others to the instances of decisions.

The private automated decision making with regard to explainability is regulated by the growing, mostly with successive EU acts, body of specific rules (see e.g. art. 13.2(f) and 14.2(g) of the GDPR and amended art. 6(a) of Directive 2011/83 on Consumer Rights). Even those rules do not, however, contain a unified concept of explainability. According to Bibal et al. (2021) they may be interpreted in technical terms and categorized as requirements for revealing of:

- the main data features used in the model or in a decision,
- all data features used in a decision,
- the combination of data features used in a decision,
- the whole model.

On the other hand, automated administrative and judicial decision making, if at all permissible, seems to be bound by the general rules adopted in administrative and court procedures. In general, the decisions should be motivated (justified) with reasons: facts, legal provisions and, in case of judicial decisions, reasoning related to the

[1] It is also called interpretability, though some authors differentiate between those terms. For example, according to Rudin (2019) interpretability is an inherent feature of a model, whereas explainability entails post hoc explanations for black box models incomprehensible to humans (see also Lipton 2018).

arguments of the litigation parties. From the technical perspective, the system should be able to provide not only the decision, but also the relevant legal rules and proper arguments (Atkinson et al. 2020; Bibal et al. 2021).

Such considerations have also led to the development of the explainable AI movement (XAI) (Gunning and Aha 2019). It is based on the assumption that AI algorithms already perform tasks of such importance that they should be 'white boxes', transparent not only to their creators but also to end-users (or anyone who may be affected by the algorithm's decision). In this context, the US National Institute of Standards and Technology developed four principles of XAI:

> (1) The system should be able to explain its output and provide supporting evidence (at least).
> (2) The given explanation has to be meaningful, enabling users to complete their tasks. If there is a range of users with diverse skill sets, the system needs to provide several explanations catering to the available user groups.
> (3) This explanation needs to be clear and accurate, which is different from output accuracy.
> (4) The system has to operate within its designed knowledge limits to ensure a reasonable outcome (Phillips et al. 2021).

A similar idea is sometimes proposed in the context of the laws of robotics, which are easily transferable to the general AI research, when it is suggested that in addition to the famous three laws identified by Asimov a fourth criterion should be added – the explainability of AI actions (Murphy and Woods 2009, Wachter et al. 2017b).

Are these proposals (legal regulations, XAI movement, etc.) the right way to proceed? Are AI algorithms really black boxes? And even if they are – do we really need to worry about it? In this short essay we would like to suggest that the black box problem is, in fact, a superficial one as it results from an overlap of four different – albeit interconnected – issues: the opacity problem, the strangeness problem, the unpredictability problem, and the justification problem. Let us consider all those issues one by one.

2 Minds as blackboxes

Imagine one has designed a sophisticated algorithm which aids judges to determine the optimal sentence for a repeat offender. The algorithm is fed with data pertaining to the offender's personal life, the circumstances of their first as well as subsequent offenses, but also with a huge database on other repeat offenders: what were the circumstances of their actions, what were their sentences, how they behaved in prison, what was their life after they were released, etc. The size of the database is such that no man would be able to get acquainted with it, not to mention analyze it, within their lifespan. Meanwhile, the algorithm, run on a fast computer, needs only seconds to come up with a verdict. How should we perceive such verdicts? Do we know what happens when the algorithm does its magic?

The answer is *yes* and *no*. *Yes*, because we have *designed* the algorithm in such a way that it looks for patterns in the huge database. The pattern may be, for example, that repeat offenders who are highly intelligent and have no permanent employment are more likely to keep breaking the law, therefore a longer sentence (and hence a longer isolation) is called for in their case. What we know is that the algorithm *looks for such patterns*. What we *do not know* is what the pattern on which the algorithm ‘based its decision’ is, and what exactly were the steps that led the algorithm to such a conclusion.

Is this problematic? At the surface, it seems like a very bad way to make important decisions. In public life, and in the law particularly so, we strive for transparency, and there seems to be no transparency in the machine learning ‘magic’. But let us compare our algorithm with a real judge, who makes a decision in a similar case. Do we *really know* what is going on in the judge’s head? Can we be sure what is the pattern they base their decision on? The last decades of research in experimental psychology and neuroscience suggest a plain answer to this question, and the answer is ‘no’ (Bargh and Morsella 2008, Guthrie et al. 2001, Brożek 2020). The way people make most – if not all – of their decisions is unconscious. In our decision-making, there are no clearly identifiable ‘steps’, which we are aware of and can control. Usually, the decision simply appears in our minds, ‘as if from nowhere’ (Damasio 2006).

The interesting thing is that – in experimental psychology and neuroscience – what we are trying to figure out is *what is the mechanism* of the unconscious decision-making. Thus, in relation to the functioning of the human mind we are looking for the kind of knowledge we *already have* in the case of any AI algorithm. For example, it is very likely that the human mind is a powerful pattern-finder. What we are interested to learn is *how such patterns are found*, what is the pattern-detection mechanism, and not what patterns are *actually* found, how do they look, or why the mind has based its decision on this rather than a different pattern.

From this perspective, the human mind is much more of a black box than the most sophisticated machine learning algorithm. For the algorithms, we at least know *how* they work, even if we cannot explain why they have arrived at a particular decision. In the case of the human mind, we have only a tentative outline of the answer to the question how it works (Bonezzi et al. 2022).

Our inability to understand exactly what an AI algorithm does is sometimes referred to as the opacity problem (Burrell 2016; Zednik 2021). However – when compared with our knowledge and understanding of the way the human mind works – the algorithms are not really that opaque. The opacity problem does not seem to be a genuine issue. At the same time, we do not question the decisions humans make, or at least not in the way we put into doubt the decisions made by AI. We do not treat minds as black boxes, even if they seem to be black boxes *par excellence*. Why is it so?

3 Stranger things

A simple answer to the question posed at the end of the previous section is that we do not consider our minds ‘black boxes’ because we are familiar with them. AI algorithms, and in particular machine learning algorithms, seem like black boxes, because they are unfamiliar: they are ‘strange things’ we are not yet accustomed to. The operative word here is ‘yet’. Ever since the beginning of human civilization, we have created many artifacts which – initially strange to us – have become familiar companions in our daily lives. Writing, print, steam engines, electricity, automobiles, radio, television, computers, mobile phones – they all once were sources of fear and awe, mysterious black boxes, but, with time, we have got accustomed to them. The same may be true of AI algorithms providing us with practical (legal, medical, technical) decisions: give us some time together and we will get familiar with them.

This answer – that we consider AI algorithms ‘black boxes’ because they are ‘stranger’ than other things – may be true, but it is at the same time somewhat shallow. We believe that there is a deeper reason for human beings to be perfectly happy with the decision-making processes of the human mind, while feeling uneasy when letting AI algorithms decide. In order to understand it, we need to spare a few words on folk psychology.

Folk psychology is the ability of mind-reading, i.e., of ascribing mental states to other people. A more detailed characterisation – albeit not an incontestable one – has it that folk psychology is a set of the fundamental capacities which enable us to *describe* our behavior and the behavior of others, to *explain* the behavior of others, to *predict* and *anticipate* their behavior, and to produce *generalizations* pertaining to human behavior. Those abilities manifest themselves in what may be called *the phenomenological level* of folk psychology as “a rich conceptual repertoire which [normal human adults] deploy to explain, predict and describe the actions of one another and, perhaps, members of closely related species also. (…) The conceptual repertoire constituting folk psychology includes, predominantly, the concepts of belief and desire and their kin – intention, hope, fear, and the rest – the so-called propositional attitudes” (Davies and Stone 1995).

One can also speak of *the architectural level* of folk psychology which consists of the neuronal and/or cognitive mechanisms which enable ascribing mental states to others. Importantly, this level is not fully transparent or directly accessible to our minds – while we are able to easily describe the conceptual categories we use to account for other people’s behavior (at the phenomenological level), we usually have no direct insight into the mechanisms behind mind-reading (Brożek and Kurek 2018).

It follows that we understand and explain behavior – including decision-making – not as it happens, but as seen through the lenses of the folk-psychological conceptual scheme. Moreover, we are in principle not aware that this interpretive mechanism is at work, since the architectural level of folk psychology is not something we may observe. This fact explains *why* we have no problem in accepting decisions made by other people – even if actually their minds are *black boxes* to us. We do not see it that way, because the conceptual apparatus of folk psychology makes us interpret the behavior of others as an outcome of a decision-making process which *seems* to be transparent and perfectly understandable. At the same time, we have a problem accepting a decision made by an AI algorithm, because this is *not* the way decisions are made, at least from the point of view of folk psychology.

Thus, our thesis is that the opacity problem is *not* the real problem with AI algorithms. The real issue lies somewhere else and may be deemed *the strangeness problem*. Moreover, the strangeness in question is not superficial – it will not dissolve once we get accustomed to the AI algorithms making decisions for us. Such algorithms are different from cars, airplanes and mobile phones, because they seem to be doing what – according to folk psychological conceptual scheme – only real people, equipped with rational minds and free will, can do.

Is it possible to overcome this difficulty? It is extremely difficult to answer this question. On the one hand, the research in psychology and anthropology shows that the folk psychological conceptual scheme is a cultural creation – it differs from culture to culture (Lillard 1998). For example, the very concept of agency and decision-making is different in the Western culture and Eastern cultures or the cultures of indigenous peoples of the Amazon (Morris and Peng 1994). From this perspective, it may be possible for the folk psychological conceptual scheme to evolve with time into something different, e.g., a framework which accepts AI algorithms as capable of decision-making.

On the other hand, at least some mechanisms behind the folk-psychology seem to be inborn. In particular, as suggested by research in the developmental psychology, it seems that folk psychology is deeply rooted in the human ability to spontaneously distinguish between two kinds of interactions (causality) in the world – physical and intentional (Bloom 2004). We perceive the interactions between physical objects as goversned by a different set of laws than the intentional actions of other people. Only in the latter case one can speak of genuine decision-making processes. The question is, therefore, whether AI algorithms can fall into this second category – can we perceive them as intentional?

It seems that the answer is negative as long as we *do* understand how the algorithm functions. Historically, humans attributed intentionality to physical objects – stones, trees or rivers (Hutchison 2014). Even today, we have a tendency to *anthropomorphize* inanimate objects (e.g., robots which perform some mundane tasks). However, given our current worldview, once we *know* that a vital decision (e.g., to a legal case or pertaining to the medical treatment) is generated by a ‘soulless’ algorithm, such anthropomorphic attributions may not be possible. They would require a major shift in our worldview, which – given its nature – is difficult if not impossible to imagine. Thus, it does not seem likely that we will perceive AI algorithms as making genuine decisions, especially the more vital ones. The strangeness problem is an enduring one.

4 Cognitive safety

Let us imagine now that someone has developed an extremely complex AI algorithm based on deep machine learning which has one goal: to provide an answer to the question in the form ‘what is the sum of $x+y$’, where x and y are natural numbers in the range from 1 to 100. The algorithm uses a huge dataset and is highly accurate – in fact, it has been used several million times and has always given a correct answer. Disregarding the fact that there are much less complex computational ways for adding natural numbers, we would probably never question the algorithm. The reason is that it gives answers which are *expected*. What is expected gives us no headache.

The human mind is a wonderful mechanism, capable of maneuvering in a highly complex and unpredictable world. Because of this complexity and unpredictability, the mind naturally gravitates towards often used and well-tested behavioral patterns and previously accepted beliefs. It is somewhat cognitively rigid, or to put it differently, it is a cognitive conservative (Kruglanski 1989; Webster and Kruglanski 1994; Brożek 2020). Revolutions in our individual cognitive spheres as well as in our culture do not happen too often and take some time to exercise real influence on what we think and do.

One important lesson which comes from psychology and the cognitive sciences is that the human mind strives for certainty (Kruglanski 1989). This need is deeply rooted in us by the evolutionary processes and manifests itself, *inter alia*, in our emotional mechanisms (Kruglanski 1989; Brożek 2020). In the recent years much attention has been paid to the so-called epistemic emotions (Gopnik 2000). Contradiction or some other inconsistency in our experience – be it Einstein's uneasiness that the observed motion of Mercury minimally deviates from the predictions of Newton's theory, or the feeling that 'something would be wrong' if we just went for holiday ignoring the fact that our uncle is terminally ill, always generates an emotion: of curiosity, disorientation or even anxiety. It motivates us to seek an explanation where the cognitive dissonance comes from. The feeling that 'something is wrong' is the main driving force behind Einstein's search for the general theory of relativity; but it is the same force which makes us skip vacation in face of a serious illness in our family. Without epistemic emotions there would be no discoveries, whether spectacular or small. The reduction of anxiety and disorientation, satisfying one's curiosity, and sometimes amusement or revelation, are the rewards we get for making our worldview more coherent (Hurley et al. 2011).

One would be mistaken, however, claiming that emotions have only positive influence on our cognitive processes, motivating us to search for better answers to the questions we pose operating in the physical and social environment. Emotions can also significantly disturb the thinking process – not only in extreme cases, where strong emotions like fear enter the stage, but also in quite ordinary decision processes. For example, it is difficult to 'work' for a longer time with two alternative and mutually inconsistent hypotheses. The mind has an inclination to quickly settle such a conflict rather that analyze the consequences of each hypothesis and systematically compare them. It is connected to the fact that – as we have observed above – our emotions drive us to certainty and reward us for it. It is difficult to accept that we base our beliefs and actions on conjectures rather than on solid and unshakable foundations. It is easier to believe that we have reached certainty even if objectively we are far from it (Kruglanski 1989).

This drive for cognitive safety makes it difficult for us to accept outcomes of a decision-making process which are unexpected. It doesn't matter who or what is making the decisions: the mere fact that the outcome is unexpected makes us uneasy.

For the same reason, a decision-making mechanism – be it a human being, an AI algorithm or an oracle – which produces expected outcomes is much easier to accept.

This is problematic in the context of our discussion for the following reason: the role of AI algorithms is not only to *replace us* in some cognitive tasks such as making a medical diagnosis or delivering a legal decision. Given that AI algorithms – and in particular machine learning – are capable of analyzing huge datasets in ways far exceeding the abilities of the human mind, our hope is that the algorithms *will* produce better outcomes than humans are capable of. However, this means that these outcomes will be unexpected.

In this way we arrive at the tension inherent in what we may call the *unpredictability problem*: we do not welcome surprises, while this is exactly what the AI algorithms are made for.

5 Justify me

In the law – and, more generally, in the social life – we expect decisions to be justified or at least justifiable. The typical perception of how lawyers – and, in particular, judges – operate is based on three tenets:

(1) Legal reasoning has a clearly identifiable structure.
(2) Legal reasoning consists in carrying out operations on sentences (beliefs) in an algorithmic way.
(3) Legal reasoning is based – in one way or another – on the rules of classical logic. As a consequence, legal reasoning aims at providing a solution to a legal case which is justified (rational) (Wróblewski 1992; Alexy 2009; Stelmach and Brożek 2006, Hage 2005).

Meanwhile, the research in cognitive science shows that the actual processes of practical reasoning are a far cry from such an ideal model. Although there is no one single, commonly accepted theory of actual legal thinking, the existing approaches seem to share (to a greater or lesser degree) the following assumptions (Brożek 2020; Brożek et al. 2021):

(1) Most (if not all) legal decisions are made in a way which has no identifiable structure nor consists of algorithmic steps. The decisions appear in one’s mind as if from nowhere.
(2) Most, if not all, decisions are made in (a) an *unconscious* way, where the unconscious processes are largely an effect of (b) social *training* and are based on (c) *emotional* reactions.
(3) In practical decision-making, reason (rational argumentation) has a secondary role. It either serves as an *ex post factum* rationalization of the decisions made (to defend those decisions against the criticism of others) or, in the best case, it has an indirect or otherwise hugely limited influence on the decision-making process (Haidt 2001).

These facts underline another difficult tension: between the way we perceive rationality and justification, and the way in which justification is usually produced. The classical stance is that conscious, rational deliberation is what *precedes* the decision (Kant 1909). From this perspective, it is quite understandable that an outcome of the work of an AI algorithm, which cannot be traced back and repeated, and hence remains 'mysterious', cannot be treated as rational or justified. In other words, the decision reached by the algorithm does not meet our standards of justification – at least as long as *the way* of reaching the decision is considered constitutive of its justificatory power.

However, a different approach – the one that takes seriously the actual mechanisms of (human) decision-making – opens the way for a different understanding of rationality and justification. It is *not* important how the decision was reached; the only question is whether the decision can be defended (justified) *ex post*. From this perspective, decisions made by AI algorithms can be rational, when an appropriate (acceptable) justification can be adduced in their favor.

In fact, this perspective paves the way for a reconceptualisation of how algorithms for legal decision-making (or for aiding legal decision-making) should be structured. The general idea is to 'mimic' the behavior of the human mind. The envisaged system would consist of two components or modules: the 'intuitive' and the 'rational'. The intuitive module would enable the system to learn from experience (i.e., large datasets) what are the patterns connecting types of legal problems with the corresponding legal decisions. For such an architecture, some machine learning seems to be the best option. The rational module, in turn, would be based on the existing (mainstream) models in AI & Law, i.e. it would be based on the use of some appropriate logical system (e.g., a kind of defeasible logic). However, the goal of the module would be different than usually assumed in the AI & Law literature: instead of *producing* a legal decision in the case at hand, it would aim at *justifying* a decision reached by the intuitive module. Thus, the rational module would 'work backwards': given a decision (based on the knowledge accumulated in the datasets and 'uncovered' by machine learning algorithms), it would search for a proper justification for it (this is similar to the idea of post-hoc explanations; however, the goal of the rational module would be to look for justifications, not explanations). Such a procedure is not mysterious; in fact, this is what the original ancient Greek meaning of 'analysis' is. As the great mathematician Pappus put it:

> For in analysis we suppose that which is sought to be already done, and we inquire from what it results, and again what is the antecedent of the latter, until we on our backward way light upon something already known and being first in order. And we call such a method analysis, as being a solution backwards (*anapalin lysin*) (quoted after Hintikka and Remes 1974).

The proposed architecture requires one more element. If the process of constructing a justification for the 'intuitive solution' cannot be completed (e.g., there appears a contradiction), one cannot simply accept the intuitive decision. What is needed is a kind of 'feedback loop' between the intuitive and the rational components (in this, the proposed solution further differs from the ideas pertaining to post-hoc explanations).

One can envisage it in various ways. In particular, it may function as a veto (if the intuitive decision cannot be justified, it is simply rejected and the intuitive module is activated to search for another solution), or it may be a more constructive mechanism (i.e., some modifications to the intuitive solution are introduced and tested in a back-and-forth procedure between the intuitive and the rational modules).

The soundness of the computational architecture outlined above notwithstanding, the moral of our considerations is that the justification problem in relation to the decisions made by AI algorithms remains a genuine one as long as we claim – for example after Kant – that the *way* the decision is made generates the justificatory power of that decision. Once this requirement is abandoned, and the method of the ex-post justification is allowed, the decisions of AI algorithms may be rendered rational; moreover, it is possible to construct the computational system in such a way that it takes advantage of the incredible power of pattern-finding in large datasets, while at the same time providing us with a justification for the decision made.

6 Conclusion & perspectives

Let us repeat the question which propelled our analysis: is there really a black box problem in relation to the AI algorithms? We believe that this question involves four different, although interconnected issues: the opacity, the strangeness, the unpredictability and the justification problems. Our analysis suggests that – contrary to the often expressed opinion – opacity problem is not significant. In fact, we do understand and can explain the operations of AI algorithms in a much better and more complete way than the functioning of the human mind. However, there is an additional problem here. The algorithms involved may be quite sophisticated so that only well-trained specialists may fully grasp their mechanics. From this perspective, the XAI and related postulates seem reasonable: the algorithm user (or someone whose legal or economic interest may be influenced by the decision made by the algorithm) should have access to an understandable (even if simplified) description of the functioning of the algorithm. If this is not the case, at least for some users the algorithms would remain 'black boxes'.

Similarly, the justification problem discussed above does not seem to be a genuine one as long as we do not consider justification to be generated by the *way* in which a decision is made. This is a crucial observation. As we have seen, in the short discussion pertaining to the explainability of AI, it is sometimes claimed that an explainable algorithm for solving legal issues should provide us with *reasons* for the given decision. Once we admit that justification may be constructed *ex post*, this requirement can be met.

The other two issues – of strangeness and unpredictability – are more problematic. The unpredictability of the decisions made by AI algorithms is what generates our distrust in them in the first place; however, it also represents what is really powerful in machine learning and related methods. They were designed to find patterns in datasets which cannot be analyzed by humans with their limited computational capacities. There is a genuine tension here. Fortunately, it does not seem to affect the question of explainability of AI as relevant to law. When we address properly the other problems – of opacity by providing (a simplified and) understandable description of the algorithm, and of justification by generating an acceptable *ex post* justification for the decision – even an unpredictable decision may be acceptable: ultimately, in such a case we would know *what happened* (i.e., how algorithm works) and that the outcome is justified.

The strangeness problem is also troublesome. It seems that – given our folk psychological conceptual apparatus – we cannot treat AI algorithms as genuine decision-makers. But this creates a kind of cognitive tension which is difficult to alleviate. There will always be a sense of strangeness and the transpiring need for a better understanding and justification of what AI algorithms do. Therefore, the postulates of the XAI movement as well as other suggestions pertaining to explainability of AI in the context of law are relevant. Our insight is, however, that the need for them does not arise from the opacity of AI algorithms but rather from their strangeness.

We believe that our observations provide a new perspective on the discussions currently taking place in the AI&Law literature and pertaining to XAI. Arguably, most of the issues dealt with within the field of XAI & law stem from two fundamental questions:

1. When, if ever, should explainability be required by the law?
2. What kind of explanation would be optimal from the legal perspective?

In the contemporary literature it is by far easier to find answers to the second question. As Bilal et al. (2021) point out, the general opposition is between explaining the mathematics of models and providing explanations that make sense to humans. The AI & law researchers seem to prefer the latter solution (Pasquale 2017; Selbst and Barocas 2018; Mittelstadt et al. 2019; Hacker et al. 2020) and offer its variants (Ye et al. 2018; Prakken 2020; Prakken and Ratsma 2021). This line of research also includes numerous papers interpreting the existing (or, as in the case of EU AI Act, pending) legal requirements on explainability, criticizing them or proposing new mechanisms and provisions (Goodman and Flaxman 2017; Malgieri and Comandé 2017; Wachter et al. 2017a; Selbst and Powles 2018; Casey et al. 2019; Zuiderveen Borgesius 2020; Grochowski et al. 2021; Kaminski 2021; Hacker and Passoth 2022; Sovrano et al. 2022). It is worth noting that *de lege lata* objections state that the rules are vague, too weak or incompatible with AI conceptual grid rather than unnecessary.

On the other hand, the first question (on the need for explainability in law) is tackled less often and mostly indirectly. Most prominently, Rudin (2019) advocates replacing explainable black boxes with inherently interpretable models, at least in the case of high-stakes decisions. It may be seen as abandoning one requirement (explainability) in favor of another, arguably a stricter one (interpretability). Similar concerns have recently been raised in the context of non-discrimination law by Vale et al. (2022).

The account we presented above suggests that more balance is needed between the two issues. We agree with Rudin that explainability may not be the holy grail of AI & law. However, our premise is different: explainability is valuable, but does not have to be required to a greater extent than in the case of human decision making (at least when black boxes perform no worse than humans). Strangeness is more problematic than opacity, hence the criteria for explanation (justification) should be rather psychological (understanding, trust) than technical. Moreover, one should not dismiss various approaches to providing explanations without testing them (see Prakken and Ratsma 2021) in different legal contexts, as the needs and risks differ between sectors. The same applies to the development of legal rules on explainability (see Zuiderveen Borgesius 2020) – we should shape the law on the basis of existing problems, not adjust the problems to the law.

Declarations

Statements and declarations This article has been supported by the Polish National Agency for Academic Exchange (NAWA) under the Grant "Digitalized! Society in the era of digital revolution", contract no. PPI/APM/2019/1/00016/U/00001".

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Alexy R (2009) A theory of legal argumentation. Oxford University Press, Oxford

Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, Herrera F (2020) Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inform Fusion 58:82–115

Atkinson K, Bench-Capon T, Bollegala D (2020) Explanation in AI and law: past, present and future. Artif Intell 289:103387

Bargh JA, Morsella E (2008) The unconscious mind. Perspect Psychol Sci 3(1):73–79

Bibal A, Lognoul M, De Streel A, Frénay B (2021) Legal requirements on explainability in machine learning. Artif Intell Law 29(2):149–169

Bloom P (2004) Descartes' Baby: how the Science of Child Development explains what makes us human. Basic Books, New York

Bonezzi A, Ostinelli M, Melzner J (2022) The human black-box: the illusion of understanding human better than algorithmic decision-making. J Exp Psychol Gen 151(9):2250–2258. https://doi.org/10.1037/xge0001181

Brożek B (2020) The legal mind: a new introduction to legal epistemology. Cambridge University Press, Cambridge

Brożek B, Hage J, Vincent N (eds) (2021) Law and mind: a Survey of Law and the Cognitive Sciences. Cambridge University Press, Cambridge. https://doi.org/10.1017/9781108623056

Brożek B, Kurek Ł (2018) Folk psychology and explanation. In: Brożek B et al (eds) Explaining the mind. Copernicus Center Press, Kraków, pp 149–170

Burrell J (2016) How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data & Society Jan-Jun 2016, 1–12

Casey B, Farhangi A, Vogl R (2019) Rethinking Explainable Machines. Berkeley Technol Law J 34(1):143–188

Castelvecchi D (2016) Can we open the black box of AI? Nature 538:20–23. https://doi.org/10.1038/538020a

Damasio A (2006) Descartes' Error. London, Vintage

Davies M, Stone T (1995) Folk psychology: the theory of mind debate. Blackwell, Oxford

Doshi-Velez F, Kortz M, Budish R, Bavitz C, Gershman S, O'Brien D, Wood A (2017) Accountability of AI under the law: The role of explanation. arXiv preprint arXiv:1711.01134

Goodman B, Flaxman S (2017) European Union regulations on algorithmic decision-making and a "right to explanation". AI magazine 38(3):50–57

Gopnik A (2000) Explanation as orgasm and the drive for causal understanding. In: Keil F, Wilson R (eds) Cognition and explanation. MIT Press, Cambridge, MA, pp 299–324

Grochowski M, Jabłonowska A, Lagioia F, Sartor G (2021) Algorithmic transparency and explainability for EU consumer protection: unwrapping the regulatory premises. Crit Anal Law (CAL) 8(1):43–63

Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D (2018) A survey of methods for explaining black box models. ACM Comput Surv (CSUR) 51(5):1–42

Gunning D, Aha D (2019) DARPA's explainable Artificial Intelligence (XAI) Program. AI Magazine 40(2):44–58. https://doi.org/10.1609/aimag.v40i2.2850

Guthrie C, Rachlinski JJ, Wistrich AJ (2001) Inside the judicial mind. Cornell Law Rev 86:778–830

Hacker P, Krestel R, Grundmann S, Naumann F (2020) Explainable AI under contract and tort law: legal incentives and technical challenges. Artif Intell Law 28:415–439

Hacker P, Passoth JH (2022), April Varieties of AI Explanations Under the Law. From the GDPR to the AIA, and Beyond. In xxAI-Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers (pp. 343–373). Cham: Springer International Publishing

Hage J (2005) Studies in legal logic. Springer, Dordrecht

Haidt J (2001) The emotional dog and its rational tail: a social intuitionist approach to moral judgement. Psychol Rev 108:814–834

Hintikka J, Remes U (1974) The method of analysis. D. Reidel, Dordrecht

Hurley M, Dennett D, Adams R (2011) Inside jokes: using humor to reverse-engineer the mind. MIT Press, Cambridge, MA

Hutchison A (2014) The Whanganui River as a legal person. Altern Law J 39:179–182

Jain S, Wallace BC (2019) Attention is not explanation. arXiv preprint arXiv:1902.10186v3

Kaminski ME (2021) The right to explanation, explained. Research Handbook on Information Law and Governance. Edward Elgar Publishing, pp 278–299

Kant I (1909) Kant's critique of practical reason and other works on the theory of Ethics. Longmans, Green & Co., London

Kruglanski A (1989) The psychology of being "right": the problem of accuracy in social perception and cognition. Psychol Bull 106:395–409

Lillard A (1998) Ethnopsychologies: Cultural Variations in theories of mind. Psychol Bull 123:3–32

Lipton ZC (2018) The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. Queue 16(3):31–57

Malgieri G, Comandé G (2017) Why a right to legibility of automated decision-making exists in the general data protection regulation. Int Data Priv Law 7(4):243–265

Miller T (2019) Explanation in artificial intelligence: insights from the social sciences. Artif Intell 267:1–38

Mittelstadt B, Russell C, Wachter S (2019), January Explaining explanations in AI. In Proceedings of the conference on fairness, accountability, and transparency (pp. 279–288)

Morris M, Peng K (1994) Culture and cause: american and chinese attributions for Social and physical events. J Person Soc Psychol: 949–971

Murphy R, Woods DD (2009) Beyond Asimov: the three laws of responsible Robotics. IEEE Intell Syst 24(4):14–20. https://doi.org/10.1109/MIS.2009.69

Pasquale F (2017) Toward a fourth law of robotics: preserving attribution, responsibility, and explainability in an algorithmic society. Ohio St LJ 78:1243

Phillips PJ, Hahn CA, Fontana PC, Broniatowski DA, Przybocki MA (2021) Four Principles of Explainable Artificial Intelligence. Draft NISTIR 8312. https://doi.org/10.6028/NIST.IR.8312-draft

Prakken H (2020) A top-level model of case-based argumentation for explanation. In Proceedings of the ECAI 2020 Workshop on Dialogue, Explanation and Argumentation for Human-Agent Interaction (DEXA HAI 2020)

Prakken H, Ratsma R (2021) A top-level model of case-based argumentation for explanation: formalisation and experiments. Argument & Computation, pp 1–36. Preprint

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. https://doi.org/10.1038/s42256-019-0048-x

Selbst AD, Barocas S (2018) The intuitive appeal of explainable machines. Fordham L Rev 87:1085

Selbst A, Powles J (2018) January "Meaningful information" and the right to explanation. In conference on fairness, accountability and transparency (pp. 48–48). PMLR

Sovrano F, Sapienza S, Palmirani M, Vitali F (2022) Metrics, explainability and the european AI act proposal. J 5(1):126–138

Stelmach J, Brożek B (2006) Methods of legal reasoning. Springer, Dordrecht

Vale D, El-Sharif A, Ali M (2022) Explainable artificial intelligence (XAI) post-hoc explainability methods: risks and limitations in non-discrimination law. AI and Ethics: 1–12

Vilone G, Longo L (2021) Notions of explainability and evaluation approaches for explainable artificial intelligence. Inform Fusion 76:89–106

Wachter S, Mittelstadt B, Floridi L (2017a) Why a right to explanation of automated decision-making does not exist in the general data protection regulation. Int Data Priv Law 7(2):76–99

Wachter S, Mittelstadt B, Floridi L (2017b) Transparent, explainable, and accountable AI for robotics. Sci Rob 2(6):eaan6080

Webster D, Kruglanski A (1994) Individual differences in need for cognitive closure. J Personal Soc Psychol 67:1049–1062

Wiegreffe S, Pinter Y (2019) Attention is not not explanation. arXiv preprint arXiv:1908.04626v2

Wróblewski J (1992) The judicial application of Law. Kluwer Academic Publishers, Dordrecht

Ye H, Jiang X, Luo Z, Chao W (2018) Interpretable charge predictions for criminal cases: learning to generate court views from fact descriptions. In: Proceedings of the conference of the North American chapter of the association for computational linguistics: human language technologies. pp. 1854–1864

Zednik C (2021) Solving the black box problem: a normative framework for explainable artificial intelligence. Philos Technol 34(2):265–288

Zuiderveen Borgesius FJ (2020) Strengthening legal protection against discrimination by algorithms and artificial intelligence. Int J Hum Rights 24(10):1572–1593

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.