Explainable AI and the Philosophy and Practice of Explanation

Kieron O'Hara

Electronics and Computer Science, University of Southampton, SO17 1BJ, United Kingdom

kmoh@soton.ac.uk

https://orcid.org/0000-0002-9051-4456

Accepted version of paper for *Computer Law and Security Review*, August 2020

Abstract: Considerations of the nature of explanation and the law are brought together to argue that computed accounts of AI systems' outputs cannot function on their own as explanations of decisions informed by AI. The important context for this inquiry is set by Article 22(3) of GDPR. The paper looks at the question of what an explanation is from the point of view of the philosophy of science – i.e. it asks not what counts as explanatory in legal terms, or what an AI system might compute using provenance metadata, but rather what explanation as a social practice consists in, arguing that explanation is an illocutionary act, and that it should be considered as a process, not a text. It cannot therefore be computed, although computed accounts of AI systems are likely to be important inputs to the explanatory process.

Keywords: Explanation, AI, GDPR, explainable AI, XAI

1. Introduction

The question considered in this paper is whether explanations of the outputs of artificial intelligence (AI) or machine learning (ML) can be wholly computable or not. Many AI or ML processes are highly complex, especially when performed over big data. Furthermore, many, particularly using methods such as neural nets or deep learning, are referred to as 'opaque' or being concealed within a 'black box'. This is a misleading description, however, because the decision-making may be transparent, and the weights and outputs of the various nodes clear and accessible.

The problem with such information is that it may not be explanatory, in a sense outlined in section 3. Roughly, the real-world relevance of the operation of the system is in terms meaningful in a social context (e.g. a person may or may not be judged creditworthy), whereas the parameters of the system in operation (weights and outputs) are not meaningful in the same way. An ML system is designed to uncover patterns or regularities in datasets that would be practically impossible for humans to discern, and a system based on deep learning is modelled (roughly) on neuronal processes in the brain, which themselves are barely understood, rather than conscious psychological inference. Hence systems of this type, be they ever so transparent, will produce outputs that will to some extent have to be 'taken on trust'.

Explaining AI outputs may have a number of important goals, including enabling the management and improvement of systems. However, the goal that is the focus of this paper is that of justifying the output to an external and possibly sceptical audience. The reason this is a tricky problem is discussed in section 3. One area of research in AI is that of explainable AI (XAI), that is, using the parameters relevant to decision-making to compute an account of the output that is expressed in meaningful and explanatory terms (Adadi & Barrada 2018, Monroe 2018), for example based on provenance metadata gleaned from the operation of the AI system (Groth & Moreau 2013). In this paper, I will argue that this research programme, while arguably necessary, is not sufficient to achieve its goals.

The computation of an ‘explanation’ based on provenance will doubtless be useful and relevant to explaining the decision in a real-world social context, and the computed ‘explanation’ may be a valuable exhibit during the process of explaining the decision (as, in fact, the non-socially-meaningful weights and outputs might also be). However, the more ambitious claim that the explanatory task might end when such an account has been given is, I shall argue, false. If the computed ‘explanations’ have proven reliable in some specific context (say, in an industrial process), or if not very much hangs on them, then we might treat them as effectively final and adequate. But in a context of socially-sensitive decision-making, this would be (a) a descriptive error, misunderstanding the process of explanation, and (b) a normative error, failing to see what evaluative standards are appropriate in sensitive contexts.

I shall assume that, in a social context, AI produces *output* which feeds into a *decision*. AI has no *decision-making role*, in the sense that it has no *responsibility* for any decision taken as a result of the output. Its quantitative output is interpreted during some social process, and the interpretation feeds into decision-making. It may be that the decision-making process is highly streamlined, so that the output is never questioned, and its interpretation very straightforward (e.g. *if x>0.5 then creditworthy, else uncreditworthy*). But even so, two points remain clear. First, the administrators of the system retain responsibility for the decisions, even if they in practice never intervene (and even if the architecture of the system denies them the opportunity to intervene). Second, to have real-world effect, there has to be some kind of actuation mechanism which is conceptually separate from the production of the AI output. This mechanism also has to be accounted for by the explanation.

The paper will have four substantive sections. Section 2 will briefly characterise the important context for this inquiry as set by the General Data Protection Regulation (GDPR) of 2018. Section 3 will consider the question of what an explanation is from the point of view of the philosophy of science – i.e. it will ask not what counts as explanatory in legal terms, nor what might be possible within XAI, but rather what explanation as a social process consists in. Section 4 will add some considerations brought in by the legal context, and section 5 will consider the opening question of whether AI explanations can be considered as wholly computable.

2. Context: the requirements set by GDPR

Explaining AI output has long been a research programme (Swartout 1983, Southwick 1991, Berry 1997). During the days of rule-based expert systems, there were concerns that explanations would be required in order for their recommendations to be taken seriously. Simply tracing the rules that fired was a beginning, because the rules which were used in expert systems programs were often taken from an expert’s practice, and so were understandable to an extent by outsiders. However, the justifications for the rules, which are also relevant to explanation, were used by coders in writing the programs, but did not appear in the code itself, and so the inferences of systems, the models upon which they were based and the principles governing expertise in the domain were traced and mined for illuminating and justificatory accounts of why a certain output was produced (Swartout 1985, O’Hara 1994, Schreiber et al 2000).

XAI faces a different problem, which is that the operation of an ML system does not map on straightforwardly to human-readable concepts at all, so the programme described above could not be explanatory of that technology. I have some sympathy with Robbins’ contention that if we had such an explanation, we wouldn’t need the technology, since we could do the inference using rule-based AI, but I won’t pursue this issue here (Robbins 2019).

However, XAI has become more imperative because the EU’s GDPR has brought explanation into its framework of data protection. GDPR provides for punitive fines for transgressors, and so has gained attention; nevertheless what it counts as an adequate explanation will not be properly understood until we have amassed sufficient case law. The term ‘explanation’ appears only in Recital 71:

> In any case, such processing [e.g. automatic profiling] should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision.

This recital is best understood as commenting on Article 22(3), which states that, in the cases where the data subject has consented to the automatic decision-making, or where it is essential for performance of a contract between data subject and data controller:

> … the data controller shall implement suitable measures to safeguard the data subject’s rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.

Veale and Edwards (2018) remark on the failure to include ‘explanation’ specifically in the binding part of the regulation. The Article 29 Working Party (WP29 2018) glosses it as requiring the data controller to *explain* the significance and the consequences of the processing *in terms meaningful to the data subject*. This is a sensible requirement, but arguably at odds with the text of Recital 71, which makes no such demand unless it is already implicit within the idea of ‘explanation’. In any case, as Veale and Edwards note, it may well be that such an explanation does not correspond well with the sorts of explanations produced by XAI.

Part of our problem is that the notion of explanation has not itself been explained yet, so in the next section we need to consider what the properties of an explanation are.

3. Explanation

To triangulate between the various disciplines at loggerheads here, I will develop an account of explanation using disciplinary apparatus distinct both from the legal context and XAI. Explanation has traditionally been treated philosophically as a topic in the philosophy of science, and so has often been understood as something derived from the deductions and methods of the scientist. Aristotle in the *Posterior Analytics* takes the explanation of an event to be its *cause*, as do more recent social scientists such as Jon Elster (2015). In his ground-breaking work, Carl Hempel takes it to be a deductive or inductive *argument* for the truth of the conclusion (Hempel 1962, 1965). Note that neither of these classic accounts stresses the need to accommodate an audience, although they do suggest that explaining involves the production of a text or object – the explanation. To this extent, XAI adheres to this orthodoxy.

There are, however, problems with such accounts, especially in the context in which we are interested.

3.1 Explanation and understanding

Firstly, explanation has a *pragmatics*, with an aim or goal, in which it may or may not succeed. To explain something successfully is to enable *understanding* of that thing. This does imply (as WP29 hinted) that explanation is relative to an audience (the data subject in the GDPR case), by whom it needs to be understandable (Van Fraassen 1980, Achinstein 1983, Edwards & Veale 2017).

Understanding goes beyond a state of knowledge, although philosophers differ about how the distinction should be characterised; metaphorically at least, Strevens suggests that the relevant cognitive attitude in understanding is a *grasp* of a phenomenon, rather than simply knowing about it (2013). Hills argues that understanding of a phenomenon entails practical knowledge of a set of abilities (2009), while Grimm suggests that understanding is to do with a wider set of knowledge claims about various modal relationships between the parts of a phenomenon (2014). Even reductionists who are sceptical that understanding is a special epistemological or cognitive state agree that it requires a good and deep supply of relevant knowledge, including relevant theory and perhaps counterfactuals (Sliwa 2015). I would characterise an understanding of a proposition in terms of the supporting web of information and knowledge, both of relevant specifics and background knowledge, such that the specific knowledge about the phenomenon to be explained is supported, so that it is unlikely to be forgotten easily, it can be justified to others, it can be translated into different (simpler or more complex) terms, lessons can be drawn from it, questions can be answered about it, and so on.

To take an example, someone who reads an article by a reliable author he trusts finds out that hydrogen cells are inefficient, and so unlikely to be part of the solution to climate change, even though they are non-polluting. Our reader has learnt something about the phenomenon, but cannot really be said to understand it. The author does understand it, because she has a good deal more relevant knowledge about it. For example, she is aware that burning hydrogen produces energy and the only by-product is water, but also knows that the most practical source of hydrogen that does not produce large quantities of greenhouse gases in the first place is through the electrolysis of water, which requires more energy than would be released in its combustion. Therefore one might as well use the electricity one would have used for electrolysis directly in whatever one wanted the hydrogen cell for. This is why she wrote what she wrote, and she will be able to dress up the lesson in different forms, and use her knowledge and understanding in a range of tasks – unlike her reader.

The boundary between knowledge and understanding is no doubt fuzzy, but it seems clear that to explain something, and therefore produce understanding in the inquirer, it is not enough simply to impart some propositions or a model of inference. Some sort of relevant support or background must also be imparted, and what background is appropriate will depend in part upon the capacities and existing knowledge of the inquirer. The understanding subject must also be able to do more than merely parrot the proposition that he or she has just learned. This is why a university education goes beyond setting out sets of propositions to be memorised, but also includes lab exercises, moots, individual and collaborative projects, tutorials, work placements, mock exams, and even criticising the orthodox knowledge that has been learned. It is also why someone who can recite the monarchs of England for a pub trivia quiz does not thereby qualify as a historian.

Indeed, being handed a read-out of the output of XAI hardly even constitutes knowledge, never mind a deeper state of understanding. Compare someone who is handed a copy of Einstein's paper 'Does the inertia of a body depend upon its energy content?' – such a person may or may not be said to know why mass and energy are equivalent as a result, but certainly cannot be said to understand why. Similarly, someone handed the XAI readout, however sweetened with syntactic sugar, may possibly be said to *know* why a decision has been taken, but surely cannot be said to *understand* why. This is what Edwards and Veale call the "fallacy of transparency" (2017).

3.2 Explanation as process versus explanation as text

Secondly, there is an important ambiguity in the term 'explanation'. An explanation may be an object or a text (such as the output of XAI), but on another interpretation is a process or performance that exists through time, has a beginning and an end, and is distinct from other explanations that occur in different times and places, even of the same phenomenon using identical terms. The *process* of explanation is an *illocutionary act* (Austin 1962), an utterance which performs a task, similar to a request, a promise or a command. The goal of an explanation is the achievement of understanding of a phenomenon by the audience.

In contrast to much of the literature in the philosophy of explanation, I argue that the process of explanation is the prior notion, and the text secondary. To see this, consider a basic description of an explanatory event: 'X explained Q to Y by saying P'. Here, Q is a problem or question, and P is a series of propositions giving an account of Q. X is the explainer, and Y the inquirer in search of understanding. The whole event described by the statement is an explanation of the process-type, and the explanation of the text-type is P.

However, as anyone who has ever explained anything to a child, student, spouse, policeman, or in a court of law will know, this idealised statement conceals a lot of complexity. One very rarely utters a series of propositions, or hands over a piece of paper with an explanatory account, to an inquirer so that the job is done. X may well begin with an idea of what he wants to say, but there typically will follow a conversation, with some debate, questions, clarifications and challenges. Y will try out X's explanation: will it cover counterfactual circumstances? She might compare it with alternative explanations. It might go against some of her own prior understanding of Q that X does not have access to. X may have to alter or adjust his own point of view in the face of such a challenge.

In the end, if X is successful, Y will have achieved understanding of Q, but the text P which some want to call the 'explanation' is co-produced by explainer and inquirer, as a result of some trial and error, some adjustments to cater for Y's interests, capacities and circumstances, and possibly wholesale changes, in a process which may prove as enlightening to X as to Y, as Y reveals and tests X's assumptions and brings her own specific perspective to bear. And then note that if X sets out to explain Q to a new inquirer Y*, she may make a wholly different set of demands on X during this second process of explanation, resulting in a different text-type explanation P*. Hence the text-type explanation P cannot be assumed to be given in advance by superior X and handed over to deficient Y and Y*. Rather, P (and P*) are the *products* of process-type explanation-events, and dependent on their successful execution. No process, no product (and no understanding).

In the rest of this paper, I shall therefore take 'explanation' to refer to the process or illocutionary act, which is the prior concept. There is a technical vocabulary which can be used to avoid ambiguity: the phenomenon to be explained (Q, in the above schema) is often referred to as the *explanandum*, and the text used to explain it (P) as the *explanans* (Hempel 1965). I will refer to the output of XAI as an explanans, and so our question is whether an XAI explanans can be sufficient to explain the explanandum of the decision reached about a data subject after data processing.

Explanation as performance is important, because the production of understanding in the audience will be *facilitated* rather than hindered by a set process (e.g. in education, scientific research or courts of law) in which the audience also participates, to test the explainer's reasoning and see how it emerged. Mere presentation of the explanans in the absence of a supplementary process of explanation will be less responsive to the needs of the data subject. The explainer will be unable to see what is meaningful to the data subject, and unable to adjust the explanation accordingly, and so pragmatically less likely to facilitate understanding.

Note in contrast that the presentation of a computable explanans without further process removes the possibility of evaluation from this stage of a dispute. The data subject will be unable to question further the logic of the decision, or indeed to verify whether the explanans contains all the information needed to understand it. The explainer will also be unable to verify whether understanding has been reached by the data subject.

It was the role of performance in the production of understanding that led Plato to argue in the *Phaedrus* (in writing, through the mouth of Socrates) that spoken discourse was superior to writing.

> SOCRATES: You know, Phaedrus, writing shares a strange feature with painting. The offspring of painting stand there as if they are alive, but if anyone asks them anything, they remain most solemnly silent. The same is true of written words. You'd think they were speaking as if they had some understanding, but if you question anything that has been said because you want to learn more, it continues to signify just that very same thing forever. (Plato 1997, 552, 275de)

3.3 The interests of the inquirer

Thirdly, the explanandum is not objectively presented, but rather appears as a problem from the perspective of a questioner. Why did X happen? This can be qualified in a number of ways depending on the interests of the audience. *Why did X happen rather than Y?* may need a different explanans to the explanandum *How did X happen?* The explainer and the questioner might easily have different questions or contrastive cases in mind, which may well not be evident without structured conversation.

This matters in big data analysis. Where ML techniques have extracted regularities from a very large dataset, we find non-obvious, often interesting and sometimes important correlations between parameters. However, they are merely correlations, as the boosters of big data are happy to concede (Mayer-Schönberger & Cukier 2013). That M and N are correlated does not tell us whether M caused N, N caused M, or whether they both had some common but exogenous cause. Because the sample size is so large, we can be sure of the correlation, which for policy purposes may be all we need.

But that means that an account of the inference of the correlation between M and N cannot answer the question *why* the decision was justified, or in other words, the cause of the correlation. It can only be used to explain *how* the decision was made. This may well be important and valuable, especially from the company's point of view to show they have not been negligent, without meeting the needs of the inquirer.

Consider someone who is denied car insurance because he wears red trousers, who asks the data processors for an explanation of the decision. The how-explanation will tell the man how the full, unbiased data concealed a significant correlation between car accidents and red-trouser-wearing, and hence that the company's decision was commercially justified, carefully made, and properly evaluated. This is important. But the unfortunate inquirer is likely to be more concerned with the question of *why* red-trouser-wearing is correlated with accidents, not simply to reassure himself that the company has not made a frivolous decision. What he wants to know, in effect, is whether changing his trousers will enable him to get insurance, and if not, what he else he needs to do.

The aim of an explanation is often to facilitate future action. Explaining human biochemistry in terms of the genome facilitates personalised medicine. Explaining expertise facilitates the construction of expert systems (O'Hara 1994). GDPR recital 71 makes clear that the purpose of the explanation owed to the data subject is to allow challenge of decisions that depend on AI or other processing in law. This will certainly require understanding of the explanation (if not by the data subject, then by his or her lawyer).

One extra purpose not mentioned in GDPR, but which may be the legitimate and valuable purpose of an explanation, is for data subjects to be able to understand what it was about their past behaviour as represented in the data that led to the irksome decision, and to change their behaviour accordingly. For instance, if one was refused a loan because of past failure to keep up payments, then one might learn to make regular payments and keep lenders informed of changes of circumstances. If a decision is effectively arbitrary, then one is no better off for an explanation because one has no idea what to do next time to produce a different outcome. And an explanation that does not refer to concrete behavioural concepts that complainants can understand and interpret in the context of their personal lives *is* effectively arbitrary relative to this purpose, as in the *Little Britain* comedy sketches where "computer says no".

An explanation, then, is not a text such as the output of XAI. It is an illocutionary act which produces an explanans tailored to the capacities and intentions of the inquirer. That is not to say that XAI is not valuable – quite the opposite, it may be the only way of getting to know what happened when the program ran. It is clear, however, that, even if necessary, it is not sufficient; it can only be the *input* to an explanatory process, not the explanation itself

4. Law

Legal processes are another kind of speech act, distinct from what Brownsword (2019) calls *technological management*, the use of design to prevent certain undesired actions being possible. Rather than being of this categorical form, law is written, either as legislation or judgments, to be interpreted in a specific legal context. It is therefore a hermeneutic practice, involving the *interpretation* of the evidence in terms of written law, which itself is open to interpretation. When someone in the appropriate role speaks (e.g. a judge), this has legal effect, and more law is produced (Hildebrandt 2015).

The regulatory function of law requires it to be broadly predictable. If it is to guide and coordinate our actions, then we must have a reasonable idea of how a new case would be interpreted in the light of past decisions. Prior to actual dispute, GDPR, like any law, should help us know where we stand. It is intended to enable AI-supported decision-making. The decision-maker is made aware that decisions must be made meaningful (and be fair, non-discriminatory, etc.). Data subjects should similarly be aware that compliant decisions will be explainable, and in most cases could take that on trust. Where a decision is particularly irksome, they have the option of testing its rationale through legal action, via national regulators and ultimately in court.

Contestation as envisaged by Recital 71 involves arguments being put by plaintiff and defendant about the propriety of a decision made with input from an automated system. This requires each side being able to anticipate, to an extent, how their case will be received by the court. Only when there is uncertainty this will a case actually come to court. Furthermore, contestation is the result of people having an important stake, bringing risk, as in socially-sensitive contexts. As contestation necessitates the adoption of antagonistic positions, it would hardly be unbiased unless the decision-makers' XAI explanans had standing independent of the disputed decision.

A GDPR-compliant explanation of an AI-informed decision given by a decision-maker must therefore allow the data subject (or their legal team) to understand the rationale for the decision sufficiently to contest the case by interpreting past decisions and legislation in the context of the present specifics, and to anticipate with reasonable accuracy how a court would respond. The exact range of acceptable forms of explanation will not be known until case law has been amassed, but the XAI explanans surely cannot be taken as the canonical account of what happened, for that would in effect ensure victory for the decision-makers as long as they were justified in their own terms.

5. Discussion

It does not seem plausible that the output of XAI could function as an explanation as required in a sensitive context, in the face of a regulation such as GDPR.

The first point to note is that the output of the AI is not the decision of the social system (Robbins 2019). The AI exists within a context, and it is run in order to facilitate decision-making. However, data still needs to be gathered and cleaned (and checked for bias, etc.), the AI needs to be managed, and the recommended decisions need to be actuated. A social system or organisation is needed to carry out decisions that affect individuals, requiring an explanation not only of how the AI produces its output, but also how that output is used within the wider system. That is not computable from within the AI system, even if the AI is most of the story. Something else could always be done instead of what the AI recommends. And if the data subject is to be able to contest a decision, he or she needs to understand the decision-making as a whole, not simply the AI component.

In particular, ML involves the application of one or more algorithms to some data. The algorithms will most likely be reasonably well-known, and their properties understood. Their operation on the data will be traceable via provenance metadata or other kinds of audit trails, and the verification of their propriety relatively straightforward too. Any complex explanandum in the decision-making is therefore likely to require investigation of the data in which the algorithm has found a significant pattern. Is it complete enough? Does it contain biases? Is it missing important fields? Is its quality high enough? Has data from multiple sources been integrated successfully and consistently? No doubt the properties of the data as revealed by algorithmic analysis are highly relevant to these judgments, but they cannot be the whole story, *particularly where the analysis as a whole is in question*. Most of the data wrangling will happen either before an algorithm is applied, or interleaved with exploratory work by data scientists to get a sense of what the data might reveal. This is a human process, an art as much as a science; it will be vital in explaining a decision, and is clearly not computable.

Secondly, in order to contest a decision, the data subject must understand it. To facilitate this, as argued above we should take 'explanation' in its performative sense, not in the sense of a product or text. A process of communication, ideally standardised, is far more likely to result in verifiable understanding on the part of the data subject. Simply presenting the explanans to a subject is not guaranteed to produce understanding in the required sense, and – at least until the case law tells us what kind of explanans is acceptable under GDPR – it is in the defensive interests of the decision-maker to be able to evaluate the subject's understanding of the decision-making.

Thirdly, the data subject's lawyers must be able to take their understanding of the decision into court and contest it, creating their own interpretation of past law and the current decision and presenting it before the judge for a ruling. This surely requires a perspective on the decision independent of that provided by the decision-maker (i.e. the computed account of the XAI). Hence, while the account is useful, it cannot be taken as the whole explanation.

Fourthly, if GDPR and similar legislation is to steer our actions so that we don't end up in court all the time, then we need to be able to predict what a regulator or judge is likely to conclude about a case. XAI cannot in and of itself anticipate such a judgment without supplementation. Even if a particular algorithm was extremely reliable and well-tested in court, so that a computed explanans could be seen as highly credible, there is always the possibility that the plaintiff has unusual circumstances, or other judicial norms are relevant, so that the court will look at this case differently.

Fifthly, a computed account of XAI will be an important management tool for decision-makers, but managers need to understand the overall decision-making process, the nature of their responsibilities, and the specific role of the AI, not just the output of the AI in isolation.

Finally, the nature of data science means that XAI is likely to be able to compute how-explanations that can reassure managers, regulators and data subjects that a decision was made conscientiously, carefully and according to accepted methodologies, but it cannot compute why-explanations, which require a causal account of the detected regularity in a domain, rather than merely identifying significant correlations. However, when the purpose of the explanation is to enable the data subject to change behaviour so as to receive a different decision in the future, the why-explanation is far more valuable than the how-explanation.

Hence, while a computed account of the output of an AI system may contribute a great deal of value, to call it an 'explanation' (in the sense of something that has enabled the audience to understand the explanandum) puts excessive weight upon it. At best, all such an account can do is to feed into various explanatory processes. This is no small contribution, but as well as working out how such an account can be best produced in XAI, additional research is needed to investigate how it can inform human and social decision-making, to make it meaningful and valuable to all sides.

References

Achinstein, P. The Nature of Explanation. Oxford University Press: New York; 1983.

Adadi, A., Berrada, M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 2018; 6: 52138-52160, https://doi.org/10.1109/ACCESS.2018.2870052.

Austin, J.L. How To Do Things With Words. Oxford: Clarendon Press; 1962.

Berry, F.S. Explaining managerial acceptance of expert systems. Public Productivity & Management Review 1997; 20(3): 323-335, https://doi.org/10.2307/3380981.

Brownsword, R. Law, Technology and Society: Re-Imagining the Regulatory Environment. Abingdon: Routledge; 2019.

Edwards, L., Veale, M. Slave to the algorithm? Why a 'right to an explanation' is probably not the remedy you are looking for. Duke Law & Technology Review 2017; 16(1): 18-84.

Elster, J. Explaining Social Behavior: More Nuts and Bolts for the Social Sciences, revised edition. Cambridge: Cambridge University Press; 2015.

Grimm, S.R. Understanding as knowledge of causes. In: Fairweather, F., editor. Virtue Epistemology Naturalized: Bridges Between Virtue Epistemology and Philosophy of Science, Cham: Springer; 2014. 329-345, https://doi.org/10.1007/978-3-319-04672-3_19.

Groth, P., Moreau, L. PROV-Overview: An Overview of the PROV Family of Documents; 2013. Available from http://www.w3.org/TR/prov-overview/. [Accessed 5th Aug, 2020]

Hempel, C.G. Explanation in science and in history. In: Colodny, R.G., editor. Frontiers of Science and Philosophy, London: Allen & Unwin; 1962. 7-33.

Hempel, C.G. Aspects of Scientific Explanation. New York: Free Press; 1965.

Hildebrandt, M. Smart Technologies and the End(s) of Law. Cheltenham: Edward Elgar; 2015.

Hills, A. Moral testimony and moral epistemology. Ethics 2009; 120(1): 94-127, https://doi.org/10.2307/3380981.

Mayer-Schönberger, V., Cukier, K. Big Data: A Revolution That Will Transform How We Live, Work and Think. London: John Murray; 2013.

Monroe, D. AI, explain yourself. Communications of the ACM 2018; 61(11): 11-13, https://doi.org/10.1145/3276742.

O'Hara, K. Mind as Machine: Can Computational Processes Be Regarded As Explanatory of Mental Processes? University of Oxford DPhil thesis; 1994. Available from https://eprints.soton.ac.uk/254167/. [Accessed 5th Aug, 2020]

Plato. Phaedrus. In: Cooper, J.M., editor. Plato: Complete Works, Indianapolis: Hackett; 1997. 506-556.

Robbins, S. A misdirected principle with a catch: explicability for AI. Minds & Machines 2019; 29(4): 495-514, https://doi.org/10.1007/s11023-019-09509-3.

Schreiber, G., Akkermans, H., Anjewierden, A., de Hoog, R., Shadbolt, N., Van de Velde, W., Wielinga, B. Knowledge Engineering and Management: The CommonKADS Methodology. Cambridge MA: M.I.T. Press; 2000.

Sliwa, P. Understanding and knowing. Proceedings of the Aristotelian Society 2015; 115(1): 57-74, https://doi.org/10.1111/j.1467-9264.2015.00384.x.

Southwick, R.W. Explaining reasoning: an overview of explanation in knowledge-based systems. Knowledge Engineering Review 1991; 6(1): 1-19, https://doi.org/10.1017/S0269888900005555.

Strevens, M. No understanding without explanation. Studies in History and Philosophy of Science Part A 2013; 44(3): 510-515, https://doi.org/10.1016/j.shpsa.2012.12.005.

Swartout, W.R. XPLAIN: a system for creating and explaining expert consulting programs. Artificial Intelligence 1983; 21(3): 285-325, https://doi.org/10.1016/S0004-3702(83)80014-9.

Swartout, W.R. Explaining and justifying expert consulting programs. In: Reggia, J.A., Tuhrim, S., editors. Computer-Assisted Medical Decision Making, New York: Springer; 1985. 254-271, https://doi.org/10.1007/978-1-4612-5108-8_15.

Van Fraassen, B.C. The Scientific Image. Oxford: Clarendon Press; 1980.

Veale, M., Edwards, L. Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling. Computer Law and Security Review 2018; 34: 398-404, https://doi.org/10.1016/j.clsr.2017.12.002.

WP29. Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679. Article 29 Working Party; 2018. Available from https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053. [Accessed 5th Aug, 2020]