In No Circumstances Can or Should Explanations of AI Outputs in Sensitive Contexts Be Wholly Computable

Kieron O'Hara
Electronics and Computer Science
University of Southampton
Southampton, UK
kmoh@soton.ac.uk

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.

CCS CONCEPTS

• Applied computing: law, social and behavioural sciences: law

KEYWORDS

Explanation, AI, GDPR

ACM Reference format:

Kieron O'Hara. 2020. In No Circumstances Can or Should Explanations of AI Outputs in Sensitive Contexts Be Wholly Computable. In *Workshop on Explanations for AI: Computable or Not?* at *Websci 2020.*

1 Introduction

The question set in this workshop is whether AI explanations are 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 to be explained below. Roughly, the real-world relevance of the operation of the system will be 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. The reason this is a serious problem is discussed in section 3.

One area of research in AI is that of computing explanations, that is, using the parameters relevant to decision-making to compute an account of the output that is expressed in meaningful terms. In this paper, I will argue that this research programme, which may be necessary, is not sufficient to achieve its goals.

The computation of an 'explanation' may 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 take them as final. 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.

One final definitional point is 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 designers and administrators of the system retain responsibility for the decisions, even if they in practice never 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.

2 The Requirements Set By GDPR

Explaining AI output has long been a research programme. During the days of expert systems, there were concerns that explanations would be required in order for their recommendations to be taken seriously, and so the inferences of

systems were traced and mined for illuminating accounts of why a certain output was produced [5].

However, the programme has taken off recently, partly because AI algorithms have become harder to oversee, but more concretely, the EU's GDPR has brought explanation into law. GDPR provides for punitive fines for transgressors, and so has gained attention; nevertheless its significance is not always obvious, and will not be so 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.

The Article 29 Working Party [2] glosses this 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'. This means we need to consider what the properties of an explanation are.

3 Explanation

Explanation has traditionally been treated within philosophy as a branch of the philosophy of science, and so has often been understood as something derived from the deductions and methods of the scientist. Aristotle takes the explanation of an event to be its cause. In his ground-breaking work, Hempel takes it to be a deductive or inductive argument whose conclusion is the explanandum. Note that neither of these classic accounts stresses the need for meaningfulness, although they do suggest that explaining involves the production of a text or object – the explanation [1]. To this extent, the research programme of explaining AI adheres to this classical orthodoxy.

What such accounts miss is that explanation has a pragmatics, and involves an aim or goal. 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 [1].

Secondly, they miss an important ambiguity in the idea of 'explanation'. An explanation may be an *object* or a text (which may be computed by an AI system), but equally it may be a *process*, performance or speech act. The process of explanation which information, with the goal of the achievement of understanding of the explanandum by the audience. Explanation as performance is extremely important, because the production of understanding in the audience will be facilitated much more easily by some set kind of process (e.g. in education, scientific research or courts of law) in which the audience also participates, and comes to test the explainer's reasoning and see how it has emerged. Mere presentation of a computable explanation 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 produce understanding in the data subject.

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

Note also that the person responsible for the AI-informed decision, who may be a non-technical manager, may also not understand the 'explanation' presented as a text or computable artifact. The process of explanation may be as valuable for the person responsible as for the person directly affected.

Thirdly, the explanandum is not objectively presented, but rather appears 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 explanation to Why did X happen rather than Z? The explainer and the questioner might easily have different contrastive cases in mind.

Finally, the aim of an explanation is often to facilitate future action. Explaining human biochemistry in terms of the genome is intended to facilitate applications in medicine. My own DPhil research looked at different kinds of explanation of expertise to facilitate the construction of expert systems [5]. Recital 71 makes clear that the purpose of the explanation is to allow the data subject to challenge 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). Note that survey work on what people *find* explanatory, while interesting, cannot be sufficient to *determine* whether an explanation is adequate for a given task.

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. The data

subject would not win the case, but at least would have learned a valuable lesson.

4 Law

Legal processes are another kind of speech act. Law is not a series of if-then-else rules, where the legal process involves checking whether or not the context matches the 'if' clause, and then performing the appropriate action (this is what Brownsword calls technological management [3]). It is written, either as legislation or judgments, to be interpreted in a specific legal context. Law is therefore a hermeneutic practice, involving the *interpretation* of the evidence in terms of the 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 [4].

The function of law requires it to be broadly predictable. If it is to guide our actions, then we must have a reasonable idea of how a new case would be interpreted in the light of past decisions. This is why major changes in the law (such as GDPR) can lead to temporary uncertainty and concern.

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 great uncertainty about the future decision will a case actually come to court. Furthermore, contestation is the result of people having an important stake, as is the case in socially-sensitive contexts. As contestation requires the adoption of antagonistic positions, it would hardly be unbiased unless the decision-makers' computed account of the output of the AI had standing independent of the disputed decision.

GDPR, like any law, is there to coordinate action. Except in case of actual dispute, it should help us know where we stand. GDPR is intended to enable AI-supported decision-making. The decision-maker is aware that decisions must be made meaningful (and must be fair, non-discriminatory, etc.). Data subjects should similarly be aware that compliant decisions will be explainable, and in most cases would take that on trust. Where a decision is particularly irksome, they have the option of questioning its rationale.

The understanding of the AI-informed decision that the explanation brings to the data subject (or to his or her legal team) must therefore equip them to contest the case by interpreting past decisions and legislation in the context of the specifics of the present day, and to anticipate how a court will respond.

5 Discussion

How are we therefore to answer our exam question? Can a computed account of the output of the AI system 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. Some sort of actuation is also needed, and so any explanation of the decision must not only take the AI's output into account, it needs also to include how that output results in a decision for which a person or organization is responsible. That is not computable from within the AI system, even if the system is most of the story.

Secondly, in order to contest a decision, the data subject must understand it. To facilitate this, we should take 'explanation' in its performative sense, not in the sense of a product or text. A process of communication is far more likely to result in verifiable understanding on the part of the data subject.

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 AI's output). 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 judge is likely to conclude about a case. The computation of an account of the AI output 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 'explanation' could be seen as highly credible, there is always the possibility that the plaintiff has unusual and relevant evidence, or more widely that judicial norms have shifted, and so the court will look at this case differently.

Fifthly, a computed account of the output will be an important management tool for decision-makers, and their practice may alter on the basis of their own understanding of the output. The computed account will certainly contribute, but the overall decision-making process, the nature of their responsibilities, and the specific role of the AI need to be understood as a whole by managers, not just the output of the AI system in isolation.

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, 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

[1] Peter Achinstein. 1983. *The Nature of Explanation*. Oxford University Press, New York.

[2] Article 29 Working Party. 2018. *Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679*. https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053.

[3] Roger Brownsword. 2019. *Law, Technology and Society: Re-Imagining the Regulatory Environment*. Routledge, Abingdon.

[4] Mireille Hildebrandt. 2015. *Smart Technologies and the End(s) of Law*. Edward Elgar, Cheltenham.

[5] Kieron O'Hara. 1994. *Mind as Machine: Can Computational Processes Be Regarded As Explanatory of Mental Processes?* University of Oxford DPhil thesis, https://eprints.soton.ac.uk/254167/.