February 2020

The data ethics challenges of explainable AI and their knowledge-based solutions

Mathieu D'AQUIN
Data Science Institute
Insight SFI Research Centre for Data Analytics
NUI Galway, Ireland

Abstract. Explainable AI has recently gained momentum as an approach to overcome some of the more obvious ethical implications of the increasingly widespread application of AI (mostly machine learning). It is however not always completely evident whether providing explanations actually achieves to overcome those ethical issues, or rather create a false sense of control and transparency. This and other possible misuses of Explainable AI leads to the need to consider the possibility that providing explanations might itself represent a risk with respect to ethical implications at several levels. In this chapter, we explore through a series of scenarios how explanations in certain circunstances might affect negatively specific ethical values, from human agency to fairness. Through those scenarios, we discuss the need to consider ethical implications in the design and deployment of Explainable AI systems, focusing on how knowledge-based approaches can offer elements of solutions to the issues raised. We conclude on the requirements for ethical explanations, and on how hybrid-systems, combining machine learning with background knowledge, offer a way towards achieving those requirements.

Keywords. Ethics, Explanation, Scenarios, Ethics-by-Design, Knowledge Graphs, Knowledge-Based Systems

1. Introduction, background and related work

Explainable Artifical Intelligence (AI) is a current trend which aims at making the results of AI systems more interpretable by providing "explanations" to justify them. The need for such explanations is more prominently justified in systems used for the purpose of decision making, and has at least been partially driven by recent European regulation requiring automatic decisions to be explained and interpretable [5]. While AI could refer to a wide variety of approaches, we assume here that, in the context of Explainable AI, the technique which results are explained falls in the general category of "data centric" approaches, such as machine learning or data mining, both because they are the ones for which the need for explanation is more obvious, and because they have received a lot of attention in the last few years from academia and industry. This still covers a large number of the applications of AI such as recommendation, prediction and classification, which have been considered within a large number of application domains, including healthcare, finance, retail, media, etc.

As depicted in Figure 1, in such an AI process, as considered here, the main role of the human user (besides the parts that are directly related to designing and implementing

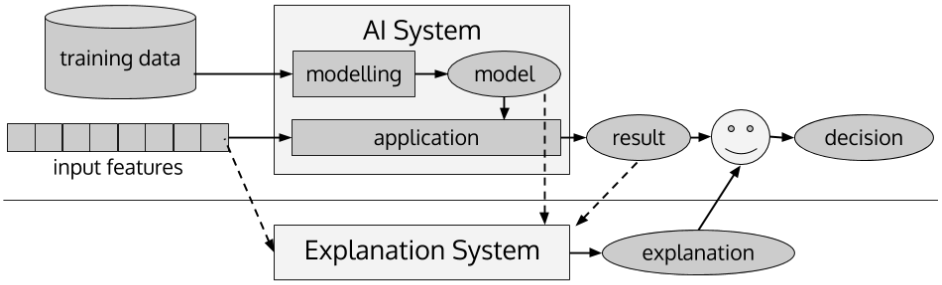

Figure 1. Simplified view of the (data-centric) AI process and the place of explanations.

the process, such as selecting the training data, the model, etc.) is in using the results of the process in order to make a decision. A base definition of Explainable AI is therefore, as also depicted, an extension of this process where the result comes with an explanation enabling the user to make a better decision. We will discuss later in this section what such an explanation can, and should, be, but we start here by looking at the general categories of approaches that exist in the literature to produce them.

Based for example on [10], we distinguish three main types of approaches to producing explanations:

- Using explainable models (e.g. decision trees, or recommendation by explanation [9]): These are approaches where the explaination is generated out of showing directly the inner working of the AI system, which uses interpretable models or has been designed to be inherently explainable.
- Reverse engineering the result (e.g. Deep Explainer [7]): These are approaches that extract from the model and its application salient features that were used to generate the result, as a way to justify this result.
- Reconnecting input and results (path-based explanations and/or sensitivity analysis): In those approaches, the model used by the AI system itself is not used to generate an explanation, but an explanation is generated out of finding credible connections between the input features and the output of the system, or through perturbing the input of the system to show where results might change (i.e. sensitiviy analysis [4]).

Other chapters in this book describe techniques that exploit knowledge graphs to provide explanations and which mostly fall under one of those categories. Here however, we focus on the ethics implications of explanations. There has been much discussion, including in relation to regulation, concerning the ethical aspects of not providing explanations. Here, we therefore rather focus on cases where ethical implications might emerge from providing such explanations. To do so, we rely on the approach advocated in [3,13,12], in particular with respect to the use of anticipatory scenarios to analysis possible outcomes of the deployment of technologies. We devise such scenarios to explore some of the possible ethical implications of deploying explainable AI, and see what role knowledge-based approaches [1] can take in alleviating negative consequences in such scenarios. To do so, we consider scenarios according to four different dimensions:

- The kind of technique used for producing explanations (i.e. the above categories).
- The level of deployment/control (i.e. whether the technology is available to/under control of a selected few, or used by millions).
- Whether the technology is used and operating in the way intended, or is subject to abuses, incorrect or misleading results.
- The particular category of ethical values that the technology might be affecting.

The first dimension was explained already earlier, and the second is self-explanatory. To understand the possible effect of varying the third dimention, it is first important to clarify what explanations are expected to deliver, and therefore what they are. According to [1], the objective of an explanation is to provide the human user, the person or group of persons making a decision, with the ability to assess the correctness, accuracy and adequacy of the result. Looking at it from a more conceptual level, [11] studied the use of the term "explanation" in various disciplines, to provide an ontological view of the foundamental components of an explanation, reproduced in Figure 2. This is relevant since, as will be discussed later in this chapter, several ways in which explanations might not be adequate, including through being misleading, is in situations where some of those components are missing or do not play their expected roles.

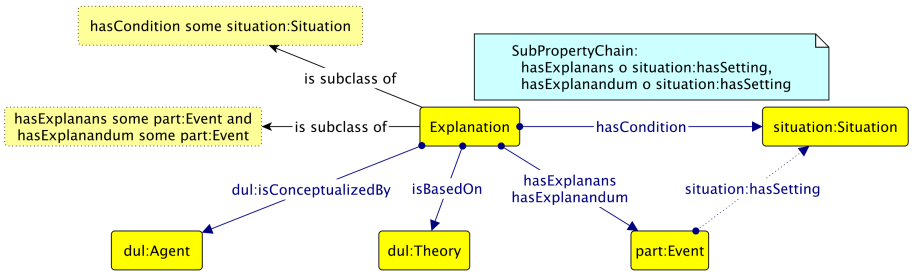

Figure 2. Conceptual model of explanations (from [11]).

Also, as mentioned earlier, to understand ethics implication of a process or technology, we need to look at the specific values that are being affected by the technology deployed, i.e. the fourth dimension in the list above. There are a number of values that can be considered, and that have been used to analyse ethics aspects in a number of domains (see for example [2]). Here, we look at the values represented in the "Trustworthy AI assessment list" (pilot phase) of the "Ethics Guidelines for Trustworthy AI" published by the High-Level Expert Group on Artificial Intelligence of the European Commission [6]: Human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal well-being; and accountability. Those have been chosen specifically to assess ethics aspects of AI technology, and are therefore appropriate to our goal. Each of the following sections looks into scenarios specifically considering the impact of Explainable AI on a particular value in this list. In each section, we present a scenario that illustrates possible ethical issues according to the considered value, extend the discussion to aspects beyond the scenario, and envision possible knowledge-based solutions to the issues raised. While this cannot be considered an exhaustive approach to uncovering possible ethical issues from Explainable AI, it provides a framework for designers, developpers and users of such sys-

tems to explore such issues, which is consistent with the Trustworthy AI assessment list and guidelines cited above.

2. Human agency and oversight

> *As every morning, Dr. Laplace, oncologist, receives the results from the previous day's analyses of his patients' data, together with recommendations for treatment. Each recommendation comes with a justification, based on the genetic profile of the patient, the tumour, as well as results from the latest literature and clinical trials. As always, most of them are quite straighforward and Dr. Laplace just taps "approved" so the treatment can go forward. As always too, a few of them seem, sometimes more than slightly, unexpected. Each justification shows to Dr. Laplace that the recommendation is valid according to latest research, of which she was not aware. She spends a bit more time on those anyway, for good conscience, before approving them.*

Human agency is the capacity of a human being, actor in a process, to act upon and within the environment of this process. The scenario described provides a sense of how explanations, in this case the justifications for the treatement recommendations, can reduce human agency, namely, the capacity of Dr. Laplace to act upon the decision of treatment.

The process in which Dr. Laplace engages can be seen as science fiction, but is far from unplausible, even within the next few years. Precision medicine has made much progress, and many researchers are working on establishing (often knowledge graph-based) approaches to achieve exactly what is described (see for example [8]). In this process, explanations have been introduced as a way to provide transparency, and to integrate the human expert in the decision process, being the ultimate decision maker. Those explanations are however based on much more information than the expert user can integrate. They are not questionable, as the human expert would have to spend enormous efforts to explore all the ramifications of the recomendation and of the associated explanation. We can expect this process to have been put in place as an attempt to reduce the effort, and time, required by the practitioner to decide on a treatment. Spending the same effort in tracing back the reasoning from its automated "assistant" would therefore be counter-productive. In addition, going against the recommendation of the tool would represent a risk for Dr. Laplace. In case of a serious problem with the treatment, justifying the decision as having followed the advise of the tool, which appeared well justified, would be much easier than justifying going against it. Dr. Laplace therefore ends up simply ticking the box, with her involvement being reduced to maintaining the pretence of oversight in a decision which was really made algorithmically.

The kind of technique used to produce the explanation here is not made explicit. It is however obvious that techniques that produce explanations independently from the process that produces the result would create additional issues. Namely, as there is no necessary alignment between the explanation and the inference performed by the AI process, tracing back this process would be immensely more difficult.

The presented scenario is a typical case of a system deployed and used by and for a select group of people. In this case, it seems reasonable, as the decision requires medical knowledge to be properly made/validated. The issue of the ability to question the expla-

nation provided would therefore be even more significant if the process involved users without the required knowledge to understand them (as discussed in Section 6).

Finally, the scenario assumes that the system works in the way intended. There are, of course, many potential issues associated with the possibility that the system could be "hacked", but those are similar to the ones associated with more or less any medical system dealing with patient treatements. Of course, the issues regarding human agency would be even more prominent if the explanations made were inaccurate, or based on false premisses, and no AI system can be considered entirely immune from such problems.

Regarding possible solutions, the main issue here might be seen as coming from the fact that only one recommendation per case is provided, with only one explanation that cannot reflect the whole of the system's inference, but only the positive indicators that have led to this result. In such cases, a solution commonly proposed is for the system to also give indicators of the level of confidence in its recommendation. In order to be effective however, such confidence indicators would have to be interpretable. An approach here would therefore be to provide, instead of one recommendation/explanation/confidence indicator, a network of possible recommendations, together with the positive and negative indicators that provide evidence in favour or against a particular course of treatment. In other words, the explanation in this case is replaced by a specific knowledge graph of the treatment options, genetic markers and research results for the practitionner to explore, in order to make an informed decision.

3. Technical robustness and safety

> *Joining the prestigious Nisachausette Institute of Technology (NIT) has always been a challenge, and many have found the selection process to be unfair and to lack in transparency. Surprisingly, to its administators, setting up an automated selection process which not only explains the reasons for the result, but also publicly shares those explanations (anonymised) for all candidates, has not made people more confident. Jeff already tried three years ago. He got rejected because the "pattern of interactions generated from his use of the online assessment tool did not match the ones of successful NIT students". He spent those three years collecting explanations from all others, analysing the features used, comparing negative cases to positive cases, figuring out what behaviour, engagement, attitude would increase his chances of acceptance. He learned how to fake successful patterns and finally got accepted. He got a great degree, as did all the other "fakers" he met at NIT (they had a secret club).*

Technical robustness and safety include a number of different aspects, including whether the system or process in place can be exploited to achieve something other than intended, as well as, according to [6], whether the system is accurate.

In this scenario, the underlying AI system relies on invalid or irrelevant signals to make a decision, which already shows a lack of robustness. The focus here however is on how the explanations add to the issue by emphasising the inaccuracy and the use of spurious signals in the decision making process, and on how the availability of large amounts of explanations can help individuals with the right understanding and skillset to

exploit those inaccuracies to their advantage. Of course, the scenario has a happy ending, with Jeff and others getting their degree at NIT. We could even imagine the head of NIT knowing about the practice and relying on a principle that if someone is clever enough to game the system, they deserve a place. This however remains an issue as it amounts to pretending that the profound lack of robustness of the system is a feature, the "real test", even though other candidates would be reluctant to use the same approach as Jeff, considering it cheating.

While this scenario might appear quite specific, it is easy to imagine others where explanations, especially the availability of multiple explanations, enable users to reverse-engineer the model used to make a decision, so to exploit its weaknesses. This, of course, can only be applied in cases where the explanation is directly related to the inner-working of the model, by being based on the model's actual inference or on extracting the most salient features used in the decision.

This scenario also expects a system available to a large number of people, where there is a way to collect information about the inner working of the model, through having access to multiple explanations. This can be achieved either through accessing explanations from multiple users (as in the scenario) or in cases where the user is not limited in the number of inputs they can enter, by systematically exploring the space of inputs, to analyse and map both the space of outputs and the space of explanations.

Finally, the case of Jeff and NIT assumes that the AI system is inaccurate and that the explanation system is being exploited in a way different from the one intended. It therefore goes against robustness and (in a very broad sense) safety at two levels. We could easily imagine however a way in which the explanation system could be used in exactly the right way, but through explaining results based on irrelevant, biased data, would end-up strenghtening those results, and therefore justifying the biases embeded in the data.

A key part of the reason why the explanation system fails in this and similar scenarios is that it provides an incomplete explanation. Indeed, it explains the link between the input of the AI system (the explanans, see Figure 2) and its output (the explainandum), but without providing grounding for it (the theory). There is no valid relationship between the two components as the model relies on spurious signals learned from historical data. In other words, the explanation amounts to nothing more than saying "you got those results because the system used those features". A solution to this would be to not only construct explanations from tracing back the behaviour of the systems, but by also intregrating external knowledge that can justify this behaviour and ground the signals on which the system relies. In other words, the explanation should rather be "you got those results because the system used those features and there is a known, causual relationship between those features and what the system is trying to predict" (in our case, whether the candidate will succeed in their studies at NIT). A knowledge-based solution to integrating those "known relationships" would consist of building a knowledge-graph of existing empirical results in the considered field from the scientific literature (in our case, robust studies of the abilities and characteristics of candidates that are more conducive to success as a student). Of course, a consequence of this would be an increased ability of the explanation system to discard spurious, ungrounded results from the AI system (as in the case of Jeff), by recognising explainations that are not justified by knowledge of the domain.

4. Privacy and data governance

Jane didn't really know what to whatch when starting Webflux. She really enjoyed the last movie she watched, "The 12 Zebras", as did her friend Nat. Nat seems to be the only one to share her tastes arround here. Nobody else they talked to in their small village had seen that movie. Most of them don't even have Webflux. But then, the first recommendation it comes up with now is a movie called "Fifty nuances of Purple". Jane checks the explanation provided and it says "This movie was watched and highly rated by someone who also watched "The 12 Zebras" in your neighbourhood." That would have to be Nat then, but Jane wonders why she didn't mention that movie when they talked about what to watch next.

Privacy and data governance relate mostly, in the context of this chapter, to the protection of personal data from unintended disclosure.

In this scenario, it appears that Nat, Jane's friend, did not want her to know that she watched and enjoyed a particular movie, or at least, the movie streaming service should not be revealing it without her knowledge and consent. This case might seem trivial, especially compared to scenarios in healthcare, education or finance, but it simply illustrates how explanations, by revealing more information about the inner working of AI systems, increase the risk of disclosure in cases when personal data is being used by those systems. We therefore expect the issue to be applicable to many other scenarios, especially in collaborative-filtering, hyperlocal and/or highly personalised recommendations.

Interestingly, the issue with disclosure of personal information might appear whether or not the information is actually used by the AI system. Indeed, the issue here is that the explanation system uses specific information that might lead to accidental disclosure. In other words, the explanation system can be treated entirely speparatly from the AI system as one that manipulates personal data, and which outputs, even though aggregated (at least in the case of our scenario), could reveal more information than intended.

The scenario relies on a case where a large number of people use and have access to the AI system and the explanations. This is necessary in this case for the personal information being disclosed to actually be available in the system, and for the expectation of anonymity brought by aggregation to be present. We could however also imagine other scenarios where explainable AI systems trained on large amounts of personal data from many individuals are used by only a few selected people (because of addressing highly specialised tasks) and still lead to similar accidental disclosures.

Also, whether the explanation system works as expected or not appears irrelevant here since the issue of accidental personal data disclosure is independent of the objectives of both the AI system and the explanation system. One could argue that, if considering anonymity and data protection as requirements, then the explanation system is not operating adequately, which leads to the main difficulty in addressing the issue –contradictory requirements– as discussed below.

While the scenario and the issue might appear trivial, finding a solution may be surprisingly complicated. Indeed, one could simply treat the personal data in input of the explanation system through any existing anonymity filter. Obfuscating the data about Nat's rating or location using k-amonymity for example would resolve the issue in our scenario. However, it would also have rendered the explanation inaccurate. According to European regulation, directly or indirectly identifiable information provided to a system

cannot be disclosed to a third party without explicit, direct consent from the data subject, and without justification of purpose. The same regulation also states that individuals have a right to be given an explanation for the output of an algorithm. Those requirements could end-up contradicting each other when the only way not to indirectly disclose identifiable information would be to produce an explanation that does not actually reflect the inference made by the AI system[1]. Balancing those two aspects, anonymity and accuracy, might end-up being a complex challenge that could involve nuanced notions of user preferences, information sensitivity, user prior knowledge, etc. This forms a complex network of varied elements which knowledge-based approaches can help manage, enabling potentially complex reasoning.

5. Transparency

> *It is the third time Claire submitted her research grant proposal to the Wakandian Science Foundation (WSF), and the third time it gets rejected by the automated proposal pre-selection process. This AI system was trained on hundreds of accepted and rejected grant proposals, relying on all aspects of the applications, from the text of the research plan to the proposed budget, so to predict the likelihood of a given proposal being accepted. Proposals that fall below a certain treshold are automatically rejected, without a human being reviewing them. The objective was to reduce the ever-increasing workload associated with assessing the ever-increasing number of proposals submitted. Short feedback is provided on the rejected proposals, justifying the rejection, based on reverse-engineering the complex model making the prediction to extract salient features and connect them with feedback provided to past rejected proposals from the training set. The first one Claire got was "Budget too high", so she reduced the budget. The second one was "Insuficient community engagement", so she added workshops with community stakeholders. The third piece of explanation Claire received was "Insufficient resources to carry out workplan". She is now lost and gives up on what would have been a very impactful project.*

Transparency, as a characteristic of a system, corresponds to the system being open and clearly communicating on the processes it carries out. To a large extent, explanations in AI are assumed to support increasing transparency.

In the scenario above however, it is less than clear that this is the case. We can assume that, to function, the AI system put in place by the WSF would rely on a large number of rich features. It would include basic numeric features, such as the amount of funding requested, the duration of the proposed project, the number of tasks and workpackages, the number of partners and the size of teams. It would likely also include structured information about the workplan and the applicants, as well as embeddings of texts and other media included in the different sections. As a result, the outcome of the prediction would be based on inferences using evidence from thousands of those features. Actually accounting for the entirety of such inferences would require an explanation almost as complex as the model itself, and therefore entirely unsuable. The approach taken here, which is common in current explainable AI research, is to reduce the explanation to the

[1] We would of course assume that, in our case, data protection would take priority over the right to explanation.

most prominent (i.e. salient) features. While this can work well in cases where specific features have a significantly stronger contribution to the result than others, we can expect that for Claire's proposal, it was not the case. In other words, while budget, engagement and workload could have been marginally more salient than others, the agregation of potentially thousands of other, slightly less salient features was actually the main driver for the outcome. The explanation system is an algorithm that outputs an estimate of the likely explanation, and is itself a black-box. It only gives the illusion of transparency, while actually further obfuscating an algorithmic decision.

This scenario is based on the assumption that the explanation system uses a method based on reverse-engineering the process captured by the model used by the AI system, to extract salient features that can be used to come up with realistic "feedback" according to past proposal review reports. Since this is done to avoid actually describing the inner-working of the AI system, it could easily be imagined that the explanation is constructed entirely indenpendently from the AI system, i.e. that the feedback would be lifted from similar rejected proposals, without even attempting to connect this to features used by the AI system. The same issue would obviously appear in this case. Using an explainable model would, on the other hand, avoid the issue of transparency. The scenario however assumes that the problem is so complex that explainable models are not applicable. Applying sensitivity analysis could also help giving a more straighforward explanation, but again, the complexity of the model/problem would likely mean that too many features are needed to be considered. Assuming it was feasible however, sensitivity analysis would transform the explanation into something that could be interpreted as "your proposal would have been accepted if you had done X, Y and Z." Besides the list of features to change and the list of possible alternative explanations being potentially very large, this would open up to issues such as the ones illustrated in the scenario of Section 3.

The particular issue described here, explanations giving a false sense of transparency for a process that is too complex to be transparent, can appear whether the system is used by millions, or only by a few people. Also, the issue appears whether or not the explanation system works as designed. It is clear however that it is an issue related to the fact that the explanation system is not designed to achieve what an explanation system should deliver: A valid, accurate representation of the process leading to the AI system's result.

It is not clear whether the problems described above can really have a solution. However, understanding the relative importance of features used to build more nuanced explanations, as well as their connection with the actual output (i.e. addressing the "theory" component of the explanation) can help alleviating those problems. In other words, as with the human agency scenario (Section 2), providing more knowledge about the aspects, entities and indicators contributing to the result, and about their connections, while not removing the issue entirely, can support the human user making the final decision in taking into account the known incompleteness of the explanation. That is however if, unlike in our scenario, a human indeed takes the final decision.

6. Fairness, societal well-being and accountability

Robert's company was going well. His platform combining Internet of Things technology with robotics and community engagement had been deployed and was now

working for several associations and local councils to support communities in some of the least favoured areas of the city, helping them overcome some of the key reasons for the issues they were facing. Also, he had just found the pefect property where to move his family: A nicely sized 4 bedroom house in one of the areas his company had helped turn arround. He believed it was going to work. Obtaining the mortgage was not supposed to be an issue: Even if he was not very rich, his status of CEO of a successful company should make him a great candidate for the bank. His application was nevertheless rejected. The explanation that came with it was rather unclear about the reasons. It talked about the projected value of the property at short term, based on extrapolating over historical data, being expected to continue declining, and that features of the profile of the applicant (occupation, career prospect) and of the property (history of occupation) leading to a low confidence in a successful relocation. Despite his principles and beliefs, Robert follows the recommendation of his bank advisor to move to a "more suitable" part of the city.

In this scenario, we consider three interelated aspects. First, on "Diversity, non-discrimination and fairness", which relates to the system gathering equaly for all users, and not favouring some users over irrelevant features, the most evident issue is that the explanation in the scenario requires mastering a vocabulary and understading a domain which only a few of the users would really know. This appears unavoidable, since the decision itself is based on features in this specialised domain (finance, the property market, etc.) on which the system cannot expect all users to be experts. While the explanation is supposed to provide a way for the user to question the reasons for the decision, it instead leads to even more confusion. The user is left with no ability to question the explanation itself. There is no oportunity for disagreeing.

Second, this scenario also strongly relates to "societal well-being". Indeed, digging a bit deeper, Robert might have noticed that both explanations provided are based on assumptions that relate to the way the area has been considered in the past and to the way people sharing his profile have acted in the past. More explicitly, the system has rejected his application because, not being aware of the interventions from Robert's company, it considered the area of the property as a risky place for investment and unlikely to be suited, as a long term home, to someone with a CEO profile. Those represents systematic biases: Unfair assessements based on generalising over aggregated, past data, not taking sufficiently into account specific cisrcunstances and additional variables affecting future prospects. While this is unrelated to the presence of an explanation, since those biases would be there anyway, the explanation fails to uncover them. Instead of clearly demonstrating to Robert how the decision was based on wrong assumptions, enabling him to question it, it justifies the biases in a way that make them appear more valid than they are. Since they drive investment and relocation decisions, those biases in this specific scenario would actively contribute to slowing down improvements in the considered area of the city.

Third, the scenario also relates to issues of "accountability". Indeed, as discussed above, it appears here that the decision is based on reasons that do not apply well. We can imagine that, reviewing those kinds of decisions later, the bank might realise that a mistake was made, and that properties in the area considered by Robert were being wrongly assessed. Ideally, we would assume that decisions are not made entirely automatically and that a human expert validates them before presenting them to the user. Indeed, in the scenario, Robert deals with an advisor who is supposedly there to help him

with his decision. However, the advisor in this case instead takes the role of supporting Robert in finding an alternative that would be more suitable to the system, rather than in understanding the decision in a way that can make it questionable. Similarly to what is described in Section 2, while the advisor should be accountable for the decision, the presence of a "categorical" explanation for the decision makes it even less possible to go against it. In the event of a review, it appears easier for the "accountable" advisor to justify his decision based on the explanation of the system, rather than taking the risk of contradicting it.

The particular technique used to come up with an explanation here is not very relevant to the issues themselves. The scenario makes assumptions about the explanations being strongly related to the actual features used by the AI system, suggesting a "reverse engineering-style" explanation, which can be seen to be at the origin of its lack of clarity. The issue of explanations masking implicit biases on which the decision relies is however even more likely to happen when explanations are not directly based on analysing the process and features applied by the AI system.

The aspect of whether the system is used and controlled by many is interesting here. The particular scenario presented is one of a system used by many, but controlled by a few. It is easy to imagine many others configurations where societal well-being and/or fairness are affected by the aspect of who controls the system, and by how it is expected, and failing, to cater for a wide variety of users.

Finally, it is clear that in this scenario, as in Section 5, while the explanation system does what it is designed to do, it is not designed to truly achieve the purpose of explanations. Instead of providing a way to understand the decision, see when they might be misguided, and question them, they achieve making them more obscure and less questionable.

As mentioned above, the decision here is not entirely automatic, and is partly handled by a human advisor who should be able to interpret the explanation based on their own backgound knowledge. This represents an obvious element of solution: In cases like this one where the features and decisions cannot easily be explained without referring to specialised background knowledge, the introduction of background knowledge is required. An explanation could indeed be made more accessible to the final user, and possibly help them understand it enough to detect biases and invalid assumptions, if reconnected to background knowledge that is accessible to them. The issue is that explanations are, in this kind of cases, framed within a particular area of background knowledge that requires a certain amount of expertise (finance, property market). Making such specialised background knowledge explicit, also explicitly encoding relevant background knowledge from the user (i.e. for Robert, what his company has been doing, his objectives, etc.) and aligning those two related but differently framed knowledge domains could help formulating explanations produced based on the features of the former in a way that is meaningful to the later.

7. Conclusion: Towards a more ethical, knowledge-based Explainable AI

There is a lot of expectations associated with Explainable AI, which are strongly related to the current trend that applications of AI rely on complex, fundamentally hard to inter-

pret models, based on picking and extrapolating from sometimes counterintuitive signals from large amounts of data. The scenarios presented above however illustrate some cases in which, if not considered carefully, the application of explanations with those models could have unintended consequences, often counter to the original purpose of explanations. There are, naturally, many other scenarios that could be considered, addressing the same or other, possibly more precise values. It is obvious already from the scenarios presented and the associated discussions that those issues are interelated, and that problems with repsect to specific values are often strongly connected to problems regarding others. For example, the scenario for human agency could be similarly used to discuss inclusion, societal well-being, accountability, etc.

In relation to this, while the objective of this chapter is to raise awareness of the need to consider the potential impact of explanations and not to assume that the mere presence of explanations is sufficient to achieve transparency and human agency, the scenarios above also show some common traits with respect to the way issues come about in the application of Explainable AI. In particular, several of the issues come from the explanations not being sufficient to provide a complete, questionable view of the decision made, or are not sufficiently interpretable themselves to truly enable informed decisions.

For this reason, we extract two main conclusions from this exercise in looking at Explainable AI from the point of view of ethics values. First, there are a number of properties that explanations need to have in order to be effective and to reduce the risk of unintended consequences. Namely, explanations should be:

Complete: As discussed in [11] (see Figure 2), an explanation requires to include not only some form of correlation between what is being explained (the decision) and what explains it (in many cases, the features used). It needs to explicitate what relates those (the theory), i.e. by which mechanism or principle those features actually impact on the decision.

Complete again: In several of our scenarios, a part of the issue is that alternatives to the decisions are not presented, and the reasons for them being discarded are not explained. In other words, the results are only positively explained, and do not present the whole background for the decision. In some cases of course, the reason for this is that the whole justification for the decision is simply too complex to be of any use as an explanation (see Section 5).

Honest: In order to serve their purpose, explanations should accuratly capture the way in which the AI system has produced its results. As shown in several scenarios, oversimplifications or indirect explanations can, paradoxally, end up being more misleading than the absence of explanation.

Understandable: While this might appear obvious, the purpose of explanations being to make a complex result interpretable, as illustrated by several scenarios, it is not trivially achieved. Indeed, to be honest, explanations might have to be complex, and by being complex and referring to processes and entities with which the user cannot relate, fail in providing any added value.

The second level of conclusion we reach is that, in many cases, to move towards achieving the properties above, it is required to integrate some elements of a knowledge-based approach. Indeed, the main issue of Explainable AI, as applied to machine learning, is that there is a wide gap between the numerical, complex, connected methods implemented by those approaches and the knowledge of the user. Interpretability is founda-

mentally the ability to integrate new information (the result of the AI system) within an existing knowledge framework (the one of the user). As discussed in the sections above, possible solutions to improving the interpretability of the results of AI systems beyond "basic" explanations therefore involve mapping such explanations, the entities and the processes to which they relate, with knowledge represented in a way that makes it manipulable and integratable by the user. This could provide a layer above the low level features on which explanations rely, bridging the gap between the elementary, numerical operations of machine learning and the understanding of the results produced, so to support informed, intelligent decision making.

Acknowledgement

This work has been partly funded by Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2, Insight SFI Research Centre for Data Analytics.

References

[1] Rajendra Akerkar and Priti Sajja. *Knowledge-based systems*. Jones & Bartlett Publishers, 2010.

[2] Sally Bean. Navigating the murky intersection between clinical and organizational ethics: A hybrid case taxonomy. *Bioethics*, 25(6):320–325, 2011.

[3] Mathieu d'Aquin, Pinelopi Troullinou, Noel E O'Connor, Aindrias Cullen, Gráinne Faller, and Louise Holden. Towards an ethics by design methodology for ai research projects. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 54–59. ACM, 2018.

[4] Andries Petrus Engelbrecht, Ian Cloete, and Jacek M Zurada. Determining the significance of input parameters using sensitivity analysis. In *International Workshop on Artificial Neural Networks*, pages 382–388. Springer, 1995.

[5] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3):50–57, 2017.

[6] High Level Expert Group on AI, European Commission. Ethics guidelines for trustworthy ai, 2019.

[7] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222, 2017.

[8] Peipei Ping, Karol Watson, Jiawei Han, and Alex Bui. Individualized knowledge graph: a viable informatics path to precision medicine. *Circulation research*, 120(7):1078–1080, 2017.

[9] Arpit Rana and Derek Bridge. Explanations that are intrinsic to recommendations. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, pages 187–195. ACM, 2018.

[10] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*, 2017.

[11] Ilaria Tiddi, Mathieu d'Aquin, and Enrico Motta. An ontology design pattern to define explanations. In *Proceedings of the 8th International Conference on Knowledge Capture*, page 3. ACM, 2015.

[12] Pinelopi Troullinou, Mathieu d'Aquin, and Ilaria Tiddi. Re-coding black mirror chairs' welcome & organization. In *Companion Proceedings of the The Web Conference 2018*, pages 1527–1528. International World Wide Web Conferences Steering Committee, 2018.

[13] Pinelopi Troullinou and Mathieu d'Aquin. Using futuristic scenarios for an interdisciplinary discussion on the feasibility and implications of technology. *Black Mirror and Critical Media Theory*, page 69, 2018.