Ethics of AI and Explainability

Stefan Buijsman

TU Delft

Philosophy and XAI

- There are, roughly speaking, two major debates in philosophy that revolve around XAI
- First: is explainability of AI systems a requirement for them to be trustworthy?
- Second: what does a (good) explanation look like?
- Let’s take these questions in turn

Philosophy and XAI

- As a case study: NarxCare is an algorithm that informs doctors of the risk that a patient has opioid addiction
- Doctors are strongly incentivised to use the algorithm, to the point where they can lose their license if they don’t demonstrably use the AI system

Philosophy and XAI

- Another example: the UWV (Dutch government organisation responsible for social benefits) wants to use AI to verify whether unemployed people are applying to jobs effectively. They'll then show the results (the people judged to be least effective) to case officers

- Their reasoning is that as little information as possible should be provided to case officers (e.g. no precise scores etc) to prevent biasing them in their follow-up discussions. They'll add in some random cases to ensure that case officers can't fully trust the model

TU Delft

Why XAI?

- If we're looking at the philosophical literature, however, then the discussions are more abstract. Mostly, they focus on whether there should be a general requirement of explainability, and the argument goes that accuracy is enough for trustworthy AI. I'm skeptical, but it's useful to debate their arguments

- To start with, Robbins (2019) argues against the need for explainability methods in using black box models. From the abstract: "a principle of explicability for AI makes the use of AI redundant."

- He presents two arguments against a requirement that AI is explainable

Why XAI? Robbins

- Against a blanket requirement of explainability on AI models:
  What requires explanations are decisions/actions, not processes. And we only care about explanations when the consequences of those actions significantly affect someone

- For example: do we need explanations for an ML algorithm that performs highly accurate quality control in a factory? (as compared to e.g. an algorithm making hiring decisions)

- When low risk (another example is an AI that detects signs of cardiac arrest on emergency phone line) we should thus, so Robbins argues, adopt opaque AI of sufficient accuracy

Why XAI? Robbins

- The second argument is a ‘Catch 22’ on the use of interpretable ML:
  “If ML is being used for a decision requiring an explanation, then it must be explicable AI and a human must be able to check that the considerations used are acceptable. But if we already know which considerations should be used for a decision, then we do not need ML.” (p.509)

- “The explanation might be something like “the applicant has a low debt to savings ratio and a high income to rent ratio”. These both seem to be relevant considerations when deciding whether to accept a loan application. … [But] if we already know which considerations are acceptable, then there is no reason to use ML in the first place.” (p.510)

- So, the idea is that we can only verify if an ML output is correct/appropriate if we can make the considerations on which that decision is based explicit into rules that could fuel GOFAI

Why XAI? Robbins

- One optional role for explanations: pointing to relevant features we were unaware of

- “For example, if a medical diagnosis algorithm used as a consideration that the patient’s eyes were a very specifc color, we would not immediately be able to tell if this was an acceptable reason or not.
  This may cause us to test the hypothesis that this specifc eye color was strongly correlated with the diagnosis. If this eye color is indeed indicative (to a medically signifcant level) then the algorithm’s explanation would have contributed to the scientifc and medical community by coming up with a consideration we had not thought of before.
  This consideration can now be used by GOFAI and/or doctors to make future diagnoses.” (p.510-511)

Why XAI? London

- A guiding idea in the arguments of Robbins is that an algorithm shouldn't be required to explain (certain) decisions if it performs better than humans. London (2019), 'Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability', similarly pursues this idea that accuracy is what matters

- One reason: in the medical context, London argues, experts also often act on uncertainty and incomplete information and have to rely on empirical findings that a treatment is effective

- "Modern clinicians prescribed aspirin as an analgesic for nearly a century without understanding the mechanism through which it works. Lithium has been used as a mood stabilizer for half a century, yet why it works remains uncertain. Large parts of medical practice frequently reflect a mixture of empirical findings and inherited clinical culture." (p.17)

Why XAI? London

- Furthermore: treatment based on (presumed) theoretical understanding is not guaranteed to be better

- “The long medical preference for radical mastectomy over less aggressive alternatives was driven by the pathophysiological theory that removing as much tissue from the breast as possible would reduce the probability of cancer recurrence. Only after a series of clinical trials was this theory shown to be false.
- The same is true for the theory of drug action that drove the use of high-dose chemotherapy with autologous bone marrow transplant as a treatment for end-stage breast cancer. In such cases, the overreliance on plausible theoretical explanations lead to treatment practices that harmed patients and consumed scarce resources precisely because key causal claims in those theories were false.” (p.18)
- “Clinicians, therefore, frequently make judgments about how comorbidities, gender, ethnicity, age, or other factors might affect intervention efficacy and toxicity that go beyond validated medical evidence” (p.18)

Why XAI? London

- In a way, the argument goes as follows: we shouldn't have to explain AI, because people aren't able to provide these explanations either. London primarily points to our lack of understanding of underlying causal mechanisms in treatment choice

- Trust in AI should therefore come from proven reliability/accuracy

- Is there a relevant difference between being unable to explain the efficacy of a treatment method plus the degree of opacity of our own decision making, and the features of black box algorithms?

Why XAI? Durán & Jongsma

- Instead of criticizing XAI by saying that people aren't able to explain either, an alternative strategy is to argue that for us accuracy is enough to trust/acquire knowledge
- That's the main goal of Durán & Jongsma (2021), 'Who is afraid of black box algorithms?', though they start with a conceptual puzzle for XAI: the problem that according to them transparency/explainability is itself based on an opaque process

- Suppose we have an interpretable predictor P that explains algorithm A. P explains A if it correctly answers the question 'why does A output a in this case?'
  But, Durán & Jongsma argue, there is then a question of how P represents the inner workings of A and the opacity of P itself. For someone to believe that A has been explained, it must also be explained how P works and why it correctly represents A
- "transparency displaces the question of opacity of *A* to the question of opacity of *P*, taking the latter as non-problematic. But *P* is, strictly speaking, still opaque" (p.3)

Why XAI? Durán & Jongsma

- They then focus on reliability in an account known as Computational Reliabilism, as a way to establish trust even if the algorithm remains opaque

- This stems from a wider account in epistemology known as Reliabilism, which argues that we are justified to believe something if it is the result of a reliable process

- Similarly, they hold that an AI that is reliable (which they understand in a broader sense than just accuracy, as they include expert knowledge, verification and validation, a history of succesful implementations and more) is one that can be trusted. We are then justified to believe what the AI tells us, and might know e.g. the diagnosis

Why XAI? Durán & Jongsma

- So for computational reliabilism: one interpretation is that as long as the AI is in fact reliable (in their more substantial conception) then we can get justified beliefs from it

- That contrasts with the claim that establishing reliability is sufficient for establishing trust in AI systems. I.e., that users/decision makers should be given information about reliability and attend to that

- We could be even less strict as proper reliabilism doesn't care about what users are aware of. If it's in fact reliable, then it's good. If it isn't, despite evidence to the contrary, then you're out of luck

Why XAI? Krishnan

- One way to discuss the over-arching question ‘why are we engaging in XAI?’ is to look at the different goals that interpretable AI/ML might fulfill. Durán & Jongsma, for example, look at what is needed to be justified to believe the output of AI. London looked more at Trust.

- Krishnan (2020), ‘Against Interpretability: a Critical Examination of the Interpretability Problem in Machine Learning’ goes through all of these

- Justification: as with Durán & Jongsma there is an appeal to reliabilism. That’s supposed to show that we have no need for further information about an output to be justified in believing it
  - As opposed to evidentialism, where we need reasons. Though for evidentialists ‘the AI output was x’ might be a reason to believe x without further information

Why XAI? Krishnan

- Anti-discrimination: Krishnan sees two reasons one might call for interpretability here
  - Spotting bias: measures of fairness can be used even if the algorithm is opaque
  - Diagnosing the source of bias; her response here is that problems often stem from the data and this dataset is not opaque

- Reconciliation problem: weighing AI outputs against other information to reach an overall conclusion
  - She argues that the features the AI tracks are what matters (shared features increase confidence in aptness, different indicators strengthen belief if conclusion is the same)
  - “scrutiny of the content of training data sets and ways of testing classifiers to see what features they actually track are viable ways of extracting this information without scrutiny of the steps that the algorithm performs in arriving at categorizations” (p.498)

Why XAI?

- Are there other goals we might have in mind when building XAI tools?

- Krishnan: use of AI by scientists to uncover causal mechanisms/generate scientific explanations
- Krishnan: Public trust (though she suggests this might only be true due to experts stressing the importance of interpretability)

- I think that it is still an open question whether explainability is a requirement or just a nice thing to have in AI systems. We can, for example, consider other goals such as:
  Accountability
  Control
  Contestability

TU Delft

Explanations from philosophy

- Still, ideally we'd like to make model behaviour more interpretable. So can we say something about what succesful explanations would look like?

- The philosophy of science has a long-standing research topic on the nature of explanation, where they are interested in what information is essential for scientific explanations, such as given in physics

- A range of accounts exists
  - Causal / Interventionists
  - Unificationist
  - Mechanistic

TU Delft

Explanations from philosophy

- Before diving in, however, there is one fairly common view on what explanations are: they are answers to *contrastive* why-questions (from Peter Lipton, ‘Inference to the Best Explanation’, 1991)

- So when we ask, for example, ‘why did you throw your plate on the floor?’ we always have a contrast in mind

- If the answer is ‘because I was full’, then the wrong contrast was picked up on

- Contrasts help us narrow down what aspect we want explained, but are often left implicit because we can pick them up from context

Causal explanations

- One popular idea is that explanations are *causal*, i.e. to explain a phenomenon all you really need to know is what caused it. This nicely explains e.g. why explanations are asymmetric:

- The flagpole has length *f* because its shadow has length *s* and the sun strikes the flagpole at angle *a*

- The shadow has length *s* because the flagpole has length *f* and the sun strikes the flagpole at angle *a*

- Miller (2019), in a survey of explanation in social sciences, also hints at this when he defines “interpretability of a model as: the degree to which an observer can understand the cause of a decision”

Causal explanations

- On the most prominent version of this (Woodward, 2003) we want to explain why variable Y has value y by appealing to a cause X
- “An explanans E will consist of:
  - (a) a generalization G relating changes in variables X and changes in Y
  - (b) a statement that the variables X take values *x*

- E is minimally explanatory iff (i) E is approximately true,
  - (ii) According to G, Y takes value *y* under an intervention in which X take values *x*
  - (iii) there is some intervention where X take values *x’*, with G correctly describing the value *y’* that Y would assume, with *y’≠y”*

TU Delft

Causal explanations

- For Woodward, we have two parts of an explanation:
  - A rule covering the actual case
  - A counterfactual also covered by the rule

- Furthermore, there is a standard on how good an explanation is:
  - The more what-if-things-had-been-different questions can be answered with the same explanation (i.e. the more counterfactual situations are covered), the better the explanation

Counterfactual explanations

- How does this relate to counterfactual explanations?

- The focus here is on *what-if-things-had-been-different* questions, i.e. explanations that also tell you what the model predicts in a range of counterfactual cases

- It thus differs from current work on counterfactuals in XAI by
  (1) using a generalization G to link the different counterfactuals together
  (2) including a contrast class

Abstraction

- Generally speaking, we prefer explanations with more abstract variables:
  - The pigeon pecked because it was presented with a scarlet stimulus
  - The pigeon pecked because it was presented with a red stimulus
- As long as both are correct, we tend to see the explanation using ‘red’ as better

Abstraction

- Why is abstractness a good thing?
  More abstraction allows us to answer more why-questions

- Here, variable x_1 is more abstract than variable x_2 when: ”the actual value of x_1 is implied by the actual value of x_2”

- If something is scarlet, it is also red.
  If something is a fridge, then it’s also a kitchen appliance, etc.
- Similarly: the input variables of a model may be far less abstract than e.g. concept-based explanations. So, *G(X) = b(X)* can score poorly on abstractness

Abstraction

- Yet there is also a limit to abstractness:
  - The pigeon pecked because it was presented with a red stimulus, or provided with food, or tickled on the chin, or its cerebellum was electrically stimulated
- Not a great explanation to get, so what's wrong?
- The suggestion from the philosophical literature: this explanation is not specific enough
- Specificity here means: you can change the values of some variables in the explanation (e.g. the tickling, when the pecking is due to the colour) without changing the output value
- There is thus a notion of *relevance* of information: having (sufficient) influence on the result

TU Delft

Generality

- Abstractness is helpful because it makes explanations more general. Yet what type of generality is important exactly?
- There are at least two relevant considerations then:
  - The breadth of a generalization
  - The accuracy of a generalization
- The more inputs an explanation covers, the better (so the idea goes) Additionally: the number of black boxes to which it applies. *G(X) = b(X)* applies only to specific model *b*
- On the other hand, an explanation pointing to more general features such as bias or adversarial cases as the cause for model behaviour applies to far more than a single model
- Counterbalance with accuracy: whether the predicted output is actually correct

TU Delft

Unificationism

- Interventionism / counterfactual accounts such as that of Woodward are not the only extant accounts of scientific explanation in the philosophical literature

- Philip Kitcher (1989) is the main proponent of this theory where: scientific explanation is a matter of providing a unified account of a range of different phenomena

- For example, consider Newton's theory of gravitation: it was a successful explanation, so Kitcher reasons, because it was the first theory to give a unified account of (falling) motions on Earth and of motions in the heavens

- One big challenge here is to make precise what is meant by *unification*

TU Delft

Unificationism

- Kitcher does this by talking about argument patterns that can be used to (deductively) derive as many different phenomena as possible

- To build this up, he starts with schematic sentences, which are sentences in which some of the nonlogical vocabulary has been replaced by dummy letters. For example, from: “Organisms homozygous for the sickling allele develop sickle cell anemia” to “For all *X* if *X* is *O* and *A* then *X* is *P*”.
- Filling instructions are directions that specify how to fill in the dummy letters in schematic sentences

- Argument patterns then are ordered sets of schematic sentences, filling patterns and clarifications stating which sentences are premises and which conclusions

TU Delft

Unificationism

- “Science advances our understanding of nature by showing us how to derive descriptions of many phenomena, using the same pattern of derivation again and again, and in demonstrating this, it teaches us how to reduce the number of facts we have to accept as ultimate”

- So, the idea about generality is already embedded into unificationism from the start, and while it doesn’t mention abstraction per se, it is one way to further limit the number of required argument patterns.

- Main difference with Woodward: there is no focus on causality/counterfactuals, reasoning is done (primarily) deductively

Unificationism

- What about asymmetry then? Earlier we had the example of a flagpole and its shadow, where we don't explain the length of the flagpole based on the length of its shadow. Unificationism has no principled reason against this

- Kitcher suggests we don't explain lengths in this manner because of a single "origin and development"(*OD*) pattern of explanation, according to which the dimensions of objects-artifacts, mountains, stars, organisms etc. are traced to "the conditions under which the object originated and the modifications it has subsequently undergone" (p.485)

- We then need similar patterns to account for e.g. the intuition that we don't explain the velocity of a ball by the breaking of the window it hits, and so on

Durán

- Durán (2021) makes a case for these kinds of (scientific) explanations for AI, he calls it bona fide sXAI

- For instance, “*why* does this medical AI suggest a dose of 50 mg of aspirin for a headache?”

- Should be answered by: “inferential dependence between prostaglandin hormones and swelling, and the effect of aspirin in stopping the production of prostaglandin.
  That is, a supported or confirmed scientific explanation needs to be supplemented with objective relations of dependence in the system, such as the causal linkage between prostaglandin hormones and stopping swelling (e.g., through showing the acting biochemical components).
  Further considerations, such as the dangers of administrating aspirin to pregnant and breastfeeding women, must also be included in the explanation.”

Durán

- So note what is missing compared to current XAI tools: a connection of the AI output to scientific theory (he writes in the medical context)

- He criticizes current XAI tools for not answering the *why* question, but only answering a *how* question: “there has been a long-standing interpretation of classifications as explanation, when these should be kept separate”

- For example, on the single counterfactuals: they “do not render information as to *why* this individual has a risk score of 0.5, but rather what needs to be altered in the variables of the algorithm to obtain the desired score. In other words, the counterfactual explanation informs *how* a given output was obtained, not *why.* Whereas in the latter case, we are demanding objective relations of dependence, in the former, we expect to convey the computational mechanisms that lead to a given output.”

Durán

- One way to see the difference: Durán wants explanations to cover the way decision depend on the situation in the real world (e.g. why the recommended medicine is appropriate for a given diagnosis)

- As opposed to explanations focussing on the way AI outputs depend on the inputs (e.g. why the algorithm recommends a treatment given the input, which is separate from scientific theory)

- He opts for unificationism as the answer to how AI output integrate with wider (scientific) knowledge about e.g. medicine, or whatever domain the task is in
- But note that unificationists can focus on just the AI: argument patterns that cover as many outputs as possible. There's no link to scientific theories required

Mechanism

- As for broad accounts, the last is the causal-mechanical account by philosophers such as Nathan Salmon (1984), which focusses on the mechanisms that led to the output (note the strong disconnect with Durán's claim that such explanations only answer *how* questions)

- The idea is that what explains are (mechanistic) models of the system that leads to the phenomena we want to understand

- "Constitutive explanations, in contrast, reveal the organized activities of and interactions among parts that underlie, or constitute, the explanandum phenomenon. More specifically, they describe features of the mechanism for a phenomenon, where the mechanism includes the set of all and only the entities, activities and organizational features relevant to that phenomenon." (Craver & Kaplan, 2018)

Mechanism

- “A constitutive mechanistic model has explanatory force for phenomenon P versus P' if and only if (a) at least some of its variables refer to internal13 details relevant to P versus P', and (b) the dependencies posited among the variables refer causal dependencies among those variables (and between them and the inputs and outputs definitive of the phenomenon) relevant to P versus P”

- One worry: doesn’t this mean you need to give every detail about the process/system?

- Salmon-Completeness (SC): The Salmon-complete constitutive mechanism for P versus P' is the set of all and only the factors constitutively relevant to P versus P’.
  Where: Causally relevant factors are those that we could intervene upon (ideally) to change the explanandum phenomenon

TU Delft

Mechanism

- If model M contains more explanatorily relevant details than M* about the SC mechanism for P versus P', then M has more explanatory force than M* for P versus P', all things equal

- So not all details matter, only those that concern the causally relevant factors

- So there's a difference here with interventionists: the internal mechanism is what the explanation is concerned with, whereas interventionists don't necessarily care about the mechanism
- Similarly, it's different from unificationism in that there's still a focus on causation, plus theory unification is only relevant in so far as it helps to include more causally relevant factors

TU Delft

Objectivist explanation

- Note that all three prominent philosophical accounts don't really talk about what the person receiving the explanation is like

- Rather, their idea is that all explanations have the same structure, regardless of the recipient and specific why-question

- Pragmatism doesn't much feature here, though it is a position we find advocated for in the XAI context, e.g. Páez (2019), "The Pragmatic Turn in Explainable Artifcial Intelligence"

Pragmatism about explanation

- One general worry that one might have with the philosophical accounts is the following: “The success of an explanation therefore depends on several critical audience factors – assumptions, knowledge, and interests that an audience has when decoding the explanation” (De Graaf & Malle, 2017)

- Likewise, Páez (2019) considers a shift from explanation to understanding to highlight “the importance of taking into account the specifc context, background knowledge, and interests of end-users and stakeholders of opaque models”

- Not only that, according to him the connection between XAI and trust also essentially depends on pragmatic factors: “predictive reliability and a post hoc explanation are not sufficient to generate trust. Trust does not depend exclusively on epistemic factors; it also depends on the interests, goals, resources, and degree of risk aversion of the stakeholders.”

Pragmatism about explanation

- Páez does still advocate for some objectivity in explanation, but his paper nicely highlights the question: how much does the succes of an explanation depend on the person receiving it?

- Durán has a nice distinction between two types of pragmatism:
  - *Pragmatism*$_1$ takes considerations about the psychology of those involved in providing and receiving explanations as irreducible facts to be referenced in the explanation. *Pragmatism*$_1$ also takes into consideration the local context as irreducible for an explanation, such as the availability of resources, the specialization of the personnel, etc.
  - *Pragmatism*$_2$, on the other hand, takes pragmatic considerations such as the utility and usefulness of an explanation to be at the service of some goal connected to human interests. Under the umbrella of *pragmatism*$_2$, the structure of explanations does not depend on irreducible psychological and contextual facts, but rather on objective relations of dependence that are embedded with pragmatic, non-epistemic information

Pragmatism about explanation

- The question is: do different people really need different *types* of explanations? Or do they need explanations that have the same structure, but answer different why-questions and provide different levels of detail? (but all in the same format)?

- If you ask most philosophers, they will tend to the latter. So that there is a common thread among all succesful explanations, though details of the provided information may vary depending on background knowledge etc.

- That doesn't mean those pragmatic factors are unimportant. And we have empirical studies on the succes of XAI tools to highlight exactly that

Philosophy and XAI

- The philosophical discussions on the role of explainability can help challenge us to formulate why we are developing these tools. That can focus the kinds of explanations we aim to offer (are they for programmers, for decision makers, for those affected by decisions)?

- At the same time, philosophical reflections on explanations can provide us with blueprints for new explainability tools

- For example, to focus on contrastiveness, but also on:
  - Causal inference
  - Unifying patterns
  - Underlying mechanisms

TU Delft