The Mythos of Model Interpretability

Zachary C. Lipton

1 Introduction

Curious to hear from Zack about:

1. How the paper came to be?
2. Why was his PhD advisor, or any other co-author for that matter, not involved?
3. Thoughts on the paper's reception and personal perspectives looking back at it after 7 years.

2 Desiderata of Interpretability Research

2.1 Trust

What is trust? Possible answers:

- confidence that model will perform well wrt real objectives and scenarios
- subjective comfort in relinquishing control to the model

2.2 Causality

Hope to generate causal hypotheses by interpreting supervised learning models optimized to learn associations.

2.3 Transferability

ML objectives often diverge from true desired objectives. Examples:

- non-stationary environments
- models's predictions alter the environment
- adversarial manipulations; trying to game the system

2.4 Informativeness

Want additional information about the model's predictions, to convey to the human decision-maker.

2.5 Fair and Ethical Decision-Making

EU regulations:

1. *Right to explanation.* "Precisely what form such an explanation might take or how such an explanation could be proven correct and not merely appeasing remain open questions."
2. *Contestable decisions.* "So in order for such explanations to be useful it seems they must (i) present clear reasoning based on falsifiable propositions and (ii) offer some natural way of contesting these propositions and modifying the decisions appropriately if they are falsified."

3 Properties of Interpretable Models

3.1 Transparency

How does the model work? Humans lack this form of interpretability.

3.1.1 Simulability

Can a human contemplate the entire model at once?

True for linear models and decision trees but only to the extent that they are not too large. Might also be true for sufficiently compact deep nets.

3.1.2 Decomposability

Can each component of the model be intuitively understood?

This includes inputs. Linear models are not interpretable in this sense, if the inputs are heavily engineered or anonymized.

3.1.3 Algorithmic Transparency

E.g., convergence guarantees for linear models, but not for deep nets.

3.2 Post-hoc Interpretability

What else can the model tell me? Humans exhibit this form of interpretability.

Risk. There is no guarantee that the provided explanation accurately reflects the true mechanism behind the decision.

3.2.1 Text Explanations

Train a model to verbally explain the predictions of another.

3.2.2 Visualization

t-SNE, saliency maps.

3.3 Local Explanations

Instead of describing the entire model, describe it in the local region of a given input.

3.4 Explanation by Example

Find similar training examples. Similar to how humans justify actions by analogy.

4 Discussion

4.1 Linear models are not strictly more interpretable than deep neural networks

- Clear for algorithmic transparency.
- Simulability is lost when the model is too large.
- Decomposability is lost when the inputs are heavily engineered, anonymized, etc.
- Trade-off between algorithmic transparency and decomposability for identical predictive performance.
- Deep nets have an advantage for some kinds of post-hoc interpretations.

4.2 Claims about interpretability must be qualified

Not a monolithic concept. Also the different forms are often discordant.

Transparency can be claimed directly.

For post-hoc interpretability, need to fix a clear objective and show that it is achieved.

4.3 In some cases, transparency may be at odds with the broader objectives of AI

Transparency may come at the cost of predictive power.

“We should be careful when giving up predictive power, that the desire for transparency is justified and isn’t simply a concession to institutional biases against new methods.”

4.4 Post-hoc interpretations can potentially mislead

Model might present misleading but plausible explanations.

Already seen in humans: “Several journalists and social scientists have demonstrated that acceptance decisions attributed to virtues like leadership or originality often disguise racial or gender discrimination.”

4.5 Future Work

1. Mitigate discrepancy between real-life and ML objectives with richer loss functions and performance metrics, e.g., *sparsity-inducing regularizers* and *cost-sensitive learning.*
2. Extension to RL. Already addresses some forms, like causality. But things like fairness are still an issue.

5 Critiques

By ChatGPT. (I don’t necessarily agree.)

> While Zachary Lipton’s paper “The Mythos of Model Interpretability” has fostered important discussions, it faces some criticisms:
>
> 1. **Oversimplification of interpretability.** Critics argue the paper creates a false dichotomy between interpretable and non-interpretable models, rather than viewing interpretability as a spectrum.
> 2. **Downplaying interpretability’s importance.** The paper is seen as understating the significance of interpretability in high-stakes contexts such as healthcare or finance, where model transparency is crucial.
> 3. **Lack of actionable recommendations.** Critics note the paper primarily highlights interpretability issues without offering specific solutions for researchers and practitioners.
> 4. **Inadequate consideration of stakeholder perspectives.** The paper may not sufficiently address the diverse concerns and requirements of stakeholders such as end-users or regulators.
> 5. **Ignoring progress in interpretability research.** Some feel the paper doesn’t fully acknowledge advancements in interpretability techniques made post-2016.
>
> Despite these critiques, Lipton’s paper has played a vital role in raising awareness and fostering discussions about interpretability in the machine learning community.

My own. Were people more or less on the right track despite not carefully articulating the nuances like this paper? I was expecting a discussion of concrete works which fall into the traps outlined (like how he often throws shade on LIME, saliency maps, etc.)

Appendix

The ChatGPT output was generated with the following sequence of prompts:

> What can you say about the reception of Zachary Lipton's paper "The Mythos of Model Interpretability" in the Machine Learning community?
>
> Write some criticisms of the paper.
>
> Can you write the above in a slightly more concise manner?

On a separate session I used this prompt:

> Write a concise summary of the critiques for Zachary Lipton's paper "The Mythos of Model Interpretability."

And it did an appalling job:

> Zachary Lipton's paper "The Mythos of Model Interpretability" has drawn several critiques. The primary concerns are:
>
> 1. lack of clarity in the definition of interpretability, leading to potential confusion and misinterpretation;
> 2. questioning the necessity and usefulness of interpretability, arguing that it may not always improve trustworthiness or performance;
> 3. overemphasis on the human simulatability aspect of interpretability, potentially disregarding other important facets; and
> 4. the possible negative consequences of prioritizing interpretability over other factors, such as model performance and generalization.

Making Things Happen: A Theory of Causal Explanation

Chapter 1: Introduction and Preview

There are many senses of ‘explain’ – explain a procedure, describe a situation, justify a course of action. This account deals with *causal* explanations, which

- explain something by showing dependence on or relation to another, distinct factor
- the relationship is a matter of empirical fact (as opposed to being due to logical/conceptual reasons)

This approach distinguishes between explanation and ‘mere’ description.

Some causal and non-causal explanations:

- Salmon (1984): a causal explanation involves tracing spatiotemporally continuous causal processes
- Nerlich (1979): a geometrical explanation in terms of structural constraints (such as that of space-time) that is non-causal
- Sober (1983): Equilibrium explanations that specify that a number of initial conditions will lead to the state we seek to explain, without tracing the causal processes

Woodward seeks to provide a normative account of causal explanations – answer the question of ‘what should a causal explanation do?’ rather than just analyze what causation *is* (a descriptive account) or analyze what people mean when they talk of causes and explanations (a conceptual account).

Woodward offers a manipulationist account of causal explanations: causal explanations ‘tell us how, if we were able to change the value of one or more variables, we could change the value of other variables’. An explanation can show how to manipulate something even in the absence of information that allows for prediction or tracing spatiotemporally continuous processes, and having these other kinds of information may not provide knowledge of how to manipulate the situation. A successful explanation should show how manipulating factors in the explanation (*explanans*) would manipulate the thing that is explained (*explanandum*).

Information relevant to manipulation should be understood in counterfactual terms: *if* it were *possible* to manipulate the situation, how would the manipulation work? This allows the manipulationlist account to extend to cases where actual manipulation is not possible, such as large-scale cosmological phenomena.

Woodward suggests that our notions of causality and explanation are a product of our ability to manipulate – ”if we had been [...] intelligent trees capable only of passive observation – then it is a reasonable conjecture that we would never have developed the notions of causation and explanation and the practices associated with them that we presently possess”. This breaks with the philosophical norm of distinguishing science and technology, with science aiming to explain and technology aiming to manipulate. Science is deeply intertwined with our desire to manipulate and control nature.

Manipulation differs from simple counterfactual dependence. Consider the example of

If the barometer reading were to fall, a storm would occur

It's clear that making the barometer reading fall (creating the counterfactual scenario) will not cause the storm. The solution to this is by distinguishing intervention and invariance. An intervention is an ideal experimental manipulation made to determine whether X causes Y – to identify that a change in Y occurs only as a result of the change in X and not some other factor. "[T]he sorts of counterfactuals that matter for purposes of causation and explanation are just such counterfactuals that describe how the value of one variable would change under interventions that change the value of another."

An generalization is invariant if it holds under some intervention. That is, if the intervention changes the value of X, then the generalization still correctly describes the relationship between X and Y. "A necessary and sufficient condition for a generalization to describe a causal relationship is that it be invariant under some appropriate set of interventions."

An invariance is similar to a law of nature, but it does not need to satisfy all the conditions of a law (exceptionlessness, breadth of scope, degree of theoretical integration). Laws are one kind of invariant, but not all invariants are laws.

The manipulationlist account also allows for distinguishing levels of explanatory depth "by tracing it to differences in the degree of invariance of the generalizations to which the explanation appeals, and differences in the range of what-if-things-had-been-different questions that the generalization answers."

Woodward also argues that causal explanation is a practical activity that people have engaged in across cultures, and that accounts that rely on culture-specific notions of deductive inference are not a satisfactory account of causal explanation. Woodward also argues that accounts of causal explanation cannot completely break from everyday practices of casual explanation we engage in, and must offer some practical value. "Explanation in science does not give us something that is fundamentally different in kind from explanation in more ordinary contexts, but rather, as it were, a better version of the latter."

The account here is not reductive (does not reduce to something "outside of" causality). It relies on the notion of an intervention which is itself a causal notion. Woodward argues that even non-reductive accounts are important, even if they may be "circular." There are non-trivial choices about how various notions in this circle should be connected that are not all equally defensible.

Causal explanatory accounts should also allow for explanatory information to be epistemically accessible – it should contribute to understanding. There must be means of ascertaining whether the causal account is true.

The Fate of Explanatory Reasoning in the Age of Big Data
Frank Cabrera

Tl;dr
- Main question: Is data sufficient for inference?
- Position: No
- Argument: “even if R&S are right that sometimes explanatoriness is irrelevant toward theory-confirmation in the Bayesian sense, nevertheless, it remains true that it is rational to rely upon correlations in any predictive inference only if one possesses at least some explanatory knowledge” (Cabrera, 2021, p. 658)

I. Big Data Claims
   A. Data driven approach will dominate the scientific field and hypothesis driven methods will become obsolete
      1. “A data-driven approach rather than a hypothesis-driven approach has the potential to **lessen the impact of cognitive biases**, such as confirmation bias, on scientific research and to reveal patterns in the data, such as hidden interesting correlations, that might not have been noticed by the human researcher.” (Cabrera, 2021, p. 647)
      2. “In this way, proponents of the new data intensive science allege that theory (and, eventually, the human scientist) will gradually recede from the scene, at last allowing the data to “speak for themselves.”2” (Cabrera, 2021, p. 647)
   B. “Screening-off Thesis (SOT)” (Roche and Sober 2017 - R&S)

$$\text{(SOT)} \quad \Pr(H|O\&E) = \Pr(H|O)$$

      1. H - Hypothesis, O - Observation, E - Explanatoriness
      2. “explanatoriness is evidentially irrelevant”

II. Against Big Data Claims
   A. Explanatoriness is Evidently Relevant (EER)

(E) If H and O were true, H would explain O

$$\text{(EER)} \quad \Pr(H|O\&E) > \Pr(H|O)$$

   B. Inference to the Best Explanation (IBE)

(i) F is some fact or collection of facts.
(ii) Hypothesis H_1, if true, would explain F.
(iii) H_1 is a better explanation of F than its competitors H_2, H_3,...H_n.
(iv) Therefore, probably, H_1 is true.

III. Why Big Data Claims are Not Sufficient

A. New Argument for Incompatibilism

Premise 1 (P1): If SOT is true in a wide range of realistic cases, then IBE and Bayesianism are incompatible.
Premise 2 (P2): SOT is true in a wide range of realistic cases.
∴ So, IBE and Bayesianism are incompatible.[7]

B. Climenhaga

1. C_1 be "sometimes lung cancer after the age of 50 is caused by heavy smoking before the age of 50."
2. Sm be "S was a heavy smoker before the age of 50"
3. Ca be "S gets lung cancer after the age of 50."
4. E_{Sm} be "S's being a heavy smoker before the age of 50 would explain S's getting lung cancer after the age of 50, if S's being a heavy smoker before the age of 50 and S's getting lung cancer after the age of 50 were true."

$$Pr(Sm|Ca\&E_{Sm}) = Pr(C_1|Ca\&E_{Sm})Pr(Sm|Ca\&E_{Sm}\&C_1) + Pr(\sim C_1|Ca\&E_{Sm})Pr(Sm|Ca\&E_{Sm}\&\sim C_1) \quad (4)$$

$$Pr(Sm|Ca) = Pr(C_1|Ca)Pr(Sm|Ca\&C_1) + Pr(\sim C_1|Ca)Pr(Sm|Ca\&\sim C_1) \quad (5)$$

$$Pr(C_1|Ca\&E_{Sm}) > Pr(C_1|Ca) \quad (6)$$

$$Pr(Sm|Ca\&E_{Sm}\&C_1) \geq Pr(Sm|Ca\&C_1) \quad (7)$$

$$Pr(Sm|Ca\&E_{Sm}\&\sim C_1) = 0 \quad (8)$$

$$Pr(Sm|Ca\&E_{Sm}) = Pr(C_1|Ca\&E_{Sm})Pr(Sm|Ca\&E_{Sm}\&C_1) \quad (9)$$

$$Pr(C_1|Ca\&E_{Sm})Pr(Sm|Ca\&E_{Sm}\&C_1) = Pr(C_1|Ca)Pr(Sm|Ca\&C_1) + Pr(\sim C_1|Ca)Pr(Sm|Ca\&\sim C_1) \quad (10)$$

C. McCain and Poston

1. "certain cases of explanatory reasoning—e.g., Newton's gravitational theory explaining the orbits of the planets and Einstein's theory of general relativity explaining the precession of Mercury's

perihelion—are different in kind from the smoking-and-cancer case exploited by R&S in their defense of SOT.” (Cabrera, 2021, p. 657)

D. SOT
    1. “the thesis that R&S endorse—SOT—is that H’s being a potential explanation of O in itself is not evidentially relevant, once the confirmatory import of O is taken into account.” (Cabrera, 2021, p. 654) – H isn’t necessarily H1
    2. This makes P1 of the New Argument false

“Therefore, I do not dispute SOT, or P2 of the New Argument. Rather, I dispute the scope and significance of SOT, or P1 of the New Argument.” (Cabrera, 2021, p. 654)

E. Four versions of explanationism
    1. Weak
    2. Sturdy
    3. Ferocious
    4. Holocaust
F. “even if R&S are right that sometimes explanatoriness is irrelevant toward theory-confirmation in the Bayesian sense, nevertheless, it remains true that it is rational to rely upon correlations in any predictive inference only if one possesses at least some explanatory knowledge” (Cabrera, 2021, p. 658)
G. Ciceronian Causal-Nomological Requirement
    1. “(CCR) For any two logically distinct event-types A and B,it is rational predict some token-event b of type B, on the basis of the presence of some token-event a of type A, only if, given one’s background knowledge, there is a plausible causal nomological connection between A and B.16” (Cabrera, 2021, p. 659)