Position: The Causal Revolution Needs Scientific Pragmatism

Joshua R. Loftus [1]

Abstract

Causal models and methods have great promise, but their progress has been stalled. Proposals using causality get squeezed between two opposing worldviews. Scientific perfectionism–an insistence on only using "correct" models–slows the adoption of causal methods in knowledge generating applications. Pushing in the opposite direction, the academic discipline of computer science prefers algorithms with no or few assumptions, and technologies based on automation and scalability are often selected for economic and business applications. We argue that these system-centric inductive biases should be replaced with a human-centric philosophy we refer to as scientific pragmatism. The machine learning community must strike the right balance to make space for the causal revolution to prosper.

1. Introduction

There is a causal revolution underway in empirical sciences. From its century-old roots in controlled experiments, in recent decades it has made inroads in observational health and social sciences. As a research topic among methodologists, causal models have gained some popularity in economics, computer science, and statistics, and may be spreading into applications in other research fields and industries. However, *there is a risk that the causal revolution will stall*, and that humanity will miss out on the potential benefits of its success. To realize these benefits, **we argue that a philosophy of scientific pragmatism is necessary for the causal revolution to flourish**. This philosophy views causal models pragmatically as tools for hypothetical reasoning.

For the sake of simplicity we discuss only two types of models: causal and predictive. Strictly speaking, causal models are a subset of predictive models, so here we will use "predictive" as shorthand for "predictive models which are not also causal." If we think of modeling as a process of making assumptions and drawing conclusions then *causal models make more assumptions than predictive models*. These additional assumptions afford causal models additional conclusions, but also open them up to more criticism.

In the scenario of Figure 1, if we only judge a model based on its predictions of some important variable Y then the same model may work equally well regardless of the direction of causality. If we expect a model to also tell us what happens when we intervene on variables, then these models with equal predictive accuracy are no longer equivalent. *We become responsible to choose one of them*.

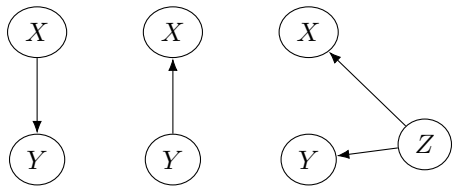

Figure 1. Structural causal models (SCMs) represented as directed acyclic graphs (DAGs). Variables are the nodes of the graph and causal effects are represented by arrows. An intervention is an operation that modifies the graph in some way. Intervening on a variable means we erase the arrows pointing into that variable, set the value of that variable arbitrarily, and then propagate the new value along directed pathways pointing out of that variable. On the left, X is a cause of Y, so intervening on X will result in a change in Y. In the other cases, intervening on X results in no change in Y. It is possible that the accuracy of some function $f(X)$ in predicting Y is equal in all cases.

We can use causal models to answer important questions that predictive models cannot illuminate. For example, if we want to know how we might change Y, then the model on the left of Figure 1 says we can intervene on X, the model on the right says we can intervene on Z (and that X would also change in that case), and the model in the middle says we would have to intervene on Y directly (which changes X as well).

The differences between these cases have important consequences when we apply models like these to reason about the real world. Suppose that Y is some variable we would like to change, but due to real world constraints we can only intervene on X. We have data and sophisticated predictive

[1]Department of Statistics, London School of Economics, London, UK. Correspondence to: Joshua Loftus <j.r.loftus@lse.ac.uk>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

models that predict Y from X with state of the art (SOTA) accuracy. Does this mean we should then spend resources on an intervention that changes X? The predictive model alone cannot answer this, we have to choose a causal model. If we choose the model on the left of Figure 1, but the real world is better described by the one on the right, then we may end up wasting resources, harming people, or pursuing a research program that will end in some type of failure.

Causal models have more assumptions and more implications. They force us to make more choices. As a result, there are more ways our choices can go wrong, but also more potential applications. To access the benefits of a causal revolution we have to be willing to pay the costs. We must learn to strike the right balance when judging proposals that involve causal modeling.

1.1. Published Context

Before elaborating our positions, we provide some context about the status of the causal revolution in empirical sciences. We would like to know if scientists care about causality and whether they are modeling and writing about it in a justified way. Fortunately for our purposes, meta-scientists have been hard at work studying these questions. Some indicative findings:

- In observational studies from 4 leading journals of obesity and nutrition (in 2006), 31% of papers had inappropriate language implying causality in the titles or abstracts (Cofield et al., 2010).
- Among all observational studies in the top 4 medical journals in 2010, authors recommend a medical practice in 56% (and only 14% of these recommended an RCT should be done to confirm their observational results) (Prasad et al., 2013).
- An NLP model applied to "about 38,000 observational studies in PubMed" showed about 32% used direct causal language, and most of these instances occurred only in the conclusion sections (i.e. not in abstracts or titles) (Yu et al., 2019).
- A study found "spin" in 84% of abstracts "of all non-randomized studies assessing an intervention published in the BioMed Central Medical Series journals between January 1, 2011 and December 31, 2013" and the most common kind of spin was causal language, occurring in 53% of abstracts (Lazarus et al., 2015).
- A study of non-RCTs in "1,170 articles from 18 high-profile medical/public health/epidemiology journals (65 per journal) published from 2010–2019" found "few studies explicitly declared an interest in estimating causal effects, the majority used language that moderately or strongly implied causality" and "action recommendations were identified in 60.3% [...] of discussion sections, about twice that in abstracts" (Haber et al., 2022).
- In a study of the 50 health studies that were most shared on social media in 2015, "34% percent of academic studies and 48% of media articles used language that [human expert] reviewers considered too strong for their strength of causal inference" (Haber et al., 2018).
- A systematic review of 199 systematic reviews of observational studies (published in 2019) found most (57%) were explicitly addressing causal issues, and among those which were not explicit about causal intent roughly half (51%) used causal language (Han et al., 2022).

The available research on "spin" and unjustified causal language is concentrated in health/medical sciences, but we believe the problem is much more widespread than directly indicated by the literature above. The health sciences have established traditions of systematic reviews, meta-analysis, and (pre-)registered trials. There are reasons to think the problem would only be worse in other fields where these accountability mechanisms have yet to catch on. For example, other fields have been shown to suffer from irreproducibility in their published literature (Open Science Collaboration, 2015), and failure to properly model causality may be one of the reasons that replication studies fail.

2. Positions

In this section we sometimes refer to the concept of an equivalent predictive model. Given any causal model, we can simply "forget" the additional causal assumptions regarding interventions and directionality and obtain an equivalent predictive model.

Positions 1-4 should prove relatively uncontroversial since they have stronger logical and/or evidential support. We still find it valuable to state and argue for them because, even though they may be uncontroversial, they may still be under-appreciated. Positions 5-6 are more speculative. We do not have evidence to cite for them. The reader will have to judge them based on their own experiences.

Position 1 (Proving too much)**.** *For any argument against using a causality in general, or a specific causal method in particular, we should check that this argument does not also apply to an equivalent predictive model.*

This is a simple logical point but may too often be forgotten.

There are many ways that probabilistic or mathematical models in general can be criticized. Sometimes people rediscover an old, general criticism and apply it to the newest wave of research. This can be a good thing. It is worth noticing when a new modeling approach is not immune to a long-standing general problem. But we should also take care to not penalize any specific class of models based on a general critique.

As an example, consider using a causal model to analyze fairness and discrimination, and suppose we have a variable in this model that represents a social category like race or gender. It is reasonable to use the concept of construct validity (Jacobs & Wallach, 2021) to criticize any model including such a variable. So it would be selective if that critique was only applied to causal models. On the other hand, it *is* reasonable to criticize causal models specifically with an argument that focuses on the experimental non-manipulability of variables like race and gender (Sen & Wasow, 2016; Kohler-Hausmann, 2018; Hu, 2023; Hu & Kohler-Hausmann, 2020). And in response to good criticisms, it may be possible to adapt and improve causal models for those purposes (Bynum et al., 2021; 2023; 2024).

Position 2 (People want causality)**.** *People usually want to reach causal conclusions. Even when they use predictive models, they often take actions or discuss the results as if the model were causal.*

We saw evidence for this position already in the literature cited in Section 1.1. Pushing beyond the evidence, we believe this problem is more prevalent and impactful than indicated by peer reviewed scientific work. This is because many scientists are aware that "association (e.g. correlation) is not causation," and hence may avoid using causal language when describing their predictive models. However, they may still be thinking about their work causally and taking actions as if the conclusions are causal. For example, they may choose to study a question that is only interesting or useful if the conclusions are causal, but they may do so using only predictive models.

Position 3 (Balanced standards)**.** *Predictive models are often used in place of causal models because they are held to too low a standard. Conversely, causal models–and methods or papers that use them–are often held to a higher standard. This is usually wrong.*

An author choosing to use a causal model makes additional explicit assumptions relative to an equivalent predictive model. At the stage of peer review, a referee may criticize any of these additional assumptions. And since causal models entail additional conclusions or implications, if the paper is published it is possible some subsequent research will falsify one of those additional conclusions, thereby calling the paper into question. Not wanting to risk it, scientists may avoid these costs of causality and opt for a model with fewer assumptions.

The principle of parsimony (or simplicity) states "*other things being equal*, simpler theories are better [emphasis added]" (Baker, 2022). Now, it is true that a causal model has more assumptions than an equivalent predictive model, so in a sense the predictive model is simpler. But it is *not* true that these two models are otherwise equal. The causal version implies more.

There are some good reasons to apply different standards to causal and predictive models. If an assumption of a causal model is difficult or impossible to test via any experiment, we should treat that assumption with caution and transparently acknowledge the limitation of any conclusions based on that model. All further conclusions or models built on that assumption must never forget the untested part of their foundation. Similarly, when reporting results of a predictive model we must acknowledge they do not necessarily imply any causal relationships.

Achieving a level playing field between causal and predictive models requires applying balanced standards. If a research project proposes using predictive models it should be questioned whether the desired goals of the inquiry only include predictive conclusions. If the work does not pay the price of making explicit assumptions required for causal conclusions it should not be allowed to suggest those conclusions for free. On the other hand, if causal conclusions are desired, then it is wrong to penalize the additional assumptions necessary for the model to able to answer our actual questions.

Position 4 (Falsifiability and the scientific method)**.** *Causal models can be falsified in more ways than predictive models. This is usually good.*

Falsificationism, one of the most well-known philosophies about the scientific method, states that science proceeds not by proving theories but rather by falsifying them (Thornton, 2023). Suppose someone does an experiment in a scenario described by Figure 1 that manipulates X and finds no significant change in Y. This falsifies the causal model $X \to Y$, but does not falsify the predictive model. If the status quo in this field is to use predictive models, then this experiment may not result in any scientific progress. Other studies will continue to support the predictive model's accuracy, and the field can remain conflicted.

We have already discussed how causality involves more assumptions and conclusions. Any of these that can be tested provide another target for potential falsification. As a result, the scientist in our publishing story above decided not to use a causal model. Irrespective of their individual

success in publishing, is this a good thing for science in general?

It is a misconception about science that the goal of designing a model or theory is to have the closest possible fit to reality in general, or to be immune to any criticism or disproof. A scientific theory can have great explanatory power and application potential even if it is known to be incorrect. For example, suppose someone builds an algorithm for autonomous vehicle control that uses classical mechanics. It would not be interesting or useful to point out that classical mechanics is "wrong" *unless*, say, that algorithm will be used for rockets traveling at relativistic speeds.

Falsification does not need to be a blindly adversarial process reaching for any conclusion that can be disproved. It can be done selectively, with attention given to those parts of a model that are consequential in the intended application. As an example consider causal analysis of discrimination. The interested parties may only be disputing the strength of causal determination along one pathway in a causal graph, and are not interested in testing every other falsifiable prediction of the model.

Assumptions can enable or block scientific progress. They can block progress if they are never questioned, which seems unlikely given the incentives to publish novel results. They cannot enable progress if they are never considered. And that may be a real risk if the scientific and machine learning communities do not learn to tolerate more assumptions in some applications.

Position 5 (Outsourcing responsibility)**.** *A preference for predictive models is often a sign of avoiding responsibility.*

If a work's value derives mainly from possible causal conclusions, but the author does not make transparent causal assumptions, they are leaving it to readers to reach the intended conclusion.

Researchers sometimes avoid making transparent causal assumptions out of fear. They do not want to be responsible for more modeling choices because each choice opens up another potential criticism. They prefer whichever methods seem to work with the least number of choices, automatically, on the default settings. And this may not be all wrong, in fact, except that the same researchers often want to interpret the results causally. If they have taken a statistics course they may know to avoid using the word "cause" when talking about the results, but they might get away with suggesting or hinting at causality with weasel words like X "drives," "predicts," or "is linked to" Y. The responsibility for making causal conclusions falls to their audience, who, due to their distance from the work, may be less familiar with the limitations.

Using causal models can be a more honest and transparent way to answer the questions we truly want to ask. Instead of pretending that we are only interested in prediction, we can admit that we want to change the world. We are almost always going to use predictions for some purpose that involves taking some actions to try to change some outcome. So we can and should use a causal model to elaborate how we think this is going to work.

Position 6 (Relevance to machine learning)**.** *The machine learning community is uniquely well-positioned to integrate causality, but also has a disciplinary bias against it.*

The overlap between machine learning and causality is growing in absolute terms (see Makhlouf et al. (2020); Yao et al. (2021); Schölkopf (2022); Kaddour et al. (2022) for some recent surveys). This is natural since applications with causal models usually involve learning functions from data. Machine learning has developed many powerful algorithms for estimating functions and learning about structure, and it is often only some additional assumptions required to transform any of these into explicitly causal methods.

However, there are reasons to be concerned about the prospects of causality in machine learning.

First, computer science as a discipline prefers general purpose algorithms, i.e. ones which are "assumption free" or "model free." This makes sense for building robust systems, since an algorithm that relies on specific assumptions can break down if conditions change so that the assumptions no longer hold. There is also an economic logic: a technological service that is more automatic and scalable has more potential customers. Algorithms using causal models are less automatic and scalable because they require more context or task-specific input. Decisions about causal directions might require domain knowledge, for example.

Second, for many years, the "secret sauce" of machine learning's success (Donoho, 2017) has been the Common Task Framework: a competition between algorithms on a fixed, well-defined task (Liberman, 2010). These tasks are usually about predictive accuracy or a set of predictive benchmarks. And, usually, the SOTA algorithms that gain the most success and attention in the field are highly complex predictive models where "the goal is not interpretability" (Breiman, 2001). It makes sense from the viewpoint of constrained optimization that if we evaluate models only on predictive accuracy then any constraints–for example, belonging to a class of certain interpretable causal models–will likely result in worse evaluations.

On the plus side, pragmatism in computer science does sometimes make room for works that pose or reformulate problems differently. This may allow some exceptional works to use causal modeling without assuming the goal of learning a "correct" model (Janzing et al., 2023).

Speculation. Allow us this brief departure from more evidenced argument. We speculate that the high degree of competition in machine learning conferences–with their low acceptance rates and focus on SOTA performance in predictive tasks–exacerbates some of the challenges for causal methods we have outlined. In particular, when referees enter a review process that they know has a low acceptance rate, they may be looking for anything they can criticize, and the additional assumptions of a causal model are a prime target.

3. Principles

3.1. Scientific Pragmatism

Pragmatism is a philosophical tradition that emphasizes usefulness (Legg & Hookway, 2021). Continuing with this tradition, we advocate thinking of scientific theories–and models in general–as tools. Instead of asking if a model is right, correct, or accurate, we should ask if it is *useful for a certain purpose*. Thus, the fitness of a theory or model depends on the goals of the people using it.

Recall our previous example of the autonomous vehicle control algorithm. If this system will only be used for automobiles on the surface of our planet, then classical mechanics is a good fit. Quantum mechanics has not been falsified under any conditions, but it would be a bad fit for the design of such an algorithm.

Causal models can be useful as tools for a variety of purposes. They can serve as a notation and language for reasoning and conveniently communicating about relationships between sets of variables (Imbens, 2020) and, hence, for interpretability and explainability (Blöbaum & Shimizu, 2017; Sani et al., 2020; Frye et al., 2020; Heskes et al., 2020; Zhao & Hastie, 2021; Loftus et al., 2023). Causality can be used to analyze algorithms for discrimination or fairness (Kilbertus et al., 2017; Kusner et al., 2017; Nabi & Shpitser, 2018; Zhang & Bareinboim, 2018; Chiappa, 2019; Yang et al., 2021).

Machine learning researchers are likely already comfortable with the idea of selecting the right tool for a task. It is important that they remember there are tasks other than prediction with SOTA accuracy. And in some application domains, the principle of scientific pragmatism supports our Positions 3 and 4 as an important counterweight to scientific perfectionism.

The greatest challenge for a scientific pragmatist is the danger of "cargo cult science" (Feynman, 1998) or its relative "cargo cult programming" (Raymond & Steele, 1997). These terms describe failures that occur when people select tools or attempt to build them without understanding how they work, based only on appearances of success. In the classic example of a cargo cult as described by Feynman, people attempt to build an aircraft, for example, by fashioning objects with similar appearances. The problem is that such an aircraft, built without the knowledge and tools of engineering, cannot function. While the problem is obvious in this example, it can be more subtle in cargo cult science where a model or method may appear to function. The problem of cargo cult machine learning may be worse still because machine learning tasks are highly standardized and its tools automated. Students of machine learning can rapidly achieve the appearance of success without understanding the tools they are using, knowing about their limitations, or developing skills to judge or build new tools.

Pragmatism can serve as a descriptive theory about practices that are already standard. We simply advocate that this is done consciously and honestly, acknowledging our purposes. When we select tools for predictive accuracy we should not expect this to necessarily increase scientific knowledge if, for example, causal interpretability is sacrificed in the process.

3.2. Value Pluralism

When selecting tools and judging their usefulness, we may not be able to reduce this judgment to the application of a single value (Mason, 2023). We have a plurality of values that are important to us, and the relevance of these values may change depending on our purposes. The challenge increases when our values are in conflict. Readers may be thinking that multi-objective optimization or multi-task learning can simply "solve" this problem for us, but different values may be incommensurable or not quantifiable (Schroeder, 2021). For example, we may desire both predictive accuracy and interpretability, and we may not have any quantified measures of interpretability.

Value pluralism is in strong conflict with machine learning's "secret sauce" as described in Position 5. The field has become more pluralistic by widening the set of standard tasks, and this is good work that should continue. However, there may still be limits on which values can be included for tasks that are standardized, or in systems built using automation and scaling (Nguyen, forthcoming; Finnveden et al., 2022). Some values may require something closer to case-by-case human judgments. For example, the usefulness of a particular causal model may not be fully determined by the available data and algorithms, but also depend on a subjective judgment involving its interpretation.

3.3. Humanism or Anthropocentrism

When we build infrastructure and systems using automation, selecting tools for their scalability and plug-and-play fitness within the existing system, it is not just causal models that are at risk of being pushed out of the picture. Another risk is evident in the terms "human-in-the-loop" or "human feed-

back," the very existence of these terms implying that there are alternatives which somehow do not involve humans.

We assert that humans are the ultimate source and arbiters of values.[1] By bringing the focus to humans using tools, pragmatism could shed light on various perennially confusing issues in machine learning. It can help us remember that the "agents" are humans, and that software processes started by humans are tools being used for a purpose.

In this human-centric view, it does not matter as much that causal methods may be less automatic or scalable. It does not matter as much when a particular algorithm is the SOTA for a certain task. We always have other purposes and values which are not codified in that task, and may not even be readily quantifiable. For example, we may sometimes wish to use a causal model within the context of a broad scientific question or business plan that involves various subjective factors. We may be motivated by those factors to make assumptions in that causal model which are not decidable in an automated way based on the available data. We can remember that this introduces limitations, and we can decide that the choice still fits with our values.

4. Conclusion

4.1. Related Ideas

Inspired by some earlier debates in statistics, Tukey (1960) proposed we distinguish between two different goals or modes of statistical analysis: conclusions vs decisions. Conclusions (the Fisherian mode) should be final, authoritative, well-tested, and we should rarely need to revise them, if ever. As such, it makes sense to have high standards of evidence for conclusions. On the other hand, decisions (the Neyman-Pearson mode) are temporary, we often make them just so we can proceed to act, and we are not surprised if we need to try something else. We speculate that our Positions 2 and 3 are often related to a failure to make the distinction Tukey recommends here. In particular, we can make the decision to use a causal model without taking that model as conclusive.

Echoing through history, the pragmatic statistician George Box reminds us: "All models are wrong, but some are useful," and "Since all models are wrong the scientist must be alert to what is *importantly wrong* [emphasis added]. It is inappropriate to be concerned about safety from mice when there are tigers abroad" (Box, 1976).

Preceding the development of structural causal models with their formal definition of intervention, philosopher of science Hacking (1983) makes a case that the practice of science involves intervention rather than representation. This makes scientists pragmatic by necessity.

In the early days of machine learning, Breiman (2001) argued in favor of its single-minded focus on predictive accuracy, and, more recently, Rodu & Baiocchi (2023) described conditions when such "outcome reasoning" may succeed or may be inappropriate.

Much recent work has focused on the problem of distribution shift or the ability of predictive models to generalize out-of-distribution. Schölkopf et al. (2021) showed causal models can be a natural fit for this task. Interestingly, Richens & Everitt (2024) established something like a converse of this result: if a predictive model can succeed under sufficiently many distribution shifts then that model entails an approximate causal model of the data generating process. This implies that if someone cares about maintaining predictive accuracy across a variety of different settings then, in a sense, they must care about causality.

4.2. Significance

Should anyone who does not personally have any uses for causal models care about challenges holding back the causal revolution? Yes, if they want humanity to continue unlocking the benefits of scientific progress, or if they want human values that are not in harmony with automated systems to be able to flourish.

A pragmatic use of causal models could revolutionize science. It has the potential to become a new standard language for the communication of scientific results, replacing the less interesting significance tests for differences between groups or for regression coefficients. For the sake of simplicity let's take the view that a scientific discovery occurs when someone falsifies a hypothesis. If that hypothesis is uninteresting, then it is not much of a discovery. Causal models are an ideal framework for specifying and communicating *interesting* hypotheses.

It is an empirical fact that humans have many different values. Systems using machine learning apparently grow most rapidly when they are built based on automation, scalability, and a narrow focus on standardized tasks. Such systems risk crowding out values which are not included in the set of tasks, or which are at odds with automation and scale.

4.3. Summary

We have argued that methods based on causality face some challenges which can put them at a disadvantage. Within machine learning specifically, this occurs because of a narrow focus on SOTA performance on prediction tasks. Most of those tasks do not encode subjective human uses like causal interpretation, for example. We advocate a philosophy combining scientific pragmatism, value pluralism, and humanism, where models and tools can be selected based

[1] We are not opposed to a view that includes other forms of biological life, but we leave that for another discussion.

on a more inclusive set of values which may not be easily quantifiable or standardized. With these views, the machine learning community can make room for a greater variety of interesting applications and methods, including more that focus on causality, and promote the flourishing of more values aside from predictive accuracy. The causal revolution needs scientific pragmatism in order to proceed beyond its first steps, and human-centric value pluralism can help us apply scientific pragmatism wisely.

Acknowledgements

The author would like to thank Pooja Loftus, Ricardo Silva, and David Watson for helpful conversations and feedback on this work.

Impact Statement

This paper presents positions intended to advance the field of Machine Learning. There are many potential societal consequences, some of which were discussed throughout the text. We do not feel we should specifically highlight any other potential impacts here.

References

Baker, A. Simplicity. In Zalta, E. N. (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2022 edition, 2022.

Blöbaum, P. and Shimizu, S. Estimation of interventional effects of features on prediction. In *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6. IEEE, 2017.

Box, G. E. Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799, 1976.

Breiman, L. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.

Bynum, L., Loftus, J., and Stoyanovich, J. Disaggregated Interventions to Reduce Inequality. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 1–13. Association for Computing Machinery, New York, NY, USA, October 2021. ISBN 978-1-4503-8553-4. URL https://doi.org/10.1145/3465416.3483286.

Bynum, L., Loftus, J., and Stoyanovich, J. Counterfactuals for the future. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

Bynum, L. E., Loftus, J. R., and Stoyanovich, J. A new paradigm for counterfactual reasoning in fairness and recourse. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, 2024.

Chiappa, S. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 7801–7808, 2019.

Cofield, S. S., Corona, R. V., and Allison, D. B. Use of causal language in observational studies of obesity and nutrition. *Obesity facts*, 3(6):353–356, 2010.

Donoho, D. 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4):745–766, 2017.

Feynman, R. P. Cargo cult science. In *The art and science of analog circuit design*, pp. 55–61. Elsevier, 1998.

Finnveden, L., Riedel, J., and Shulman, C. AGI and Lock-In. https://forum.effectivealtruism.org/posts/KqCybin8rtfP3qztq/agi-and-lock-in, 2022.

Frye, C., Rowat, C., and Feige, I. Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. *Advances in Neural Information Processing Systems*, 33:1229–1239, 2020.

Haber, N., Smith, E. R., Moscoe, E., Andrews, K., Audy, R., Bell, W., Brennan, A. T., Breskin, A., Kane, J. C., Karra, M., et al. Causal language and strength of inference in academic and media articles shared in social media (claims): A systematic review. *PloS one*, 13(5):e0196346, 2018.

Haber, N. A., Wieten, S. E., Rohrer, J. M., Arah, O. A., Tennant, P. W., Stuart, E. A., Murray, E. J., Pilleron, S., Lam, S. T., Riederer, E., et al. Causal and associational language in observational health research: a systematic evaluation. *American journal of epidemiology*, 191(12): 2084–2097, 2022.

Hacking, I. *Representing and intervening: Introductory topics in the philosophy of natural science*. Cambridge university press, 1983.

Han, M. A., Leung, G., Storman, D., Xiao, Y., Srivastava, A., Talukdar, J. R., El Dib, R., Morassut, R. E., Zeraatkar, D., Johnston, B. C., et al. Causal language use in systematic reviews of observational studies is often inconsistent with intent: a systematic survey. *Journal of Clinical Epidemiology*, 148:65–73, 2022.

Heskes, T., Sijben, E., Bucur, I. G., and Claassen, T. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. *Advances in neural information processing systems*, 33:4778–4789, 2020.

Hu, L. What is "race" in algorithmic discrimination on the basis of race? *Journal of Moral Philosophy*, 1(aop):1–26, 2023.

Hu, L. and Kohler-Hausmann, I. What's sex got to do with machine learning? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 513–513, 2020.

Imbens, G. W. Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature*, 58(4): 1129–1179, 2020.

Jacobs, A. Z. and Wallach, H. Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 375–385, 2021.

Janzing, D., Faller, P. M., and Vankadara, L. C. Reinterpreting causal discovery as the task of predicting unobserved joint statistics. *arXiv preprint arXiv:2305.06894*, 2023.

Kaddour, J., Lynch, A., Liu, Q., Kusner, M. J., and Silva, R. Causal machine learning: A survey and open problems. *arXiv preprint arXiv:2206.15475*, 2022.

Kilbertus, N., Rojas Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. Avoiding discrimination through causal reasoning. *Advances in neural information processing systems*, 30, 2017.

Kohler-Hausmann, I. Eddie Murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. *Nw. UL Rev.*, 113:1163, 2018.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. Counterfactual Fairness. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Lazarus, C., Haneef, R., Ravaud, P., and Boutron, I. Classification and prevalence of spin in abstracts of non-randomized studies evaluating an intervention. *BMC medical research methodology*, 15:1–8, 2015.

Legg, C. and Hookway, C. Pragmatism. In Zalta, E. N. (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2021 edition, 2021.

Liberman, M. Obituary: Fred jelinek. *Computational Linguistics*, 36(4):595–599, 2010.

Loftus, J. R., Bynum, L. E., and Hansen, S. Causal dependence plots. *arXiv preprint arXiv:2303.04209*, 2023.

Makhlouf, K., Zhioua, S., and Palamidessi, C. Survey on causal-based machine learning fairness notions. *arXiv preprint arXiv:2010.09553*, 2020.

Mason, E. Value Pluralism. In Zalta, E. N. and Nodelman, U. (eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2023 edition, 2023.

Nabi, R. and Shpitser, I. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Nguyen, C. T. Value capture. *Journal of Ethics and Social Philosophy*, forthcoming.

Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015.

Prasad, V., Jorgenson, J., Ioannidis, J. P., and Cifu, A. Observational studies often make clinical practice recommendations: an empirical evaluation of authors' attitudes. *Journal of clinical epidemiology*, 66(4):361–366, 2013.

Raymond, E. S. and Steele, G. L. *The Jargon File, Version 4.0. 0, 24 Jul 1996*. Project Gutenberg, 1997.

Richens, J. and Everitt, T. Robust agents learn causal world models. *International Conference on Learning Representations*, 2024.

Rodu, J. and Baiocchi, M. When black box algorithms are (not) appropriate. *Observational Studies*, 9(2):79–101, 2023.

Sani, N., Malinsky, D., and Shpitser, I. Explaining the behavior of black-box prediction algorithms with causal learning. *arXiv preprint arXiv:2006.02482*, 2020.

Schölkopf, B. Causality for machine learning. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pp. 765–804. 2022.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. Toward causal representation learning. *Proceedings of the IEEE*, 109(5): 612–634, 2021.

Schroeder, M. Value Theory. In Zalta, E. N. (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2021 edition, 2021.

Sen, M. and Wasow, O. Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics. *Annual Review of Political Science*, 19:499–522, 2016.

Thornton, S. Karl Popper. In Zalta, E. N. and Nodelman, U. (eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2023 edition, 2023.

Tukey, J. W. Conclusions vs decisions. *Technometrics*, 2(4): 423–433, 1960.

Yang, K., Loftus, J. R., and Stoyanovich, J. Causal Intersectionality and Fair Ranking. In Ligett, K. and Gupta, S. (eds.), *2nd Symposium on Foundations of Responsible Computing (FORC 2021)*, volume 192 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 7:1–7:20, Dagstuhl, Germany, 2021. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-187-0. doi: 10.4230/LIPIcs.FORC.2021.7. URL https://drops.dagstuhl.de/opus/volltexte/2021/13875. ISSN: 1868-8969.

Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., and Zhang, A. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(5):1–46, 2021.

Yu, B., Li, Y., and Wang, J. Detecting causal language use in science findings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4664–4674, 2019.

Zhang, J. and Bareinboim, E. Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Zhao, Q. and Hastie, T. Causal interpretations of black-box models. *Journal of Business & Economic Statistics*, 39(1):272–281, 2021. doi: 10.1080/07350015.2019.1624293. URL https://doi.org/10.1080/07350015.2019.1624293.