Algorithms for reliable decision-making need causal reasoning

Christoph Kern*[1,2,3,†], Unai Fischer-Abaigar[1,2], Jonas Schweisthal[2,4], Dennis Frauen[2,4], Rayid Ghani[5], Stefan Feuerriegel[2,4], Mihaela van der Schaar[6], and Frauke Kreuter[1,2,3]

[1]Department of Statistics, LMU Munich, Munich, Germany
[2]Munich Center for Machine Learning (MCML), Munich, Germany
[3]Joint Program in Survey Methodology (JPSM), University of Maryland, College Park, Maryland, USA
[4]LMU Munich School of Management, LMU Munich, Munich, Germany
[5]Machine Learning Department, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
[6]Faculty of Mathematics, University of Cambridge, Cambridge, UK
[†]Corresponding author: Christoph Kern, christoph.kern@lmu.de

Decision-making inherently involves cause-effect relationships, which introduce causal challenges. We argue that reliable algorithms for decision-making need to build upon causal reasoning. Addressing these causal challenges requires explicit assumptions about the underlying causal structure to ensure identifiability and estimatability, which means that the computational methods must successfully align with decision-making objectives in real-world tasks.

1 Introduction

Algorithmic decision-making (ADM) has become common in a wide range of domains, including precision medicine, manufacturing, education, hiring, public sector, and smart cities. At the core of ADM systems are data-driven models that learn from data to recommend decisions, often with the goal of maximizing a defined utility function [1]. For example, in smart city contexts, ADM is frequently used to optimize traffic flow through predictive models that analyze

*This version of the article has been accepted for publication, after peer review (when applicable) but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: https://doi.org/10.1038/s43588-025-00814-9. Use of this Accepted Version is subject to the publisher's Accepted Manuscript terms of use https://www.springernature.com/gp/open-research/policies/accepted-manuscript-terms real-time data, thereby reducing congestion and improving urban mobility. Another prominent application area for ADM are normative decision support systems (often subsumed under "prescriptive analytics") or, more recently, artificial intelligence (AI) agents that either inform or automatically execute managerial and operational decisions in industry. Yet, the applications of ADM to high-stakes decisions face safety and reliability issues [1, 2, 3]. Often, the objectives of ADM systems fail to align with the nuanced goals of real-world decision-making, thus creating a tension between the potential of ADM and the risk of harm and failure. Especially when deployed in dynamic, real-world environments, ADM can amplify systemic disadvantages for vulnerable communities and lead to flawed decisions.

In this Comment, we argue that *reliable* algorithmic decision-making—systems that perform safely and robustly under deployment conditions—must be grounded in causal reasoning.

2 Why ADM is rooted in causal reasoning

2.1 Decisions involve cause-effect relationships

Making the right decision—whether selecting the best treatment for a patient, choosing the most effective marketing strategy, or optimizing traffic signals in a smart city—requires understanding how a decision will influence the outcome of interest. Unlike a passive observation, the outcome is often directly shaped by the decision itself. For example, choosing a treatment to reduce a patient's mortality risk or adjusting traffic signals to minimize congestion directly affects the outcome being optimized. This distinguishes decision problems from pure associations, since decisions present interventions that eventually produce effects.

A fundamental challenge in modeling decision-making is that we can only observe the outcome for the decision that was actually taken—not what would have happened under alternative interventions (the fundamental problem of causal inference [4]). The latter is unobservable and thus referred to as the counterfactual outcome. Estimating and comparing the effects of different decisions requires reasoning about these counterfactuals. For example, if a patient receives a treatment and recovers, we cannot observe what their outcome would have been without the treatment. As a remedy, causal assumptions are needed that connect the observed data and the cause-effect relationships to express and eventually estimate counterfactual outcomes [5].

2.2 Causal assumptions are needed for modeling decision-making

Starting with clear assumptions about the data-generating process and causal relationships is crucial for establishing valid causal estimates. Such assumptions allow one to link a *causal estimand* (the quantity we aim to estimate to guide decision-making) to a *statistical estimand* (a quantity that can, in principle, be estimated from observed data). This distinction is critical: the causal estimand reflects our real objective, while the statistical estimand serves as the counterpart that we aim to estimate with our model. Without clearly defining the causal estimand first, it becomes impossible to ensure that our ADM model will align with the objective in our real-world tasks. Moreover, if the statistical estimand does not match the causal estimand, we risk producing unreliable estimates, leading to suboptimal or even incorrect decisions.

The above distinction between causal and statistical estimands highlights two key challenges: *identifiability* and *estimatability*. Identifiability refers to whether the causal estimand—the true effect of a decision—can be expressed as a function of observable data under the given assumptions [6]. Without identifiability, the effects of decisions, and therefore the corresponding outcomes, remain ambiguous, no matter how much data is available. In public policy, for example, large-scale administrative data might be available for ADM model training but these data rarely allow to recover the nuanced decision-rules based on which caseworkers have administered interventions in the past. This ambiguity is particularly problematic since it undermines the reliability of ADM models in practice. Without identifiability, any ADM model aiming to learn the correct effect of the decision remains biased. As a result, ADM models may produce incorrect or suboptimal decisions. In medicine, for example, both health outcomes and decisions for (costly) treatments may depend on unobserved factors such as patients' socio-economic status, preventing the ability to learn the true effect of treatment from the observed relationships. Estimatability, on the other hand, deals with whether the statistical estimand—once identified—can be reliably computed from finite data. Even when a causal estimand is identifiable, estimating it is usually a statistically challenging task. Practical issues such as limited sample size, measurement noise, and model misspecification can hinder the ability to produce precise and unbiased estimates. Importantly, *identifiability is a prerequisite for estimatability*: without identifiability, the task of estimation becomes meaningless, as there is no assurance that the statistical estimand corresponds to the actual effect of a decision, meaning that an ADM model may not optimize against the utility of interest.

Generally, different frameworks are used to formalize causal assumptions [7], with two prominent approaches being structural causal models (SCMs) and the potential outcomes framework. While their methods and notation for specifying and identifying causal effects differ, they are often regarded as essentially equivalent in their core principles. SCMs use tools like causal graphs together with functional causal mechanisms to explicitly represent the cause-effect relationships, while, in contrast, the potential outcomes framework focuses on defining the outcomes that would occur under alternative decisions—capturing what would happen under each choice. To enable valid causal reasoning, several key assumptions about cause-effect relationships are often made [4]. For example, unconfoundedness assumes there are no hidden confounders that simultaneously influence both the decision and the outcome. Confounders in medical treatment decisions, for example, can include disease severity, patient characteristics, and access to healthcare. Positivity ensures that every decision has a nonzero probability of being observed for all individuals, providing sufficient data for comparison. Additionally, the stable unit treatment value assumption (SUTVA) essentially posits that an individual's outcome is influenced only by their own decision and not by the decisions of others.

3 Towards causal decision-making

3.1 Utilizing causal reasoning for reliable decision-making

To ensure reliable decisions with ADM, developers must first evaluate whether identifiability holds—that is, whether the causal estimand can be expressed using observable data and that the statistical estimand corresponds to the causal estimand. This requires clearly stating and justifying the assumptions about the data-generating process and causal relationships. By taking this step, developers can help align the objectives of ADM systems with the real-world goals they are meant to support. While this alone can not guarantee safe decision-making, it is a critical prerequisite for ensuring that decision-making is both effective and reliable (see Fig. 1).

An important step is to carefully evaluate whether the necessary causal assumptions hold in practice. For instance, when developers of ADM systems have access to randomized data from experiments, unconfoundedness is often ensured by design, making causal effects identifiable. For example, in a smart city context, randomized control trials might be performed to test the impact of adaptive traffic light systems on reducing congestion. In some cases, data from natural experiments may be available, for example, due to temporal malfunctions in operational processes or regional variation in the roll-out of policy programs. However, access to randomized data is rare, particularly in high-stakes settings where randomization is impractical or unethical. Instead, many ADM systems from practice rely on observational data generated by complex processes where interventions were assigned based on historical policies (for instance, expert judgment, organizational guidelines, and institutional policies). In these cases, guaranteeing causal assumptions can be challenging and often requires input from domain experts and additional contextual knowledge.

It is important to emphasize that ignoring causal assumptions simply because they are difficult to guarantee is not a solution. Failing to make causal assumptions explicit does not mean they do not exist—on the contrary, neglecting them can lead to flawed and unreliable decision-making. For example, naively training models on observational data generated by historical decision-making processes can result in systematically underestimating true risks [8]. Even ADM models that do not explicitly target counterfactual outcomes—or that do not state the causal assumptions—often rely on such causal assumptions implicitly. A prominent example is prediction-based decision rules, where a predictive model is trained and decisions are made based on a specified threshold [9]. Here, the causal assumptions are rarely stated but such ADM models typically assume independence between the outcome and decision or rely on a known model of treatment effects given the predicted baseline outcome. A common example is hospital discharge decisions, where the decision to discharge or retain a patient typically does not influence the underlying disease progression. In such contexts, predictive associations alone may suffice for making effective decisions. Nonetheless, explicitly clarifying these assumptions is crucial, as it ensures that the decision-makers are aware of potential limitations or scenarios where associations alone might fail due to unnoticed feedback loops.

We now offer a few illustrative examples. In prediction-based decision rules, the goal is often to intervene in cases with a high predicted risk of adverse outcomes if no help were provided. For instance, in healthcare, a risk prediction model may trigger an alarm for patients where their health status is likely to deteriorate without treatment, thus allowing for targeted interventions. Here, the prediction—and thus the decision rule—is learned from historical data under interventions, and, hence, identification typically relies on the assumption of unconfoundedness or that there is sufficient variation in the historical decision-making. In offline policy learning, the objective is to maximize expected outcomes under a given decision policy, which maps observed covariates to actions. For example, the task might involve determining optimal traffic signal timings at intersections to minimize average travel time across the city. Achieving this goal requires assuming, for instance, unconfoundedness across all decisions so that intervention effects can be reliably estimated and the policy can optimize the expected utility. In reinforcement learning, the focus extends to optimizing the expected return over a time horizon, usually in an online setting. For example, adaptive traffic control systems might iteratively adjust signal timings based on real-time data to minimize congestion throughout the day. Here, identification is often naturally satisfied because decisions (for instance, signal changes) are assigned sequentially by a known policy. However, challenges arise in partially observed settings, such as unpredictable events, which require further effort to come up with an identification strategy. In sum, in each of these examples, causal assumptions provide the foundation for linking causal estimands to statistical estimands and thus for building an ADM model that allows for optimal decisions.

3.2 Why a causal lens is beneficial for reliability

We believe that adopting a causal perspective provides a systematic way to connect decision-making objectives, computational methods, and real-world deployment environments. This approach places causal identification at the center of ADM development. Hence, causal identification requires developers to clearly differentiate between causal estimands and statistical

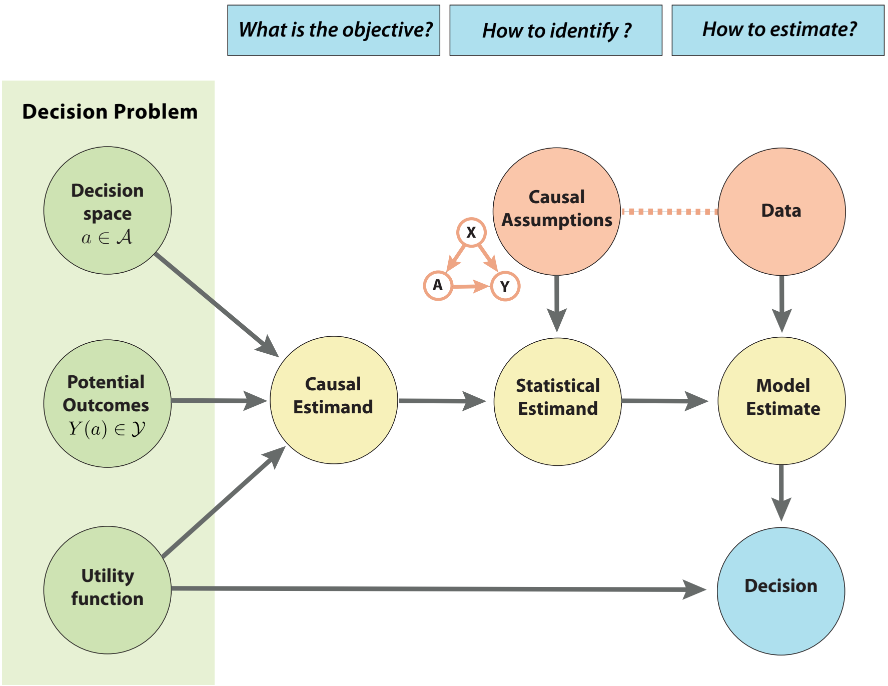

Figure 1: **Automated decision-making requires causal assumptions to ensure reliable decisions.** This figure shows the key components needed to set up an automated decision-making (ADM) system. Arrows indicate how one component informs or depends on the next. Green nodes describe the decision problem: what the goal is, what decisions can be made, and what outcomes are affected. Yellow nodes represent the quantities used to connect this problem to data: the causal estimand (the effect we want to know), the statistical estimand (what we can estimate from data), and the model estimate (what we actually compute). Red nodes represent external information needed to make this link: data and causal assumptions (such as no unmeasured confounders of decision A and outcome Y given covariates X) made about the data (dashed line). In practice, this framework can be extended to incorporate considerations such as operational constraints (for instance, resource constraints) or fairness constraints.

estimands, as well as to formalize when both are aligned. This shift in focus—from merely optimizing predictive accuracy to ensuring valid causal estimates—ensures that ADM systems can perform reliably in real-world settings.

A clear causal lens strengthens ADM systems by clarifying the assumptions required to ensure when and where decisions are reliable. Rather than seeing causal assumptions as a burden, we encourage to view them as a strength—a formal way to define the boundary conditions under which a system is designed to operate. Articulating these assumptions can also help in thinking about data requirements, such as identifying potential confounders that need to be measured. As such, transparency about causal assumptions not only strengthens trust when they hold but also clarifies when not to trust or deploy a model. Without causal reasoning, even seemingly well-performing models can fail in unpredictable ways, particularly in dynamic or complex environments. Nevertheless, we acknowledge that ADM models trained without a causal lens, or in contexts where assumptions are difficult to verify, may still perform well in certain cases. However, such models should be viewed as heuristics rather than robust solutions—which may be defendable for low-stakes tasks where causal reasoning is beneficial but not strictly needed.

So, what are the boundary conditions when a causal approach is mandatory, beneficial, or even harmful? A causal approach is generally mandatory in high-stakes scenarios where decisions directly affect safety, health, or legal outcomes. Examples include medical treatment recommendations or autonomous vehicle navigation, where understanding the causal effects of actions ensures reliability and mitigates risk. In such settings, unverifiable causal assumptions should serve as a ‘red flag’ for deploying ADM systems, as flawed causal reasoning can result in harmful consequences. Conversely, a causal lens is beneficial—but often not strictly necessary—in low-stakes or routine decisions, such as music recommendations or personalized marketing offers, where the cost of errors is minimal. Yet, a causal approach often leads to a better performance, and, hence, companies such as Spotify make broad use of causal approaches in their ADM tasks for business reasons [10]. However, in settings where causal relationships and the underlying assumptions are entirely speculative, causal claims may even be harmful by introducing unwarranted complexity, uncertainties, and even incorrect beliefs about a system’s reliability. For instance, if an ADM system for marketing decisions mistakenly assumes that all relevant confounders (such as costumer motivation or seasonal trends) have been accounted for, it may falsely attribute increased sales entirely to a specific promotion, giving decision-makers undue confidence.

Another benefit of adopting a causal perspective in ADM is the ability to address transportability—that is, the transfer of causal results learned in one environment (for instance, a training setting) to another (for instance, a deployment setting). Concepts such as causal transportability [11] provide formal tools to specify the assumptions under which ADM systems can reliably generalize across settings. For example, a model trained to optimize traffic flow in one city may encounter differences in road infrastructure, weather patterns, or driver behavior when deployed in another. A causal approach can formalize which relationships remain invariant across these environments and help adjust the ADM system accordingly. Hence, transportability hinges on causality because causal relationships are robust to shifts between environments. Likewise, such a causal approach may help avoid shortcut learning—a common issue where models exploit spurious correlations or superficial patterns in the training data that may fail under deployment conditions. By making the causal assumptions explicit, developers can systematically assess whether the model is learning meaningful, transportable relationships or merely relying on shortcuts.

4 Challenges and directions for future research

Several practical challenges remain that must be addressed to ensure ADM systems operate effectively in dynamic, real-world environments.

Data quality. The reliability of any ADM system depends on the quality of the training data. Distribution shifts—meaning the data distribution during deployment differs from that during training—pose complex challenges, especially in dynamic settings where conditions change over time. While causal reasoning improves generalizability by identifying relationships that are invariant across environments, more research is needed to develop ADM models that can safely adapt to distribution shifts and are robust to external shocks and adversarial attacks. Ultimately this requires models that can accurately quantify their prediction confidence across different deployment conditions as well as novel monitoring techniques to assess ADM performance in dynamic data streams.

Data quality issues are multifaceted and can include measurement errors, representation biases, and incomplete data. These issues are well-studied in the area of governmental survey statistics, which has developed a rich toolkit for conceptualizing and assessing errors in data from a population inference perspective. Yet, data quality is not an absolute concept but rather depends on the specific application task. While survey science traditionally focuses on requirements for valid descriptive inference, causal modeling hinges on coverage, that is, a non-zero inclusion probability of groups in the training data that differ in their treatment response. We thus call for research on new data audits, quality metrics, and validation frameworks that are tailored to the goals of (different classes of) ADMs, for which survey science provides valuable starting points.

Uncertainty quantification and robustness. Assessing the uncertainty in outcomes, decisions, and the overall model is critical for ensuring the reliability of ADM systems in practice. For example, in medical decision-making, uncertainty estimates can help determine whether the probability of a benefit from treatment is sufficiently large to outweigh the risks of adverse reactions. While substantial progress has been made in quantifying uncertainty for predictive machine learning, extending these methods to ADM settings presents new challenges such as additional uncertainty due to violations of identifiability assumptions or due to low treatment overlap. A causal perspective can help to formalize the different sources of uncertainty, that is, whether the uncertainty comes from a lack of identifiability of the causal estimand due to violations of causal assumptions (e.g., unobserved confounding) or from estimatability issues such as low overlap (e.g., certain individuals never receiving treatment) or lack of data.

A common challenge for ADM in practice is to assess whether causal assumptions, such as the absence of unobserved confounding, hold in practice. Future research should focus on developing new methods that help in assessing the plausibility of assumptions and that account for potential violations (such as partial identification and causal sensitivity analysis). Another interesting route is to derive methods that optimize decisions over an uncertainty set of potentially confounded policies and that thus find worst-case guarantees in the form of confounding-robust policies [12]. In addition, there is a need for tailored methods that account for all those sources of uncertainty simultaneously. For example, there exist various methods for partial identification and sensitivity analysis that account for uncertainty due to lack of identifiability, but more research is needed on how to augment these methods to account for finite sample uncertainty (for instance, via conformal prediction or Bayesian inference). Together, this can yield more systematic frameworks for robust ADM systems under a range of causal assumptions, which will enable practitioners to better understand and mitigate risks associated with imperfect causal models.

To improve the robustness of ADM systems, another direction is the use of causal world models [13], which abstract underlying causal mechanisms to improve generalization and adaptation across domains and different decision-making objectives, such as optimizing under varying constraints or over different outcomes. Finally, many decision-making settings involve multiple objectives, so tailored methods for multi-criteria decision-making are needed that balance a primary objective while minimizing the risk of harm. For instance, in medicine, drugs must achieve high efficacy while ensuring safety with minimal side effects [5].

Performativity. Deploying an ADM system introduces performativity—the phenomenon where predictions made by a system actively shape the target of prediction [14]. Such a feedback loop between model and data once again points to the need for causal reasoning in ADM to explicitly model how predictions influence future outcomes. For instance, in smart cities, a traffic control system that optimizes signal timing to reduce congestion might influence drivers' route choices, leading to new traffic patterns that the system did not initially anticipate. Similarly, an AI agent may shape the behavior of humans, causing the data-generating process to be different from what the AI agent was originally trained on. Traditional predictive models fail to account for this, as they assume a stable data-generating process and do not capture how interventions—such as continued deployment of the system—reshape the future training data. A recent area of research, often framed under the concept of "performative prediction", examines the causal influence of machine learning predictions on the very outcomes they aim to forecast. Various solution concepts and algorithms have been proposed, such as ensuring the stability of a system after repeated retraining. We encourage further research at the intersection of machine learning and causality—for example, using non-experimental data in settings where exploration is costly or restricted. Performative prediction also provides a valuable framework for studying more complex decision-making settings and strategic interactions, such as scenarios where multiple agents anticipate each other's actions and adapt their behavior accordingly. Such research will be especially important due to the growing adoption of AI agents in practice.

Evaluation and benchmarking. More research is needed to improve benchmarking in ADM systems, especially by developing new evaluation frameworks that reflect the complexities of real-world deployment settings. Since the estimands in ADM–such as counterfactual outcomes, treatment effects, or optimal actions—are generally unobservable, they cannot be traditionally benchmarked using standardized datasets. Rather, simulations and semi-synthetic datasets are typically the common approach to assess the performance of ADM systems [12]. Future work should prioritize the development of tailored benchmarks that incorporate realistic decision-making environments, accounting for dynamic conditions, feedback loops, and changing user behaviors. Promising directions for this include sequential A/B tests and bandit algorithms, which iteratively refine decision-making by adaptively allocating interventions based on observed outcomes, as well as adaptive clinical trials, which modify trial parameters in response to accumulating evidence to enhance efficiency and ethical considerations. These methods ensure benchmarks remain aligned with the evolving nature of decision environments. An alternative approach is to leverage real-world physical systems in a controlled environment as a testbed [15]. More research is also needed to construct better performance metrics that align closely with decision-making objectives, such as the quality of decision rules, downstream societal impacts, and performance across diverse subpopulations.

5 Conclusion

Causal reasoning is a necessary, but not sufficient condition for reliable decision-making in practice. ADM systems commonly involve deep interactions between algorithms and humans—

raising critical questions such as non-compliance with algorithmic decisions and disparate impact downstream. Yet, causal reasoning fosters transparency about the *conditions* under which ADM systems can be trusted to operate reliably, providing an invaluable building block for safe and robust decision-making. Causal reasoning offers a powerful but also necessary foundation for improving the safety and reliability of ADM. By ensuring that ADM systems capture the true effects of decisions, causal assumptions help align the systems with real-world objectives.

Acknowledgements

This work is supported by the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research.

References

[1] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806, 2017.

[2] Gabriel Dulac-Arnold, Nir Levine, Daniel J Mankowitz, Jerry Li, Cosmin Paduraru, Sven Gowal, and Todd Hester. Challenges of real-world reinforcement learning: Definitions, benchmarks and analysis. *Machine Learning*, 110(9):2419–2468, 2021.

[3] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447–453, 2019.

[4] Judea Pearl. *Causality: Models, Reasoning, and Inference.* Cambridge University Press, Cambridge, UK, 2009.

[5] Stefan Feuerriegel, Dennis Frauen, Valentyn Melnychuk, Jonas Schweisthal, Konstantin Hess, Alicia Curth, Stefan Bauer, Niki Kilbertus, Isaac S Kohane, and Mihaela van der

Schaar. Causal machine learning for predicting treatment outcomes. *Nature Medicine*, 30 (4):958–968, 2024.

[6] Charles F Manski. *Identification for prediction and decision.* Harvard University Press, Cambridge, US, 2009.

[7] Philip Dawid. Decision-theoretic foundations for statistical causality. *Journal of Causal Inference*, 9(1):39–77, 2021. doi: 10.1515/jci-2020-0008.

[8] Amanda Coston, Alan Mishler, Edward H. Kennedy, and Alexandra Chouldechova. Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, page 582–593, New York, NY, USA, 2020. Association for Computing Machinery. doi: 10.1145/3351095.3372851.

[9] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. Prediction policy problems. *American Economic Review*, 105(5):491–95, 2015. doi: 10.1257/aer.p20151023.

[10] Will Douglas Heaven. The complex math of counterfactuals could help Spotify pick your next favorite song. MIT Technology Review, 2023. URL `https://www.technologyreview.com/2023/04/04/1070885/complex-math-counterfactuals-spotify-predict-finance-healthcare/`.

[11] Elias Bareinboim and Judea Pearl. Transportability of causal effects: Completeness results. *Proceedings of the AAAI Conference on Artificial Intelligence*, 26(1):698–704, Sep. 2021. doi: 10.1609/aaai.v26i1.8232.

[12] Nathan Kallus and Angela Zhou. Confounding-robust policy improvement. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31. Curran Associates, Inc., 2018.

[13] Jonathan Richens and Tom Everitt. Robust agents learn causal world models. In *The Twelfth International Conference on Learning Representations*, 2024.

[14] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 7599–7609. PMLR, 2020.

[15] Juan L Gamella, Jonas Peters, and Peter Bühlmann. Causal chambers as a real-world physical testbed for AI methodology. *Nature Machine Intelligence*, 7:107–118, 2025.