

Causal Inference Frameworks for Trustworthy Decision-Making in Autonomous Agents

Gurumurthy S^{1*}, Natarajan P² and Subramaniam A³

¹Department of Computer Science and Engineering, Noyyal Center for Future Computing Architectures, Philippines

Email: shalini.gurumurthy@noyyal.edu

ORCID: <https://orcid.org/2822-6221-0004-0713>

²Department of Intelligent Systems, Bharathapuzha Center for Neural Computation, Indonesia Email:

priyanka.natarajan@bharathapuzha.edu

ORCID: <https://orcid.org/5926-6278-5624-8086>

³Department of Artificial Intelligence and Data Science, Muvattupuzha Center for Computational Cognition, Sri

Lanka Email: anitha.subramaniam@muvattupuzha.edu

ORCID: <https://orcid.org/3363-0923-8982-6308>

ABSTRACT

Autonomous agents that only predict, without reasoning about intervention and counterfactual outcomes, cannot reliably answer the question every acting system implicitly faces: what would happen if I did this instead. This article synthesizes the literature on causal inference frameworks as a foundation for trustworthy decision-making in autonomous agents, tracing the path from foundational estimation tools such as propensity scores, marginal structural models, and invariant prediction, through causal discovery methods that learn structure from data rather than assume it, to causal reinforcement learning architectures that fold counterfactual reasoning directly into an agent's policy loop. We give particular weight to results with concrete, verifiable numbers: counterfactual credit assignment lifting multi-agent win rates well above non-causal baselines, a targeted causal intervention recovering an imitation-learning agent from a spurious shortcut in a fraction of the queries a standard baseline needs, and a human study showing that causal explanations measurably improve a person's ability to predict an autonomous agent's next action. Read together, the evidence supports causal inference as a genuine, increasingly well-evidenced foundation for trustworthy autonomous decision-making, while leaving formal safety guarantees, causal discovery at real deployment scale, and the link between causal explanation and human trust as a substantially open research agenda.

Keywords: causal inference; autonomous agents; causal reinforcement learning; counterfactual explanation; trustworthy AI

Citation: Gurumurthy S, Natarajan P and Subramaniam A [2024]. Causal Inference Frameworks for Trustworthy Decision-Making in Autonomous Agents. DOI: <https://doi.org/10.5281/zenodo.22843893>

Copyright: © 2024 by the authors. Open access under [CC BY 4.0 license](https://creativecommons.org/licenses/by/4.0/).

Article Information: Received: 12 May 2024 Accepted: 11 August 2024 Published: 19 November 2024

Article Class: Research Article

1. Introduction

An autonomous agent that only predicts is, in an important sense, not yet making a decision at all. Predicting the next frame of a video, the next word in a sentence, or the expected reward of an action under the policy currently in force are all associative tasks: they ask what usually follows what, given how the world has behaved so far. Acting is different. An agent that recommends a treatment, reroutes a fleet of vehicles, throttles a robot's motor torque, or decides whether to intervene on a factory line is implicitly asking a counterfactual question -- what would happen if I did this, rather than that, in this particular situation -- and the associative machinery that dominates modern machine learning was never built to answer it reliably [1]. This gap between prediction and intervention has been understood in statistics and epidemiology for decades, formalized through potential outcomes and propensity scores [2] and through the graphical, structural-equation apparatus most closely associated with Judea Pearl's causal hierarchy [3]. What has changed over the past ten years is that this apparatus has moved from a largely observational, retrospective discipline -- did this policy work, did this drug help -- into something that increasingly sits inside the decision loop of software agents that act, in real time, on incomplete and shifting information.

This article synthesizes the literature on causal inference frameworks as a foundation for trustworthy decision-making in autonomous agents: reinforcement-learning policies, multi-agent controllers, imitation-learned driving systems, and the explanation layers built around them. We are not reporting new experiments. We are trying to answer a narrower and, we think, more useful question for a reader approaching this literature for the first time: what does it actually mean, mechanically, for an agent's decisions to be "causal," what has been demonstrated empirically about the benefits of that framing, and where does the evidence run out. The claim that causal reasoning makes agents more trustworthy is common in the literature and, in our reading, largely earned -- but it is earned unevenly, with strong empirical support in some corners (multi-agent credit assignment, imitation learning robustness) and considerably thinner support in others (formal safety guarantees, scalable discovery on real sensor data).

We organize the discussion around five themes. First, the foundational apparatus of causal inference -- the ladder of causation, structural causal models, and the estimation techniques (propensity scores, marginal structural models, invariant prediction) that let a system distinguish correlation from intervention effect. Second, causal discovery: how a system learns a causal structure from data in the first place, rather than assuming one is handed down by a domain expert. Third, causal reinforcement learning, where the causal apparatus is folded directly into an agent's policy-learning loop rather than applied after the fact to a trained model's behavior. Fourth, explainability, counterfactual reasoning, and the specific question of what makes a human trust, or rightly distrust, an autonomous agent's decision. Fifth, applications in autonomous driving and multi-agent coordination, where several of the clearest empirical results in this literature originate. We close with an honest accounting of open problems, because a synthesis that only reports successes would misrepresent a field that itself is candid about how much remains unresolved.

It is worth being explicit about scope from the outset. Causal inference is a large field with a much longer history than deep learning, spanning epidemiology, econometrics, and social science, and we do not attempt to survey it in full. Our focus is specifically on the subset of that literature that has been adapted for, or was designed for, autonomous decision-making systems: agents that must act under uncertainty, often with incomplete causal knowledge of their environment, and increasingly under the expectation that a human overseeing them can understand and contest the decisions those systems make [4,5].

2. Foundations of Causal Inference

The Ladder of Causation and Structural Causal Models

The organizing idea that runs through nearly every paper discussed in this article is Pearl's three-level hierarchy of causal reasoning: association (what is $P(Y|X)$, what correlates with what), intervention (what is $P(Y|\text{do}(X))$, what happens if an external agent forces X to a value), and counterfactuals (what would Y have been, for this specific unit, had X been different, given what we already observed about that unit). Schölkopf and colleagues' widely cited treatment of causal representation learning frames the entire deep learning

enterprise as having, so far, mostly climbed only the first rung of that ladder, and argues that many of the well-documented failure modes of modern machine learning systems -- brittleness under distribution shift, poor few-shot generalization, susceptibility to spurious correlations -- are symptoms of stopping at association when the underlying task actually requires intervention or counterfactual reasoning [1]. This framing matters for autonomous agents specifically because an agent's core operational loop is, definitionally, an intervention: it selects actions that change the state of the world, and then has to reason about consequences it has not directly observed for the specific action it is contemplating.

Formalizing this reasoning requires machinery beyond a purely statistical model. Structural causal models represent a system as a set of variables connected by directed edges encoding functional, not merely correlational, dependence, together with an explicit assumption about which variables are exogenous (outside the model's control) and which are endogenous (determined by other variables in the model). Bareinboim and Pearl's treatment of the data-fusion problem -- how to combine observational data, experimental data from a different population, and partial interventional data into a single valid causal estimate -- is a useful illustration of why this matters operationally rather than just philosophically: an autonomous agent deployed in the field routinely faces exactly this fusion problem, learning partly from logged historical behavior (observational), partly from its own exploratory actions (interventional), and partly from simulation (a third, separately biased data source), and naively pooling all three without accounting for how each was generated can silently corrupt the resulting policy [3].

Estimating Causal Effects Without Full Experiments

Much of applied causal inference exists because running a full randomized experiment for every decision an agent needs to make is either impossible, unethical, or prohibitively slow. Rosenbaum and Rubin's propensity-score framework, one of the most widely used tools in observational causal inference, addresses this directly: rather than requiring randomized assignment, it estimates the probability that a unit would have received a given treatment given its observed covariates, and uses that estimated probability

to construct comparisons between treated and untreated units that approximate what a randomized trial would have produced, under the assumption that all relevant confounders have been measured [2]. This idea generalizes naturally to sequential decision-making, which is closer to what an autonomous agent actually does: Robins, Hernán, and Brumback's marginal structural models extend the same logic to settings where treatment (or, for an agent, action) is chosen repeatedly over time and each choice can itself be confounded by variables affected by earlier choices -- precisely the structure of a reinforcement-learning trajectory, where an agent's state at time t is shaped by its own actions at times before t [6].

A complementary estimation strategy, invariant causal prediction, approaches the same underlying problem from a different angle: rather than modeling the assignment mechanism directly, Peters, Bühlmann, and Meinshausen ask which predictive relationships remain stable (invariant) across multiple environments or experimental conditions, and argue that only genuinely causal relationships should exhibit that stability, while merely correlational relationships will vary as the environment changes [7]. For an autonomous agent operating across multiple deployment contexts -- different cities for a delivery robot, different patients for a clinical decision-support system, different traffic conditions for a driving policy -- this invariance criterion offers a practical, testable way to separate features the agent should rely on from features that happen to correlate with good outcomes only in the training distribution it happened to see.

3. Causal Discovery from Data

None of the estimation machinery above is useful if the causal structure itself -- which variables cause which, and through what pathway -- is unknown, and for most real autonomous systems it genuinely is unknown; nobody hands a delivery robot or a recommendation engine a validated causal graph of its environment. Causal discovery is the subfield concerned with learning that structure from data, typically observational data, and it has its own decades-long history predating the recent surge of interest in causal machine learning. Glymour, Zhang, and Spirtes's review of graph-based causal discovery methods surveys the two dominant families of approach: constraint-based methods, which

use conditional independence tests to rule out edges inconsistent with the observed data (in the tradition of the PC algorithm), and score-based methods, which search directly over candidate graph structures using a fitness criterion that trades off model fit against complexity [8]. Vowels, Camgoz, and Bowden's later survey of the same territory, aimed specifically at a machine-learning rather than a statistics audience, adds a third, more recent family to this taxonomy -- continuous-optimization methods, discussed below -- and is useful precisely because it catalogues the practical failure modes (sensitivity to sample size, to the choice of independence test, to violations of the acyclicity assumption) that a reader coming from deep learning is less likely to already know to look for [9]. Spirtes and Zhang's companion review situates these methods within the broader landscape of causal discovery and inference, and is candid that both families depend on assumptions -- causal sufficiency (no unmeasured confounders), faithfulness (the data's independence structure exactly reflects the graph's structure), and, in many practical implementations, acyclicity -- that are convenient for proving theorems but frequently violated by the sensor-rich, feedback-laden environments autonomous agents actually operate in [10].

A more recent development reframes causal discovery as continuous optimization rather than combinatorial search over the (super-exponentially large) space of possible directed acyclic graphs. Zheng, Aragam, Ravikumar, and Xing's NOTEARS method introduces a smooth, differentiable characterization of the acyclicity constraint, which allows structure learning to be posed as a continuous optimization problem solvable with standard gradient-based methods rather than the discrete search that constraint- and score-based methods traditionally require [11]. On the Sachs et al. protein-signaling benchmark, a widely used real-data test case for causal discovery methods, the authors report that NOTEARS recovers a graph with 16 edges and a structural Hamming distance of 22 against the accepted ground-truth network, essentially matching the 17-edge, SHD-22 result obtained by the FGS score-based baseline on the same data -- a useful concrete illustration that reframing the search as continuous optimization does not sacrifice discovery accuracy relative to established discrete methods, while opening the door to the same gradient-based tooling (automatic differentiation, GPU acceleration) that has

driven progress in deep learning more broadly [11]. The broader causal-machine-learning survey by Kaddour, Lynch, Liu, Kusner, and Silva situates NOTEARS-style continuous-optimization discovery within a wider taxonomy of open problems connecting causal discovery, causal effect estimation, and downstream machine learning tasks, and argues, in a point we return to in the open-challenges section, that discovery methods validated on small synthetic or curated benchmarks like Sachs et al.'s protein network have been tested far less thoroughly on the high-dimensional, non-stationary, partially observed data streams that real autonomous agents actually generate [12].

Yao, Chu, Li, Li, Gao, and Zhang's survey on causal inference, and Guo, Cheng, Li, Hahn, and Liu's companion survey on learning causality with data, both make a related methodological point that is easy to lose sight of when reading individual discovery papers in isolation: causal discovery and causal effect estimation are frequently treated as separate stages in a pipeline -- first learn the graph, then estimate effects conditional on it -- but errors in the first stage propagate, often silently, into the second, and relatively few papers in either survey's coverage area report how sensitive their downstream effect estimates are to plausible errors in the discovered graph [13,14]. For an autonomous agent that discovers structure online, incrementally, as it collects new experience, this compounding-error problem is not a peripheral concern but close to a central design constraint. The difficulty compounds further once the data an agent observes is genuinely time-indexed rather than a set of exchangeable samples, which is the normal case for a sensor-driven agent rather than the exception: Runge and colleagues' survey of causal discovery from time series in Earth system science, developed outside the autonomous-agents literature entirely but directly applicable to it, catalogues the additional assumptions (about time lags, about whether the underlying process is stationary, about whether feedback loops close within the observed sampling interval) that time-indexed discovery methods require beyond the already substantial list needed for the static, cross-sectional case [15].

4. Causal Reinforcement Learning for Autonomous Decision-Making

Reinforcement learning is, on its face, already an intervention-based paradigm -- an agent takes actions

and observes their consequences, which is structurally closer to Pearl's second rung than to pure association. In practice, however, standard reinforcement learning algorithms frequently reason about the value of an action using signals that are confounded by the very policy generating the data, by other agents acting in the same environment, or by spurious features of the observation that happen to correlate with reward during training but do not cause it. Causal reinforcement learning is the strand of work that makes the causal structure of the decision problem explicit, rather than leaving it implicit in a value function or a policy network, and Zeng, Cai, Sun, Huang, and Hao's survey of the field organizes existing approaches broadly into methods that use causal structure to improve sample efficiency and generalization, and methods that use it to support explanation, fairness, or robustness -- a division that maps reasonably well onto the estimation-versus-explanation split this article follows more generally [16].

Table 1: Representative causal-inference techniques for autonomous-agent decision-making and the quantitative effect each cited study reports for its causal mechanism relative to a non-causal baseline. [17,19,11,24].

Technique	Source	Causal Mechanism	Reported Effect vs. Baseline
Counterfactual multi-agent policy gradients (COMA)	Foerster et al. [17]	Per-agent counterfactual reward baseline for credit assignment	87% win rate (3m map) vs. 47-56% for independent actor-critic baselines
Causal-confusion-aware imitation learning	de Haan et al. [19]	Targeted intervention to remove a spurious causal shortcut	Reward recovers from -160 to -130 (MountainCar) within 10-15 queries vs. hundreds for DAgger
NOTEARS continuous-optimization causal discovery	Zheng et al. [11]	Smooth acyclicity constraint replacing combinatorial graph search	16 edges, SHD 22 on Sachs protein network, matching the 17-edge, SHD-22 FGS baseline
Causal-lens explanation of RL agent behavior	Madumal et al. [24]	Structural causal model of agent action-influence for counterfactual explanation	Task-prediction score 10.90 (causal) vs. 8.53 (no explanation), $p = 0.003$

A particularly instructive line of work addresses credit assignment in multi-agent reinforcement learning, where the central difficulty is causal in a very literal sense: when a team of agents receives a single shared reward signal, no individual agent can easily tell how much of that reward its own action actually caused, as opposed to the actions its teammates happened to take at the same time. Foerster, Farquhar, Afouras, Nardelli, and Whiteson's counterfactual multi-agent policy

gradient method (COMA) addresses this directly by computing, for each agent, a counterfactual baseline: what the team's expected reward would have been had that one agent taken a different, fixed default action while every other agent's action stayed exactly as executed. Subtracting this counterfactual baseline from the actual outcome isolates the marginal, causal contribution of each agent's specific choice, and on the StarCraft unit-micromangement benchmark this yields a clear empirical gain -- on the three-marine (3m) map, COMA agents win 87% of episodes against decentralized independent actor-critic baselines that win 47% and 56%, and the improvement holds, though with a smaller margin, on larger and more heterogeneous unit compositions such as five wraiths (82% for COMA versus 57-70% for the tested baselines) [17]. That the specific mechanism responsible for the gain is an explicitly counterfactual computation, rather than a generically larger or more expressive network, is exactly the kind of causal ablation this literature was slower to provide before COMA and has become more common since.

A second, closely related line of work asks what happens when an agent's training data is confounded not by other agents but by unobserved variables shaping both the actions it took and the outcomes it observed -- the classic setup a causal-inference researcher would recognize immediately, applied to a sequential decision problem. Buesing and colleagues' counterfactually-guided policy search reframes off-policy evaluation and policy improvement as a structural counterfactual inference problem: given a trajectory generated under one policy, the method estimates what would have happened under a different, counterfactual policy by explicitly modeling the exogenous noise variables that, together with the policy, determined the observed trajectory, rather than relying purely on importance sampling over trajectories [18]. The practical payoff the authors report is that this counterfactual approach can produce lower-variance policy-improvement estimates than standard model-based or importance-sampling baselines in the partially observed environments they test, precisely because it uses the specific noise realization behind each observed trajectory rather than treating that trajectory as one exchangeable sample among many.

De Haan, Jayaraman, and Levine's work on causal confusion in imitation learning identifies a failure mode

that is, in a sense, the mirror image of the credit-assignment problem COMA addresses: an imitation-learning agent given access to more information about the expert's state -- including, notably, the expert's own previous action -- can perform worse, not better, than an agent given less information, because the additional feature lets the learned policy latch onto a spurious causal shortcut (imitating the previous action rather than the state that actually justified it) that fails catastrophically once the agent is deployed and has to act on its own previous actions rather than the expert's. Concretely, the authors report that on the MountainCar task, a behavioral-cloning policy confounded in this way collapses to roughly -160 average reward, compared to roughly -130 for a policy without the confounding shortcut available, and that their proposed causal intervention procedure -- querying either the environment or an expert to disambiguate the true causal structure -- recovers performance back to approximately -130 within about ten to fifteen targeted queries, versus the "hundreds of queries" the authors report a standard DAgger-style baseline needs to reach comparable performance [19]. This result is a clean, small-scale demonstration of exactly the risk the causal-inference literature warns about in the abstract: more data, or more features, is not unconditionally good for an agent's decision quality if that additional signal is causally spurious, and the fix is not more data of the same kind but a targeted causal intervention that disambiguates cause from confound.

5. Explainability, Counterfactuals, and Trust in Autonomous Agents

Trustworthy decision-making is not only a property of whether an agent's internal reasoning is causally sound; it is also, unavoidably, a property of whether the humans overseeing, deploying, or subject to that agent's decisions can understand and, where appropriate, contest them. This is the point at which the causal-inference and explainable-AI literatures converge most directly. Rudin's widely cited argument for interpretable-by-design models over post-hoc explanations of black boxes is not, on its face, a causal-inference paper, but its central claim -- that a post-hoc explanation can be an unfaithful, potentially misleading account of what a black-box model actually did, particularly for high-stakes decisions -- is precisely the concern that motivates using an explicit causal model as the explanation mechanism rather than a

correlational approximation fit after the fact [20]. Doshi-Velez and Kim's earlier call for a more rigorous science of interpretability makes a related, complementary point: that "interpretability" is not a single well-defined property but a cluster of different desiderata (simulatability, decomposability, algorithmic transparency) that need to be evaluated against the specific task and audience at hand, rather than assumed to transfer automatically from one explanation method to another [21]. Bhatt and colleagues' study of explainability in deployment adds an empirical, organizational dimension often missing from the purely technical literature: interviewing practitioners at a range of companies actually shipping explainable-ML systems, they find that explanations are used internally far more often for model debugging by engineers than for end-user-facing transparency, and that many deployed "explanation" pipelines are not tied to any rigorous evaluation of whether the explanations are actually faithful to the model's decision process [22] -- a sobering data point for anyone assuming that adding an explanation layer to an autonomous agent automatically makes it more trustworthy.

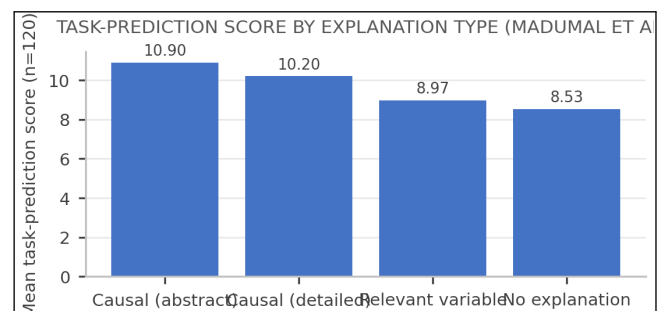


Figure 1: Mean task-prediction scores from Madumal et al.'s 120-participant human study, comparing an abstract causal-model explanation of an RL agent's behavior against a relevant-variable baseline without causal structure and a no-explanation control -- causal explanations let participants predict the agent's next action more accurately ($p = 0.003$ overall) [24].

Counterfactual explanation is the specific explanation format that sits closest to the causal-inference apparatus described earlier in this article, because a counterfactual explanation is, definitionally, an answer to a causal question: what is the minimal change to this situation that would have produced a different decision. Wachter, Mittelstadt, and Russell's foundational treatment of counterfactual explanations, developed in the context of the EU's General Data Protection Regulation's right to explanation, argues that counterfactual explanations have a specific practical advantage over feature-importance-style explanations: they can be

actionable (telling a person what they could change to obtain a different outcome) without requiring the model's internal logic to be disclosed at all, which matters both for protecting proprietary models and, separately, for producing explanations a non-technical recipient can actually use [4]. Verma, Boonsanong, Hoang, Hines, Dickerson, and Shah's review of counterfactual explanations and algorithmic recourse surveys the substantial body of work that followed, organizing methods along dimensions including whether the counterfactual respects the causal structure of the underlying features (changing "years of education" without correspondingly changing "age," for instance, may describe a change that is counterfactually meaningless) and whether the recommended counterfactual is actually actionable by the person or system receiving it [5]. Karimi, Barthe, Schölkopf, and Valera's complementary survey of algorithmic recourse pushes this causal-structure requirement further, arguing that recourse recommendations generated without respecting the causal relationships between features can recommend changes that are either physically impossible or that fail to produce the promised outcome once the full causal structure -- rather than a naive feature-by-feature perturbation -- is taken into account, and that recourse generation is therefore, properly understood, a causal inference problem rather than a purely optimization-based one [23].

For autonomous agents specifically, rather than static classifiers, Madumal, Miller, Sonenberg, and Vetere's work on explainable reinforcement learning through a causal lens is among the most directly relevant results in this literature, because it evaluates causal explanation not just conceptually but with a human study measuring whether causal explanations actually improve a person's understanding of an agent's behavior. The authors build a structural causal model of an RL agent's action-influence relationships and generate explanations by tracing counterfactual contrasts through that model, then evaluate the resulting explanations against several baselines (a more detailed causal-chain explanation, a relevant-variable baseline without causal structure, and a no-explanation control) in a human study with 120 participants across three domains (Cartpole, StarCraft II, and MountainCar). Task-prediction scores -- how well a participant could predict what the agent would do next after receiving an

explanation -- averaged 10.90 for the abstract causal-model explanation and 10.20 for the more detailed causal explanation, against 8.97 for the relevant-variable baseline and 8.53 for no explanation at all, a difference the authors report as statistically significant ($p = 0.003$ overall, $p = 0.006$ for the causal-versus-relevant-variable comparison specifically) [24]. Notably, and worth reporting honestly rather than selectively, the same study found that self-reported trust ratings were higher on average for the causal-explanation conditions but did not reach statistical significance -- a useful reminder that causal explanations measurably improving a person's ability to predict an agent's behavior is not automatically the same thing as measurably improving how much that person trusts the agent, and the literature has, so far, demonstrated the former considerably more convincingly than the latter.

6. Applications: Autonomous Driving and Multi-Agent Systems

Several of the clearest empirical demonstrations of causal reasoning's value for autonomous agents come from domains where the cost of a wrong decision is immediate and physical, which is also where the case for treating causal soundness as a trustworthiness requirement, rather than a nice-to-have, is easiest to make concretely. Autonomous driving is the paradigm case: de Haan and colleagues' causal-confusion work, discussed above primarily through its MountainCar results, was directly motivated by, and also tested against, real driving data, citing external results showing that a history-based imitation-learning driving policy -- one given access to its own recent action history, exactly the kind of feature that can encode a spurious causal shortcut -- was measured at 2.94 (± 1.79) collisions on average, compared to 3.38 (± 2.55) for a history-free variant that lacked this potentially confounding feature but performed worse on offline validation metrics, illustrating precisely the gap the authors' central argument warns about: a model can look better by conventional offline validation while performing worse, or in this case only ambiguously better, once causal structure (in this case, the dependence of the policy on its own action history) is taken seriously [19]. This is a useful corrective for anyone inclined to treat offline validation accuracy as a sufficient proxy for deployed decision quality in a safety-relevant autonomous system.

Multi-agent coordination is the second domain where the causal-inference literature has produced its clearest wins, largely because the underlying problem -- disentangling one agent's causal contribution from a shared, entangled outcome -- has no adequate non-causal solution; averaging or crediting reward proportionally to some heuristic simply does not answer the "what would have happened if only I had acted differently" question that credit assignment requires. COMA's win-rate improvements on the StarCraft micromanagement benchmark, reported above, are the best-documented instance of this, and the authors' broader argument -- that the best-performing COMA agents are competitive with centralized controllers that have access to information no individual decentralized agent could actually observe -- is a meaningful trustworthiness claim in its own right: a multi-agent system whose per-agent decisions are grounded in a genuine counterfactual credit signal is, at least in principle, more auditable after the fact than one whose credit assignment is an opaque byproduct of gradient descent through a shared value function [17].

It is worth noting what this application literature does not yet show, because the gap is informative. Almost none of the empirical results synthesized in this section come from field-deployed systems operating on physical hardware in uncontrolled environments; COMA's evaluation is a StarCraft micromanagement simulator, and de Haan et al.'s driving results, while grounded in real driving data collected by others, are evaluated primarily in the MountainCar, Hopper, and Pendulum simulated-control settings for their own proposed method, with the driving numbers imported from an external study rather than generated as part of the causal-confusion intervention experiments themselves. The pattern here echoes the causal-discovery literature discussed earlier: methods validated cleanly in simulation, with a real-world data point cited as motivation or partial corroboration rather than as the primary evidence base.

7. Open Challenges

Several gaps recur across the literature synthesized in this article, and naming them plainly is, we think, more useful to a reader than smoothing over them. First, and most pervasive: the overwhelming majority of the strongest empirical results here -- COMA's win rates, the causal-confusion recovery curves, NOTEARS's

structure-recovery accuracy -- come from simulated benchmarks (StarCraft micromanagement, MountainCar, Hopper, the Sachs protein network) chosen for experimental tractability rather than measured on deployed autonomous systems operating in physically or organizationally realistic conditions [17,19,11]. Kaddour et al.'s survey makes this point at the level of the field as a whole, arguing that causal machine learning's benchmark culture has not yet caught up to the complexity of the deployment settings the field's motivating examples invoke [12].

Second, causal discovery at the scale and non-stationarity real autonomous agents actually face remains substantially harder than the benchmark literature suggests. Glymour, Zhang, and Spirtes's and Spirtes and Zhang's reviews are both candid that the core assumptions underlying constraint- and score-based discovery -- causal sufficiency, faithfulness, and (for most methods, including NOTEARS) acyclicity -- are convenient rather than realistic for the feedback-rich, partially observed, continuously drifting data streams an autonomous agent generates as it operates [8,10,11]. An agent that discovers causal structure online, incrementally, from its own experience faces a version of this problem that the largely batch-oriented discovery literature has not yet addressed in depth.

Third, the relationship between causal explanation and human trust is measured far less rigorously, and with considerably weaker results so far, than the relationship between causal explanation and task-prediction accuracy. Madumal et al.'s human study is unusual in this literature precisely because it measured trust directly rather than assuming it follows automatically from improved understanding, and the fact that its trust results did not reach statistical significance, even as its prediction-accuracy results did, is worth taking seriously rather than treating as a minor footnote [24]. Bhatt et al.'s finding that deployed explanation systems are used mostly for internal debugging rather than external accountability compounds this concern: an explanation layer built for one purpose (helping an engineer understand a model) may simply not be the artifact needed for a different purpose (helping an affected person trust or contest a decision), and causal structure alone does not resolve that mismatch [22].

Fourth, counterfactual explanation and algorithmic

recourse, despite the causal rigor Karimi et al. and Verma et al. both argue for, still largely lack agreed-upon evaluation standards for what makes a recourse recommendation good -- actionable, minimal, causally valid, and robust to the recipient's actual capacity to act on it are all criteria the literature invokes, but rarely all simultaneously in a single evaluation [23,5]. Fifth, and more foundational: Amodei and colleagues' catalogue of concrete AI safety problems -- reward hacking, unsafe exploration, robustness to distributional shift, and scalable oversight, among others -- predates most of the causal-reinforcement-learning literature discussed here, and while causal structure plausibly helps with several of these problems (distributional shift robustness in particular, given the invariant-prediction argument of Peters et al.), the connection between causal RL specifically and formal safety guarantees remains, in Zeng et al.'s own survey's assessment, one of the least developed corners of the field relative to sample efficiency and interpretability, which have received comparatively more attention [25,7,16].

We do not think these gaps undermine the case for causal inference as a foundation for trustworthy autonomous decision-making; if anything, the specificity and reproducibility of results like COMA's counterfactual credit assignment and Madumal et al.'s task-prediction improvements suggest the underlying mechanism is real and useful. But an honest synthesis has to report that the strongest evidence to date is concentrated in simulated benchmarks and in the "does this help humans understand the agent" question, while the "does this make the agent formally safer" and "does this scale to real, messy, non-stationary deployment data" questions remain considerably more open.

8. Conclusion

The literature reviewed here supports a fairly specific and, we think, defensible claim: framing an autonomous agent's decision-making explicitly in causal terms -- structural causal models to represent the agent's environment, causal discovery to learn that structure rather than assume it, counterfactual reasoning inside the reinforcement-learning loop itself, and counterfactual explanation as the interface between the agent's decisions and the humans overseeing them -- produces measurable, reproducible improvements over purely associative alternatives in the specific settings where it has been carefully tested. Multi-agent credit

assignment is the clearest case: COMA's counterfactual baseline produces large, consistent win-rate gains over non-causal multi-agent actor-critic baselines precisely because it answers a genuinely causal question (what did this agent's action contribute) that no purely correlational credit-assignment scheme can answer as directly [17]. Imitation learning's causal-confusion problem is the clearest cautionary case in the other direction, showing concretely that more information handed to a learning agent can degrade decision quality when that information encodes a spurious causal shortcut, and that a targeted causal intervention, not more data of the same kind, is the fix [19].

None of this amounts to a claim that causal inference has solved trustworthy autonomous decision-making. The foundational estimation and discovery machinery -- propensity scores, marginal structural models, invariant prediction, constraint- and score-based discovery, and their more recent continuous-optimization descendants -- carries assumptions (causal sufficiency, faithfulness, acyclicity, correctly measured confounders) that are frequently violated by the feedback-laden, non-stationary, partially observed environments real autonomous agents inhabit [2,6,7,8,10,11]. The explanation and trust literature has demonstrated, carefully and with real human-subjects data, that causal explanations measurably improve a person's ability to predict what an agent will do, but has not yet demonstrated with the same rigor that this translates into measurably greater trust, still less into better outcomes for the people affected by an agent's decisions [24,22,20]. And the connection between causal reasoning and the kind of formal safety guarantees the field has long said it wants remains, on the evidence synthesized here, one of the least mature parts of the whole research program [25,16]. What we can say with confidence is that independent groups working on quite different problems -- multi-agent StarCraft coordination, robotic and driving imitation learning, algorithmic recourse for high-stakes classifiers, human-subjects evaluation of RL explanations -- have converged on structurally similar causal machinery and have, in the cases where they measured it carefully, found real, replicable benefits. That convergence, more than any single result, is what makes causal inference frameworks a genuine and still-developing foundation for trustworthy decision-making in autonomous agents, rather than a fashionable relabeling of tools the field

already had.

9. References

- Schölkopf, B., Locatello, F., Bauer, S., Ke, N.R., Kalchbrenner, N., Goyal, A., Bengio, Y. (2021) Toward Causal Representation Learning. *Proceedings of the IEEE*, 109(5): 612-634. DOI: [10.1109/JPROC.2021.3058954](https://doi.org/10.1109/JPROC.2021.3058954)
- Rosenbaum, P.R., Rubin, D.B. (1983) The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70(1): 41-55. DOI: [10.1093/biomet/70.1.41](https://doi.org/10.1093/biomet/70.1.41)
- Bareinboim, E., Pearl, J. (2016) Causal Inference and the Data-Fusion Problem. *Proceedings of the National Academy of Sciences*, 113(27): 7345-7352. DOI: [10.1073/pnas.1510507113](https://doi.org/10.1073/pnas.1510507113)
- Wachter, S., Mittelstadt, B., Russell, C. (2017) Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2): 841-887. DOI: [10.48550/arXiv.1711.00399](https://doi.org/10.48550/arXiv.1711.00399)
- Verma, S., Boonsanong, V., Hoang, M., Hines, K.E., Dickerson, J.P., Shah, C. (2020) Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review. *arXiv preprint*. DOI: [10.48550/arXiv.2010.10596](https://doi.org/10.48550/arXiv.2010.10596)
- Robins, J.M., Hernán, M.A., Brumback, B. (2000) Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology*, 11(5): 550-560. DOI: [10.1097/00001648-200009000-00011](https://doi.org/10.1097/00001648-200009000-00011)
- Peters, J., Bühlmann, P., Meinshausen, N. (2016) Causal Inference by Using Invariant Prediction: Identification and Confidence Intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5): 947-1012. DOI: [10.1111/rssb.12167](https://doi.org/10.1111/rssb.12167)
- Glymour, C., Zhang, K., Spirtes, P. (2019) Review of Causal Discovery Methods Based on Graphical Models. *Frontiers in Genetics*, 10: 524. DOI: [10.3389/fgene.2019.00524](https://doi.org/10.3389/fgene.2019.00524)
- Vowels, M.J., Camgoz, N.C., Bowden, R. (2022) D'ya Like DAGs? A Survey on Structure Learning and Causal Discovery. *ACM Computing Surveys*, 55(4): 82. DOI: [10.1145/3527154](https://doi.org/10.1145/3527154)
- Spirtes, P., Zhang, K. (2016) Causal Discovery and Inference: Concepts and Recent Methodological Advances. *Applied Informatics*, 3: 3. DOI: [10.1186/s40535-016-0018-x](https://doi.org/10.1186/s40535-016-0018-x)
- Zheng, X., Aragam, B., Ravikumar, P., Xing, E.P. (2018) DAGs with NO TEARS: Continuous Optimization for Structure Learning. *Advances in Neural Information Processing Systems (NeurIPS)* 31. DOI: [10.48550/arXiv.1803.01422](https://doi.org/10.48550/arXiv.1803.01422)
- Kaddour, J., Lynch, A., Liu, Q., Kusner, M.J., Silva, R. (2022) Causal Machine Learning: A Survey and Open Problems. *arXiv preprint*. DOI: [10.48550/arXiv.2206.15475](https://doi.org/10.48550/arXiv.2206.15475)
- Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., Zhang, A. (2021) A Survey on Causal Inference. *ACM Transactions on Knowledge Discovery from Data*, 15(5): 74. DOI: [10.1145/3444944](https://doi.org/10.1145/3444944)
- Guo, R., Cheng, L., Li, J., Hahn, P.R., Liu, H. (2020) A Survey of Learning Causality with Data: Problems and Methods. *ACM Computing Surveys*, 53(4): 75. DOI: [10.1145/3397269](https://doi.org/10.1145/3397269)
- Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., Kretschmer, M., Mahecha, M.D., Muñoz-Mari, J., van Nes, E.H., Peters, J., Quax, R., Reichstein, M., Scheffer, M., Schölkopf, B., Spirtes, P., Sugihara, G., Sun, J., Zhang, K., Zscheischler, J. (2019) Inferring Causation from Time Series in Earth System Sciences. *Nature Communications*, 10: 2553. DOI: [10.1038/s41467-019-10105-3](https://doi.org/10.1038/s41467-019-10105-3)
- Zeng, Y., Cai, R., Sun, F., Huang, L., Hao, Z. (2023) A Survey on Causal Reinforcement Learning. *arXiv preprint*. DOI: [10.48550/arXiv.2302.05209](https://doi.org/10.48550/arXiv.2302.05209)
- Foerster, J.N., Farquhar, G., Afouras, T., Nardelli, N., Whiteson, S. (2018) Counterfactual Multi-Agent Policy Gradients. *Proceedings of the AAAI Conference on Artificial Intelligence*. DOI: [10.48550/arXiv.1705.08926](https://doi.org/10.48550/arXiv.1705.08926)
- Buesing, L., Weber, T., Zwols, Y., Racanière, S., Guez, A., Lespiau, J.B., Heess, N. (2019) Woulda, Coulda, Shoulda: Counterfactually-Guided Policy Search. *International Conference on Learning Representations (ICLR)*. DOI: [10.48550/arXiv.1811.06272](https://doi.org/10.48550/arXiv.1811.06272)
- de Haan, P., Jayaraman, D., Levine, S. (2019) Causal Confusion in Imitation Learning. *Advances in Neural Information Processing Systems (NeurIPS)* 32. DOI: [10.48550/arXiv.1905.11979](https://doi.org/10.48550/arXiv.1905.11979)
- Rudin, C. (2019) Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*, 1(5): 206-215. DOI: [10.1038/s42256-019-0048-x](https://doi.org/10.1038/s42256-019-0048-x)
- Doshi-Velez, F., Kim, B. (2017) Towards A Rigorous Science of Interpretable Machine Learning. *arXiv preprint*. DOI: [10.48550/arXiv.1702.08608](https://doi.org/10.48550/arXiv.1702.08608)
- Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J.M.F., Eckersley, P. (2020) Explainable Machine Learning in Deployment. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*)*. DOI: [10.1145/3351095.3375624](https://doi.org/10.1145/3351095.3375624)
- Karimi, A.H., Barthe, G., Schölkopf, B., Valera, I. (2022) A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations. *ACM Computing Surveys*, 55(5): 95. DOI: [10.1145/3527848](https://doi.org/10.1145/3527848)
- Madumal, P., Miller, T., Sonenberg, L., Vetere, F. (2020) Explainable Reinforcement Learning Through a Causal Lens. *Proceedings of the AAAI Conference on Artificial*

Intelligence. DOI: [10.48550/arXiv.1905.10958](https://doi.org/10.48550/arXiv.1905.10958)

25. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D. (2016) Concrete Problems in AI Safety. arXiv preprint. DOI: [10.48550/arXiv.1606.06565](https://doi.org/10.48550/arXiv.1606.06565)

10. Declarations

Funding

This literature synthesis received no dedicated grant funding from public, commercial, or not-for-profit agencies.

Conflict of Interest

The authors declare that they have no competing interests relevant to this work.

Data Availability Statement

This article is a literature synthesis and did not involve the generation of new data; all numerical results discussed are drawn from, and attributed to, the cited sources, which readers should consult for the underlying data and code.

Ethical Approval

Not applicable. This article involved no human participants, animal subjects, or primary data collection by the authors.