Exploiting Causal Structure for Transportability in Online, Multi-Agent Environments

Axel Browne
Loyola Marymount University
Los Angeles, CA, USA
damianabrowne@gmail.com

Andrew Forney
Loyola Marymount University
Los Angeles, CA, USA
andrew.forney@lmu.edu

ABSTRACT

Autonomous agents may encounter the transportability problem when they suffer performance deficits from training in an environment that differs in key respects from that in which they are deployed. Although a causal treatment of transportability has been studied in the data sciences, the present work expands its utility into online, multi-agent, reinforcement learning systems in which agents are capable of both experimenting within their own environments and observing the choices of agents in separate, potentially different ones. In order to accelerate learning, agents in these Multi-agent Transport (MAT) problems face the unique challenge of determining which agents are acting in similar environments, and if so, how to incorporate these observations into their policy. We propose and compare several agent policies that exploit local similarities between environments using causal selection diagrams, demonstrating that optimal policies are learned more quickly than in baseline agents that do not. Simulation results support the efficacy of these new agents in a novel variant of the Multi-Armed Bandit problem with MAT environments.

KEYWORDS

Graphical Causal Models; Transportability; Multi-Agent Learning; Reinforcement Learning

ACM Reference Format:
Axel Browne and Andrew Forney. 2022. Exploiting Causal Structure for Transportability in Online, Multi-Agent Environments. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Online, May 9–13, 2022*, IFAAMAS, 9 pages.

1 INTRODUCTION

The *transportability problem* [3] is broadly defined as the task of taking data obtained in one environment, and using it to support inference in another, potentially different, setting. This task has been historically approached by the empirical sciences, largely in the endeavor of generalizing effects from laboratory settings as they apply to less controlled reality; in this domain, transportability has been studied under a number of names, including *generalizability* [7, 14] and *external validity* [6, 8]. Yet, recent developments in the study of causality have provided a graphical formalization that can be used to proceduralize the transport of causal effects across environments with shared causal structure [4, 23], yielding tools that may feature prominently in the design of more dynamic artificial agents that can adapt to differences between environments.

Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), P. Faliszewski, V. Mascardi, C. Pelachaud, M.E. Taylor (eds.), May 9–13, 2022, Online.

Various endeavors in modern machine learning mirror the objectives of transportability, including *transfer learning*, [5, 9, 27, 29] and *model generalizability*, [13, 15, 18, 25], with examples ranging from game playing agents taking successes on certain levels and using those experiences to succeed in others, to autonomous driving agents that are trained in simulations and expected to perform on actual roads.

As intelligent agents become increasingly integrated into daily life across many diverse environments, such as self-driving cars in different climates and traffic conditions, it may behoove the community of agents to not only learn from the observed experiences of one another, but to also acknowledge which parts of their environments are similar or different, thus harnessing the transportable portions of observations for the purpose of accelerating learning of optimal policies in their own domain. This work thus endeavors to examine the transportability problem in online, multi-agent systems by concerting the often disparate schools of graphical causal inference (which has primarily operated in the offline data-scientific domain), model-based data fusion for online reinforcement learners (in which causal tools have recently gained traction [11, 12, 21]), and multi-agent systems [1, 28] (in which causal models may yield promising future study).

The novel contributions of this work are thus as follows:

- Formalizes the novel problem of an online, Multi-agent Transport (MAT) environment as a variant of a Multi-Armed Bandit (MAB) problem: the MATMAB.
- Demonstrates the utility of graphical causal selection diagrams for agents with finite-sample concerns in MATMAB settings.
- Introduces and compares causally-empowered agent policies on both individual and community metrics of success.
- Provides a roadmap for future study of more complex MAT environments with relaxed assumptions from the those presented in this introductory work.

In support of the above, this paper is thus structured into the following outline of topics:

- Sec. 2 reviews the necessary background from causal inference that supports the tools employed by our agents in MAT settings.
- Sec. 3 states the definitions and assumptions of the current work's MAT formalizations alongside a motivating example.
- Sec. 4 describes the MAT agent and environment constructions that are used in the MATMAB simulations that follow.
- Secs. 5, 6 discuss the results of these simulations in support of the novel agents that exploit local causal structure, and give prescriptions for future directions.

2 PRELIMINARIES

In order to specify the causal assumptions about the relationships between variables in some environment, depict its intuitive graphical interpretation, and to formalize the structural localities of transportable effects, we employ the vocabulary of Structural Causal Models, or SCMs:

Definition 2.1. ***(Structural Causal Model)*** *[20, pp. 204-207] A Structural Causal Model is a 4-tuple, $M = \langle U, V, F, P(u) \rangle$ where:*

1. *U is a set of background variables (also called exogenous), that are determined by factors outside the model.*
2. *V is a set $\{V_1, V_2, ..., V_n\}$ of endogenous variables that are determined by other variables in $U \cup V$.*
3. *F is a set of functions $\{f_1, f_2, ..., f_n\}$ such that each f_i is a mapping from (the respective domains of) $u_i \cup PA_i$ to V_i where $U_i \subseteq U$ and $PA_i \subseteq V \setminus V_i$ and the entire set F forms a mapping from U to V. In other words, each f_i in $v_i = f_i(pa_i, u_i), i = 1, ..., n$ assigns a value to V_i that depends on (the values of) a selected set of variables.*
4. *$P(u)$ is a probability density defined on the domain of U.*

Each SCM details a corresponding *causal diagram*, which is a directed, acyclic graph depicting the model that is constructed with the following properties:

Definition 2.2. ***(Causal Diagram)*** *Given any SCM M, its associated causal diagram G is a directed, acyclic graph (DAG) that encodes:*

1. *The set of endogenous variables V, represented as solid nodes.*
2. *The set of exogenous variables U, represented as hollow nodes (sometimes omitted for brevity).*
3. *A directed edge connects two variables $V_c \rightarrow V_e$ for $V_c, V_e \in V$ if V_c appears as a parameter in $f_{V_e}(V_c, ...)$ (i.e. if V_c has a causal influence on V_e).*
4. *A bidirected, dashed edge connects two variables $V_a \dashleftarrow\dashrightarrow V_b$ if their corresponding exogenous parents U_a, U_b are dependent, or if f_{V_a}, f_{V_b} share an exogenous variable U_i as a parameter to their functions.*

Causal diagrams also serve as formalizations of the independence relationships between variables via the *d*-separation criterion.

Definition 2.3. ***(d-separation)*** *Given a causal diagram G, the independence relationship $X \perp\!\!\!\perp Y | Z$ holds (meaning that "X, Y are directionally separated given Z") for variables X, Y, Z if every path p that connects X, Y is blocked. A path is blocked whenever there exists a triplet of nodes along p such that:*

1. *Chains (mediators): patterned $X \rightarrow Z \rightarrow Y$, meaning that X affects Y through the mediator Z, are blocked when Z is conditioned upon.*
2. *Forks (common causes): patterned $X \leftarrow Z \rightarrow Y$, meaning that Z is a common cause of X and Y, are blocked when Z is conditioned upon.*
3. *Colliders (common effects): patterned $X \rightarrow Z \leftarrow Y$, meaning that Z is a common effect of X and Y, are blocked when neither Z nor any of its descendants are conditioned upon.*

The *d*-separation criterion's recipe for decoding a causal graph's independence claims maps to the causal notion of "explaining away" relationships between variables. This formalization plays an important role in an agent's feature selection, being wary to control for any possible back-door paths that introduce non-causal associations between variables, and to ensure that any causal pathway between some action and outcome is unperturbed [19, 22].

SCMs also disambiguate causal *acts* taken by agents able to perform their own experiments (interventions) from associational *observations* made by other agents whose policies may be unknown.

Definition 2.4. ***(Intervention)*** *An intervention represents an external force that fixes a variable to a constant value (akin to the random assignment of an experiment), and is denoted $do(X = x)$, meaning that X is fixed to the value x. This amounts to replacing the structural equation for the intervened variable with its fixed constant such that $f_X = x$ (eliciting the "mutilated submodel" M_x). This operation is also represented graphically by severing all inbound edges to X in G, resulting in an "interventional subgraph" G_x.*

Causal diagrams serve another purpose for determining what data (either observations or interventions) collected in one environment can be transported to another without introducing bias. *Selection diagrams* are used when the causal graph between environments is the same, but in which there may exist differences in certain structural localities.

Definition 2.5. ***(Selection Diagram)*** *[2] Let $\langle M, M^* \rangle$ be two SCMs relative to environments $\langle \pi, \pi^* \rangle$ sharing a causal diagram G. By introducing selection nodes, boxed variables representing causes of variables that differ between source and target environment, $\langle M, M^* \rangle$ is said to induce a selection diagram $\mathcal{D}$ if $\mathcal{D}$ is constructed as follows:*

1. *Every edge in G is also an edge in $\mathcal{D}$.*
2. *$\mathcal{D}$ contains an extra edge $S_i \rightarrow V_i$ (i.e., between a selection node and some other variable) whenever there might exist a discrepancy $f_i \neq f_i^*$ or $P(U_i) \neq P^*(U_i)$ between M and M^*.*

Traditionally, selection diagrams have been employed in offline data analysis to determine if already-collected datasets can transfer between environments with the assumption that selection nodes have already been encoded in the diagram to license or forbid said transport. In the present endeavor, agents will instead be tasked with learning the locations of these selection nodes over time, licensing or forbidding transport of observations from other agents.

The above preliminaries from causal inference will be useful in structuring the environments in which agents will make choices and learn from one another in the multi-agent setting. To further formalize how agents learn over time, we extend the definition of a traditional sequential reinforcement learning task:

Definition 2.6. ***(Multi-Armed Bandit (MAB) Problem)*** *A Multi-Armed Bandit problem is an online, sequential, decision-making task in which an agent is tasked with maximizing cumulative reward received over time. A MAB problem instance is characterized by:*

- *Trials: some $T \in \mathcal{N}^+$ (possibly infinite) sequential trials at which the agent makes a choice and receives some reward.*
- *Actions: some choices (also known as "arms") $x \in X$ (with $|X| \geq 2$), at each trial $t \in T$.*
- *Rewards: some distribution of rewards Y associated with each choice $x \in X$, and received at each trial $t \in T$. In the simplest, Bernoulli bandit settings, the reward is binary, $Y \in \{0, 1\}$, in which case the optimal action $x^* = \text{argmax}_x P(Y = 1 | do(X))$.*

A key facet of a MAB problem is that the reward distributions $P(Y = 1|do(X))$ are initially unknown to an agent, requiring it to maximize reward by managing a game of "explore vs. exploit," sampling arms to determine their merit until confident that it has found the best to continuously exploit thereafter. In *contextual* variants of MAB problems, the objective is to choose actions that maximize $z-$specific rewards for some pre-treatment contextual state Z, i.e., to choose $x^* = \text{argmax}_x P(Y = 1|do(X = x), Z = z)$.

3 MULTI-AGENT TRANSPORTABILITY

We now extend the preliminaries in the previous section to *multi-agent environments*, beginning with a motivating example in the classic MAB setting of online advertising.

Motivating Example: consider the task of *simultaneously deploying* several advertising agents to platforms with different audiences (e.g., different streaming platforms, websites, etc.) that may non-trivially differ in their traits and responses to the selection of ads available to the agents. Because the agents are newly interacting within these communities or with a new selection of ads, they will need to learn which ads have the maximal clickthrough rates predicated on each viewer's characteristics. Notably, the task is the same for each agent (as may be characterized by an SCM like in Figure 1 in which X is an agent's ad choice, Y is a user's clickthrough, and Z, W are pre- and post-treatment covariates like age and fear-of-missing-out, respectively), though by modeling each community as a separate agent interacting in its own environment, we can govern how (if at all) observations from one community may aid another.

Definition 3.1. ***(Multi-Agent Transport Environment (MAT-E)*** *A Multi-Agent Transport Environment (MAT-E) is a 2-tuple $\mathcal{E} = \langle M, A \rangle$ where:*

- *M is an SCM characterizing the causal mechanics of an agent's choices and received rewards in this environment.*
- *A is a set of agents belonging to and acting within the environment by the dictates of their individual policies, ϕ_i: $A = \{A_0^{\phi_0}, ..., A_a^{\phi_a}\}$.*

Definition 3.2. ***(Multi-Agent Transport World (MAT-W))*** *A Multi-Agent Transport World (MAT-W) $\mathcal{W}$ consists of some set of Multi-Agent Environments composing the entire domain of environments and associated agents in a given setting: $\mathcal{W} = \{\mathcal{E}_0, ..., \mathcal{E}_e\}$.*

Agents in MAT-Es iteratively learn from both their, and other, environments as characterized in a new type of MAB problem:

Definition 3.3. ***(Multi-Agent Transport MAB (MATMAB) Problem)*** *A Multi-Agent Transport MAB problem is a MAB variant in which:*

- *Agents independently make sequential choices in the context of some given MAT-W, $\mathcal{W}$.*
- *An episode e at a given trial t and for agent A_i consists of the state of any environmental covariates Z_t, chosen action X_t, and received reward Y_t and is denoted: $e_t^{A_i} = \{Z_t, X_t, Y_t\}$*
- *At each choice at trial t, agents possess observations of all agents' previous episodes $e_{0:t-1}^{A_i}$ $\forall A_i \in A$.*

MATMAB problems offer both challenges and opportunities for agents attempting to learn the optimal policy as quickly as possible in order to maximize cumulative reward: relying on one's own data (in which source and target environment are guaranteed to be the same) is the safest approach akin to what traditional agents would attempt, but misses the opportunity to incorporate observations from other agents to accelerate discovery. These observations come with their own risks, as they may come from heterogeneous environments that are not naïvely transportable. This challenge is complicated by potential conflicts in the causal tiers of data: each agent is free to choose actions experimentally (generating causal samples of $do(X = x)$ by intervention), but obtains only observations of other agents' episodes (wherein the governing policies of other agents may not be known).

4 METHOD

There are many possible perturbations on MATMAB properties that can lead to performance differences of agents within; the current work makes the following simplifying assumptions for the introduction of this domain, inviting future studies to relax them:

- *Markovian SCMs*, in which MAT-E SCMs possess no unobserved confounders. Furthermore, we herein specify each MAT-E using a Causal Bayesian Network, a type of SCM in which the causal structure is known, but the governing structural equations are not. Environment parameters are thus specified via $P(V|pa(V))$ for all endogenous variables V and their parents $pa(V)$ in the causal graph. These distributions are fixed in the underlying MAT-E $\forall t \in T$.
- *Known and Fixed Causal Structure*, in which each agent knows the underlying causal structure of their MAT-E, though not the parameters nor which are shared between MAT-Es.[1]
- *Perfect Observations*, in which each agent may observe the episodes of other agents with perfect fidelity (though policies of other agents may not be known).
- *Binary Rewards* $Y \in \{0, 1\}$ associated with each trial.

With these assumptions, we construct simulations to test the performance of agents in differently-parameterized MAT-Ws.

4.1 Simulation Constructs

Definition 4.1. ***(MATMAB Simulation)*** *A MATMAB Simulation is parameterized by the following:*

- *$N \in \mathcal{N}^+$ Monte Carlo repetitions over which agent performance is averaged. Each of $n \in N$ repetitions is pairwise independent.*
- *A mapping $n \mapsto \mathcal{W}_n$ of MC repetitions to a particular MAT-W, $\mathcal{W}_n$. A single MAT-W is specified for all T trials.*
- *$T \in \mathcal{N}^+$ sequential decision trials for each $n \in N$ MC repetition. At each trial, all pre-treatment covariates Z_t are presented as input to the agents, who then make a decision X_t, and are then presented with post-treatment covariates W_t and optimization target outcome Y_t.*

[1]The assumption of a known causal structure may appear to be a strong one, but can be assuaged by virtue of each agents' assumed ability to perform experiments in their own domains; if the task of *causal discovery* is added on top of the tasks outlined in this work, the ability to distinguish decision variables under each agents' control (i.e., amenable to intervention $do(X = x)$) from covariates of action and outcome empowers causal discovery algorithms far beyond that of causal discovery from offline observational datasets [16, 30].

Success of agents in MATMAB simulations is characterized at both the individual and community levels due to the fact that the utility of information from observing episodes from other agents may improve inter-agent performance.

Definition 4.2. ***(MATMAB Cumulative Pseudo-Regret (CPR))*** *Cumulative pseudo-regret $CPR(A,t)$ measures the difference between the optimal likelihood of successful outcome $Y_t^*(Z_t) = 1$ (for any contextal covariates Z_t) and reward received by agents $a \in A$ across all trials $1 : t < T$, and is thus defined as:*

$$CPR(A,t) = \sum_{a \in A} \sum_{i=1}^{i=t} \max_x P(Y=1|do(X=x), Z_i = z_i) - Y_i^a$$

4.2 Agent Policies

Effective policies at minimizing regret in MATMAB settings face unique challenges in addition to the traditional explore vs. exploit dilemma implicit in typical MAB problems, including: how to determine which agents are in similar or different environments, where these structural similarities and dissimilarities exist, and how to perform finite-sample data fusion between one's own experiments and observations of others. Agent policies in this context are thus a composite of subpolicies defined as follows:

Definition 4.3. ***(MAT Agent (MAT-A))*** *In the current study, a MAT Agent's policy ϕ is a composite of two subpolicies:*

- *Observational Transport Policy (OTP): determines how agent A_i incorporates observations of other agents' episodes, $e_{0:t-1}^{A_k}$, $i \neq j$, into its own history, if at all.*
- *Action Selection Rule (ASR): determines how the agent maps its history of episodes to an action.*

4.2.1 Action Selection Rules (ASRs): Given the discrete, contextual reward distributions assumed in the present study, and the objective function of minimizing regret in Def. 4.2, we compare performance of agents across four traditional ASRs [24] given the tradeoffs that each leverages in the MAB "explore vs. exploit" dilemma:

- *ϵ-Greedy (EG)*: at each trial, the agent chooses randomly with probability ϵ or greedily based on accumulated samples with probability $1-\epsilon$.
- *ϵ-Decreasing (ED)*: same as EG but with a cooling schedule for the value of ϵ.
- *ϵ-First (EF)*: the agent chooses randomly for the first $T * \epsilon$ trials, then greedily thereafter.
- *Thompson Sampling (TS)*: self-correcting ASR that maximizes contextual reward by sampling the believed reward distribution [26].

4.2.2 Observation Transport Policies (OTPs): We compare four agent OTPs to demonstrate both the risks of naïvely incorporating observed samples into one's own environment as well as the opportunities to exploit structural similarities between even different environments to speed learning.

- *Solo:* (no OTP) ignores sample data from other agents, making choices based on its own experience alone.
- *Naïve:* incorporates *all* episodes from other agents into its own history, ignoring environmental differences.
- *Sensitive:* incorporates *whole* episodes from other agents, but only if an admissibility criteria is met for similarity of environments (detailed later).
- *Adjusting:* incorporates *local/partial* episodes from other agents for each node lacking a selection-node in its learned selection diagram (detailed later).

The first two OTPs (Solo and Naïve) lack causal underpinnings whereas the latter causally-empowered policies (Sensitive and Adjusting) compose this work's unique agents that consider causal transportability of observations into their environments.

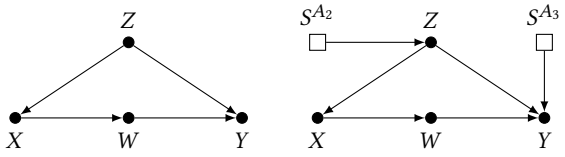

Figure 1: (Left) SCM causal graph G employed in experimental simulations with confounder Z and mediator W. (Right) Example agent selection diagram with square transport nodes added to G.

To demonstrate these causally-empowered OTPs, we refer to the causal graph in Figure 1 (Left) that all MAT-Es in the present experiment employ. This model defines agent-action X, outcome Y, covariates Z, W, and is the simplest structure that, in pursuit of measuring the causal effect of X on Y, possesses one covariate that should be controlled-for (the confounder, Z, whose control homogenizes action effects X on Y), and one that should not be (the mediator, W, whose control blocks the effect of X on Y).[2]

Note that although all MAT-Es in this experiment possess this structure, the distributions across each variable may differ between them. Consider an example MAT-W depicted in Fig. 2 in which there are 3 MAT-Es with 4 agents $A_0, A_1 \in \mathcal{E}_0$, $A_2 \in \mathcal{E}_1$, and $A_3 \in \mathcal{E}_2$. From the perspective of agent A_0, observed episodes from agent A_1 should be directly transportable, but those from A_2 and A_3 may require more careful examination.

However, by the assumptions of the MATMAB problem, no agent begins with knowledge of environmental similarities or differences compared to other agents under observation, and so causally-empowered OTPs must also learn where selection nodes exist between other agents' environments. This is accomplished by defining three components: (1) the MAT selection diagram, (2) a selection node discovery rule, (3) and an observation incorporation rule.

4.2.3 MAT Selection Diagrams: each causally-empowered agent A_i maintains a MAT selection diagram, $\mathcal{D}_\rangle$, that behaves the same as those traditionally defined in Def. 2.5, except that instead of selection nodes pertaining to causal differences between each MAT-E (which are not known a priori), they are added to variables for each other agent A_k with observed differences in local distributions $P(V|\text{pa}(V))$. For instance, if Fig. 1 (Right) represents the MAT selection diagram $\mathcal{D}_\prime$ maintained by A_0, the selection node $S^{A_2} \rightarrow Z$

[2]The simplicity of this model does not compromise the generalizability of this study's results, as the estimation of causal effects from known structures (including which covariates make unbiased controls when an effect is identifiable) is complete by the rules of *do*-calculus [10, 19].

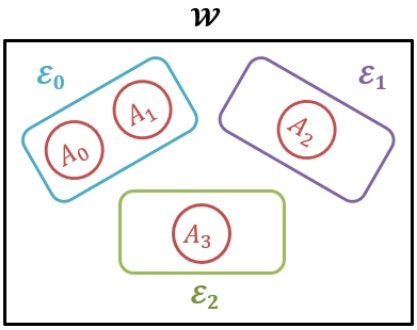

Figure 2: Example MAT-W $\mathcal{W}$ with 3 composite MAT-Es, $\{\mathcal{E}_0, \mathcal{E}_1, \mathcal{E}_2\}$ each with their own contained agents A_i.

indicates that A_0, A_2 differ on the distribution $P(Z)$. Our causally-empowered agents begin MATMAB problems by assuming that they differ from all other agents' environments by placing selection nodes upon every variable for every agent; the selection node discovery rule that follows dictates when, through experience, a selection node would be removed.

4.2.4 Selection Node Discovery Rule: because agents in MATMAB settings are sensitive to the finite-sample biases of limited experience at each trial, they require means of determining where selection nodes belong in $\mathcal{D}$ as more episodes are acquired. The present work defines a simple criteria using Hellinger Distance (H^2), such that for every node's distribution of the format $P^{A_i}(V|\text{pa}(V))$ from the samples of agent A_i, a selection node is placed upon V by agent A_i for agent A_k if $H^2(P^{A_i}(V|\text{pa}(V)), P^{A_k}(V|\text{pa}(V)))$ exceeds some threshold τ.

In the present work, we adopt an empirically well-performing threshold for each variable V of $\tau(V) = 0.02 * 2^{|pa(V)|}$, though future studies are invited to employ more sophisticated methods. With new observations collected by each agent at every trial, the selection diagram can be updated with selection nodes in $O(|A||V|)$ time for $|A|$ agents and $|V|$ variables in the underlying causal graph.

Choice of selection node discriminance may have a powerful influence on causally-empowered agent performance; conservative discovery rules requiring node distributions to be very close means that the defacto selection-nodes will be removed slowly, and thus observations incorporated later in the learning process when they may not be as useful. However, a liberal discovery rule that allows for greater distributional flexibility means that defacto selection-nodes may be removed too quickly, and heterogeneous distributions may be combined to the detriment of the learner. That said, in systems with many observable agents, we hypothesize that combinations of observations from sufficiently similar local structure will accelerate convergence.

4.2.5 Observation Incorporation Rule: with the most up-to-date selection diagram $\mathcal{D}$ available to a causally-empowered agent, they may choose if and how to incorporate observations from other agents into their own histories, which are then employed by their ASRs to make a choice. Algorithm 1 describes the general procedure by which causally-empowered MATMAB agents make decisions; the difference between our two causally-empowered implementations of the Sensitive and Adjusting agents can be found at the OTP step, which dictates how each employs their respective selection diagram and learned selection nodes. Once these nodes have been learned, the OTP behaviors depend on the following criteria:

Definition 4.4. ***(MAT-Ignorability)*** *In a MAT-E with decision variable X, outcome Y, covariates Z, and controlled covariates $C \subseteq Z$, a selection node S^{A_i} pertaining to the environment of agent A_i is said to be "MAT-Ignorable" if it is d-separated from Y given C, viz., if $S \perp\!\!\!\perp Y \mid C$.*

MAT-Ignorable selection nodes thus represent environmental differences that do not affect the maximization target in MATMAB settings. In the selection diagram of Figure 1 (Right), S^{A_2} would be MAT-Ignorable because Z (a confounder) is controlled, making it d-separated from the outcome Y, whereas S^{A_3} is not. How the causally-empowered OTPs treat MAT-Ignorable selection nodes affects their ability to incorporate external observations by the following distinction between global and local transportability.

Definition 4.5. ***(MAT-Global vs. Local Transport)*** *In a MAT-E with decision variable X, outcome Y, covariates Z, and controlled covariates $C \subseteq Z$:*

- *From the perspective of agent A_j, samples over $P^{A_k}(X, Y, Z)$ from another agent A_k are said to be globally-transportable only if all selection nodes S^{A_k} in $\mathcal{D}^{A_j}$ are MAT-ignorable.*
- *From the perspective of agent A_j, samples from another agent A_k over local distributions $P^{A_k}(V|pa(V))$ are said to be locally-transportable if there is no direct selection node $S_{A_k} \rightarrow V$.*

In other words, globally-transportable environments are either precisely the same or differ only for variables whose selection mechanics are rendered independent from the outcome. Samples from these environments are thus safe to incorporate across the full joint distribution, but can be difficult to find if even a single selection node is not rendered MAT-Ignorable. More practically, locally-transportable samples allow an agent to substitute its own experiences and those from environments that it believes to be similar in place of those it believes to be different. As such, we now more precisely define our causally-empowered OTPs in terms of the above and our toy model in Figure 1 (Left):

- *Sensitive* OTP agents incorporate observations across *all* variables, sampled from $P^{A_k}(X, Z, W, Y)$ for another agent A_k only if A_k is amenable to global transport, i.e. *all* selection nodes from A_k are MAT-Ignorable.
- *Adjusting* OTP agents incorporate *local* observations at conditional distributions $P(V|pa(V))$ from another agent A_k if $P^{A_k}(V|pa(V))$ is amenable to local transport. For any nodes that are *not locally-transportable* this OTP will instead substitute samples from its home distribution and any other locally-transportable environment for V in order to obtain a more accurate estimate. Samples from environments without any locally-transportable nodes are ignored entirely.

Again referring to the selection diagram in Figure 1 (Right) and MAT-W in Figure 2, a *Sensitive* OTP A_0 would incorporate global observations from agents A_1, A_2 but none from A_3. An *Adjusting* OTP A_0 would likewise incorporate all observations from A_1, A_2

Algorithm 1 Pseudocode for the Sensitive and Adjusting agents' choices at each trial $t \in T$, parameterized by its current MAT-E, $\mathcal{E}$. Each of the two causal OTPs will approach line 3 differently.

1: **procedure** CAUSAL_CHOOSE($\mathcal{E}$)
2: $d_t \leftarrow update\_\mathcal{D}(h_t)$ ▷ Update selection diagram
3: $h_t \leftarrow update\_H(h_t, d_t)$ ▷ OTP updates history
4: $z_t \leftarrow \mathcal{E}.f_Z(u_t)$ ▷ Covariates observed
5: $x_t \leftarrow \phi(h_t, z_t)$ ▷ ASR selects arm
6: $y_t \leftarrow \mathcal{E}.f_Y(z_t, x_t)$ ▷ Observe reward
7: $h_{t+1} \leftarrow \mathcal{E}.observe\_e_t^A()$ ▷ Update history

into each $P^{A_0}(V|pa(V))$ but would only incorporate $P^{A_3}(W|X)$ from agent A_3 (and not $P^{A_3}(Y|W,Z)$ due to the presence of a selection node). This incorporation of observations into structural localities empowers Adjusting agents to scavenge transportable information from pieces of the SCM, even if the observations from other agents are not entirely transportable. Local transportability can then exploit enhancements to accurate estimates of CPT factors of the maximization target, e.g., for the model in Figure 1 (Left):

$$P(Y=1|do(X),Z) \tag{1}$$
$$= \sum_w P(Y=1|do(X),Z,W=w)P(W=w|do(X),Z) \tag{2}$$
$$= \sum_w P(Y=1|Z,W=w)P(W=w|X) \tag{3}$$

Eqn. 1 follows from the optimization target set forth in Definition 4.2 for CPR, Eqn. 2 from the law of total probability, and 3 from the rules of d-separation ($\{X \perp\!\!\!\perp Y|W,Z\}, \{Z \perp\!\!\!\perp W|X\}$) and do-calculus ($P(W|X) = P(W|do(X))$). Consider now that an *adjusting* agent A_0 in Figure 2 can obtain a more accurate estimate of its optimization target from observations over A_1, A_2, A_3 through local-transportability, excepting the only not-locally-transportable observations from $P^{A_3}(Y|Z,W)$:

$$P^{A_0}(Y=1|do(X),Z) \tag{4}$$
$$= \sum_w P^{A_{0,1,2}}(Y=1|Z,W=w)P^{A_{0,1,2,3}}(W=w|X) \tag{5}$$

4.3 Experimental Conditions

4.3.1 Environment Construction: In order to fairly compare the performance of different combinations of OTPs and ASRs in MATMAB settings without cherry-picking favorable selection settings or reward parameterizations, all simulations were conducted in randomized MAT-Ws. The procedure for creating randomized MAT-Ws is detailed in Algorithm 2. Each experiment that follows consists of $T = 3000$ trials repeated and averaged over $N = 1600$ Monte Carlo (MC) repetitions, each taking place in a randomized MAT-W, $\mathcal{W}_n$.

4.3.2 Experiment 1 - Individual Comparison: The first experiment measures the success of the various OTPs (Solo, Naïve, Sensitive, and Adjusting) from the perspective of communities of homogenous ASRs operating in a MATMAB setting. We hypothesized that the Sensitive and Adjusting agents within each ASR-standardized community would encounter significantly less CPR compared to the Solo and Naïve, with the Naïve agents potentially failing to achieve sub-linear regret due to destructive incorporation of observations that did not transfer to their environment. Each condition in this experiment consisted of a MAT-W of 4 agents sharing the same ASR but with different OTPs.

Algorithm 2 Pseudocode for creating a randomized MAT-W given some likelihood for node mutation ϵ_m and some set of agents A; returns the set $\mathcal{E}^A$ of environments assigned to each agent $a \in A$.

1: **procedure** RANDOM_MAT_W_INIT(ϵ_m, A)
2: $M \leftarrow SCM(structure)$ ▷ Init default SCM
3: **for** $V \in M$ **do**
4: $M.P(V|pa(V)) \leftarrow randomize\_cpt()$
5: $\mathcal{E}^A \leftarrow \{M.clone() \; \forall \; a \in A\}$ ▷ $\mathcal{E}^a$ starts with default
6: **for** $\mathcal{E}^a \in \mathcal{E}^A$ **do**
7: **for** $V \in \mathcal{E}^a$ **do**
8: $s_V \sim Bern(\epsilon_m)$ ▷ Flip weighted coin for S node
9: **if** $s_V == 0$ **then** ▷ V^a to differ from default
10: $\mathcal{E}^a.P(V|pa(V)) \leftarrow randomize\_cpt()$
11: **return** $\mathcal{E}^A$

4.3.3 Experiment 2 - Community ASR Comparison: The second experiment measures the utility of heterogeneous communities of ASRs sharing the Adjusting OTP. The impetus for this experiment was a question of the degree to which causally-empowered agents rely on more explorative ones and may gain more from exploration than merely samples of arm-quality from its home environment; early exploration may earlier repeal S-nodes that then enable earlier incorporation of observations from peers in the learning process. Thus, we hypothesized that adjusting communities with ASRs prioritizing earlier exploration may outperform more "selfish," though typically individually superior, ASRs like TS. To test this hypothesis, we compared 4 different communities in MAT-Ws consisting of 4 agents each with variant blends of front-heavy exploration: $C_0 = \{TS, TS, TS, TS\}, C_1 = \{EF, EF, EF, EF\}, C_2 = \{TS, TS, EF, EF\}, C3 = \{EG, ED, EF, TS\}$. C_0, C_1 represent the homogeneous communities whose average CPR we compare to the heterogeneous C_2, C_3.

Although not discussed herein[3], the ϵ-Greedy, ϵ-First, and ϵ-Decreasing ASRs were fine-tuned with the highest-performing variants included for comparison in each experiment. In particular, the ϵ-Greedy agent set exploration rate $\epsilon_G = 1/30$, ϵ-First explored for the first $\epsilon_F * T$ trials of each context $Z = z$ with $\epsilon_F = 1\frac{2}{3}$, and ϵ-Decreasing used an exponential cooling schedule $\epsilon_{D,t} = 0.98^t$ for each context $Z = z$.

4.3.4 Experimental Metrics of Success: To compare the efficacy of agents employing different OTPs and ASRs, we examined performance on two traditional metrics of MAB agent success:

(1) *Probability of Optimal Action (POA):* the likelihood that the agent selected the contextually optimal action at trial t averaged across all N Monte Carlo repetitions.
(2) *MATMAB Cumulative Pseudo-Regret (CPR):* Def. 4.2 applied to the different agent populations described in the previous section (individually in Experiment 1 and across a community

[3] See https://github.com/axelbrowne/ECS4TOMAE for simulation code.

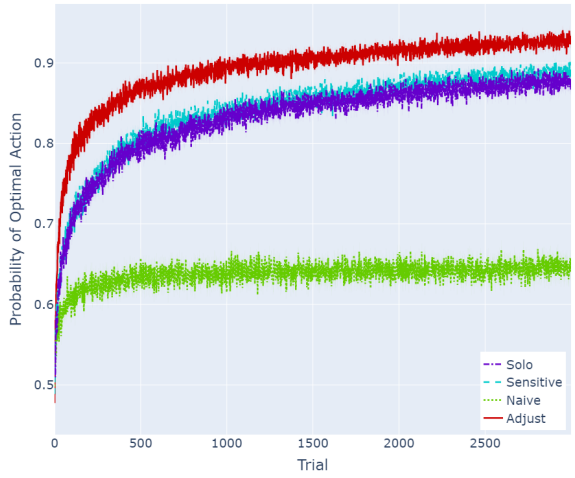

Figure 3: POA of OTPs for the TS ASR with $0.2 < \epsilon_m < 0.8$ in Experiment 1.

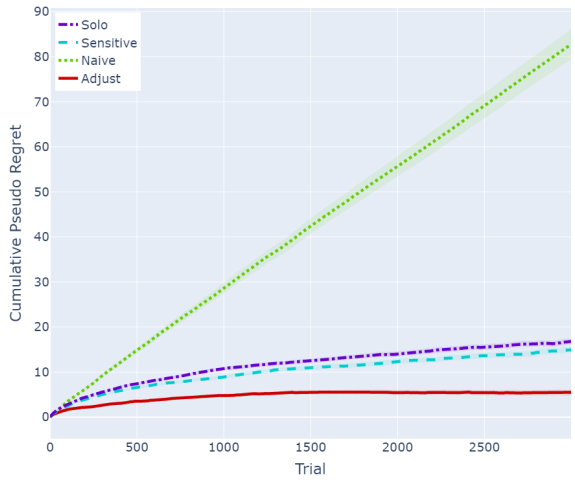

Figure 4: CPR of OTPs for the TS ASR with $0.2 < \epsilon_m < 0.8$ in Experiment 1.

of agents in Experiment 2), the CPR encountered by agents at trial t averaged across all N Monte Carlo repetitions.

5 RESULTS

Experiment 1 - Individual OTP Comparison: Table 1 summarizes the performance of all compared combinations of OTPs and ASRs in Experiment 1, with Figures 3 and 4 highlighting differences in TS performance across OTPs ($0.2 < \epsilon_m < 0.8$). For all ASRs, the causally-empowered OTPs yielded significantly less CPR than the Solo OTP, with Adjusting agents shining far above the others. While all Adjusting ASRs obtained sublinear CPR within T trials, other OTPs did not.

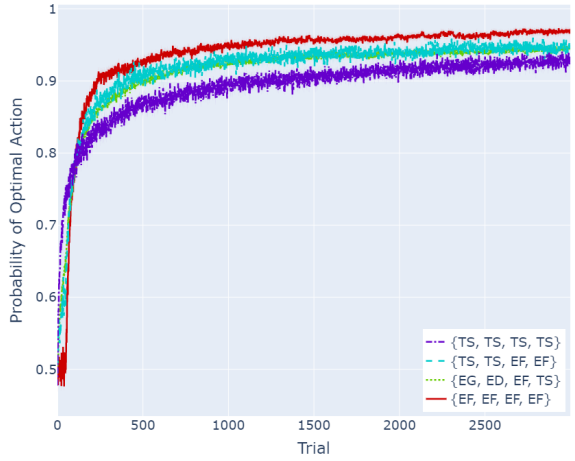

Figure 5: POA of Adjusting OTPs for the given ASR communities with $0.2 < \epsilon_m < 0.8$ in Experiment 2.

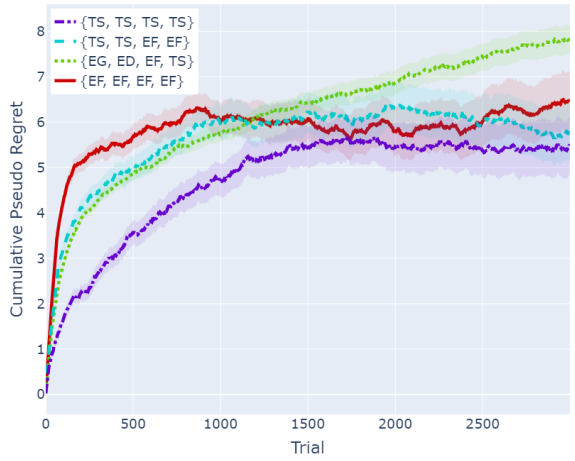

Figure 6: CPR of Adjusting OTPs for the given ASR communities with $0.2 < \epsilon_m < 0.8$ in Experiment 2.

Experiment 2 - Community ASR Comparison: Tables 2 and 3 summarizes the performance of the Adjusting-OTP-standardized communities alongside visualizations of their differing-ASR performance in Figures 5 and 6. Lower ϵ_m predicated improved performance for adjusting agents overall, though the Adjusting-TS ASR communities were least resilient to higher ϵ_m. The heterogeneous communities, $C2, C3$, failed to achieve less CPR than their homogeneous counterparts, C_0, C_1.

6 DISCUSSION

The experimental results demonstrate that MATMAB problems introduce new dimensions to traditional MAB formalizations: the opportunity to harness transportable observations from other agents can be incorporated to the benefit of agents as individuals, and that MATMAB communities benefit from larger amounts of forward exploration.

Table 1: Individual $CPR(A_i, T)$ for all ASRs in Experiment 1 with given standard-errors. Each row represents the 4-agent community in each of the 4 ASR's experiments. Bolded values are the lowest CPR in each row's community, underlined values are the lowest CPR in each column's OTP, and asterisks (*) indicate significant improvement from the Solo OTP. All simulations in this condition were run with $0.2 < \epsilon_m < 0.8$. Column means M provided at bottom row.

	Naïve	Solo	Sensitive	Adjusting
EG	82.68 ± 3.42	33.92 ± 1.30	33.90 ± 1.33	**10.02 ± 0.65***
EF	107.8 ± 3.03	23.16 ± 1.34	18.04 ± 1.27*	**6.450 ± 0.65***
ED	118.0 ± 2.86	21.82 ± 1.10	18.10 ± 1.03*	**7.761 ± 0.66***
TS	82.73 ± 3.31	16.81 ± 0.67	14.94 ± 0.66*	**5.510 ± 0.63***
M	97.80 ± 3.16	23.93 ± 1.10	21.25 ± 1.07	7.435 ± 0.65

Table 2: Average $CPR(A, T)$ for communities with Adjusting OTPs and standardized ASRs in Experiment 1 and the given standard-errors. Bolded values are the lowest CPR in each row's ASR and underlined values are the lowest CPR in each column's parameterization for the node mutation rate, ϵ_m. Column means M provided at bottom row.

Agent ASR	$\epsilon_m = 0.2$	$\epsilon_m = 0.8$	$0.2 < \epsilon_m < 0.8$
EG	**7.828 ± 0.65**	9.139 ± 0.66	9.667 ± 0.66
EF	8.058 ± 0.62	6.659 ± 0.63	**5.217 ± 0.62**
ED	**6.029 ± 0.64**	7.728 ± 0.65	8.476 ± 0.64
TS	**8.053 ± 0.62**	13.34 ± 0.63	8.728 ± 0.61
M	7.492 ± 0.63	9.217 ± 0.65	8.022 ± 0.63

Table 3: Average $CPR(A, T)$ for communities of 4 Adjusting ASRs $A = \{A_0, A_1, A_2, A_3\}$ in Experiment 2 with given standard-errors. Underlined values are the lowest community CPR.

Community Index	*Community ASRs*	$0.2 < \epsilon_m < 0.8$
C_0	$\{TS, TS, TS, TS\}$	8.728 ± 0.61
C_1	$\{EF, EF, EF, EF\}$	5.217 ± 0.62
C_2	$\{TS, TS, EF, EF\}$	5.787 ± 0.62
C_3	$\{EG, ED, EF, TS\}$	7.808 ± 0.32

Experiment 1 - Individual OTP Comparison: The Naïve agent's poor performance clearly demonstrates the risks of incorporating observations from heterogeneous environments without a causal premise to guide its selection. The Sensitive agent's (albeit modest) improvement over the Solo baseline's demonstrates a conservative approach to transportability that is difficult and slow, especially when environmental differences are high, like in the $\epsilon_m > 0.2$ conditions. That said, the Adjusting agent's exploits of locally-transportable structures in the selection diagram successfully allow it to scavenge pieces of the puzzle even when the causal effect of action on outcome is not directly transportable. This ability translates to more accurate estimation of its home environment's causal effects across ASRs, bringing clarity of optimal action earlier in the learning process.

Both causally-empowered agents appear resilient to diversities of MAT-Ws, as the results are based on many repetitions across randomized environments, and have performance improvements proportionate to environmental similarity in the given MAT-W.

Experiment 2 - Community ASR Comparison: Shifting the focus to the performance of heterogeneous vs. homogeneous communities of agents countered our hypothesized beneficial interaction between early exploring vs exploiting ASRs, as the incorporation of diverse ASRs in a single community were instead weighed down by the worse performers amongst the clique. However, our hypothesis that front-heavy exploration would yield earlier selection node discovery and translate to better performance was confirmed by C_1 obtaining the lowest CPR of any OTP-ASR community in the study. Most interesting is the performance of TS agents that appear to best benefit from communities with other ASRs. Thus, further study could be devoted to ASRs that work *as* a community rather than as individual actors to maximize community reward rather than individual reward that is averaged over the community.

Limitations. Although appropriate for modeling the causal assumptions governing the variables in a MAT-E, SCMs have issues of scalability for high-dimensional covariates, especially given the computational cost of learning selection node placement, and exacerbated for large communities of agents. Other limitations are those of scope for this introductory work of online MAT domains, including the sophistication of the selection node discovery procedure for causally-empowered agents (which may be simply enhanced to a time-delimited choice for τ); there may be even greater opportunity for sophistication in community settings wherein a new type of ASR may make exploratory choices as a function of the value of information to the community.

Future Directions. Perhaps the most meaningful fruit from this work is the number of adjacent avenues for investigation that it spawns, including (but not limited to) MATMAB variants involving: semi-Markovian MAT-Es with observations that may be affected by unobserved confounders, MAT-Es with initially unknown causal structure that must be discovered, missing data in observations (especially when not missing at random [17], as might be a premise for a game-theoretic variant wherein agents act adversarially by hiding select episodes), and scaling to more complex models beyond SCMs wherein similar structural localities can be exploited.

7 CONCLUSION

This work endeavored to employ the tools of causal graphical models within the novel, online MATMAB setting and demonstrated that causally-empowered MAT-Agents excel at learning from their peers compared to those who are not. Through simulation support, we demonstrated that success in MATMAB settings are functions of deciding agents' ASRs, OTPs, community compositions, and similarity of environments. The MATMAB formalization may serve as an important launchpad into future study for how agents learn from their own increasingly rich environments as well as those of other actors.

REFERENCES

[1] DE Asher, SL Barton, E Zaroukian, and NR Waytowich. 2019. Effect of cooperative team size on coordination in adaptive multi-agent systems. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, Vol. 11006. International Society for Optics and Photonics, 110060Z.

[2] Elias Bareinboim and Judea Pearl. 2012. Transportability of causal effects: Completeness results. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 26.

[3] Elias Bareinboim and Judea Pearl. 2013. Causal Transportability with Limited Experiments. In *Proceedings of the Twenty-Seventh National Conference on Artificial Intelligence (AAAI 2013)*, M. desJardins and M. Littman (Eds.). AAAI Press, Menlo Park, CA, 95–101.

[4] Elias Bareinboim and Judea Pearl. 2014. Transportability from multiple environments with limited experiments: Completeness results. *Advances in neural information processing systems* 27 (2014), 280–288.

[5] Olivier Bousquet and André Elisseeff. 2002. Stability and generalization. *The Journal of Machine Learning Research* 2 (2002), 499–526.

[6] Glenn H Bracht and Gene V Glass. 1968. The external validity of experiments. *American educational research journal* 5, 4 (1968), 437–474.

[7] Robert L Brennan. 1992. Generalizability theory. *Educational Measurement: Issues and Practice* 11, 4 (1992), 27–34.

[8] Bobby J Calder, Lynn W Phillips, and Alice M Tybout. 1982. The concept of external validity. *Journal of consumer research* 9, 3 (1982), 240–244.

[9] Yeounoh Chung, Peter J Haas, Eli Upfal, and Tim Kraska. 2018. Unknown examples & machine learning model generalization. *arXiv preprint arXiv:1808.08294* (2018).

[10] Carlos Cinelli, Andrew Forney, and Judea Pearl. 2020. A crash course in good and bad controls. *Available at SSRN* 3689437 (2020).

[11] Andrew Forney and Elias Bareinboim. 2019. Counterfactual Randomization: Rescuing Experimental Studies from Obscured Confounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 2454–2461.

[12] Andrew Forney, Judea Pearl, and Elias Bareinboim. 2017. Counterfactual Data-Fusion for Online Reinforcement Learners. In *International Conference on Machine Learning*. 1156–1164.

[13] Kenji Kawaguchi, Leslie Pack Kaelbling, and Yoshua Bengio. 2017. Generalization in deep learning. *arXiv preprint arXiv:1710.05468* (2017).

[14] Lawrence Leung. 2015. Validity, reliability, and generalizability in qualitative research. *Journal of family medicine and primary care* 4, 3 (2015), 324.

[15] Lucy Ellen Lwakatare, Aiswarya Raj, Ivica Crnkovic, Jan Bosch, and Helena Holmström Olsson. 2020. Large-scale machine learning systems in real-world industrial settings: A review of challenges and solutions. *Information and Software Technology* 127 (2020), 106368.

[16] Daniel Malinsky and David Danks. 2018. Causal discovery algorithms: A practical guide. *Philosophy Compass* 13, 1 (2018), e12470.

[17] Karthika Mohan and Judea Pearl. 2021. Graphical models for processing missing data. *J. Amer. Statist. Assoc.* (2021), 1–16.

[18] Andrei Paleyes, Raoul-Gabriel Urma, and Neil D Lawrence. 2020. Challenges in deploying machine learning: a survey of case studies. *arXiv preprint arXiv:2011.09926* (2020).

[19] Judea Pearl. 1995. Causal diagrams for empirical research. *Biometrika* 82, 4 (1995), 669–688.

[20] J. Pearl. 2009. *Causality: Models, Reasoning, and Inference* (second ed.). Cambridge University Press, New York.

[21] Bernhard Schölkopf. 2019. Causality for machine learning. *arXiv preprint arXiv:1911.10500* (2019).

[22] Ilya Shpitser, Tyler VanderWeele, and James M Robins. 2012. On the validity of covariate adjustment for estimating causal effects. *arXiv preprint arXiv:1203.3515* (2012).

[23] Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. 2018. Learning predictive models that transport. *arXiv preprint arXiv:1812.04597* (2018).

[24] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

[25] Victor Talpaert, Ibrahim Sobh, B Ravi Kiran, Patrick Mannion, Senthil Yogamani, Ahmad El-Sallab, and Patrick Perez. 2019. Exploring applications of deep reinforcement learning for real-world autonomous driving systems. *arXiv preprint arXiv:1901.01536* (2019).

[26] William R Thompson. 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25, 3/4 (1933), 285–294.

[27] Lisa Torrey and Jude Shavlik. 2010. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. IGI global, 242–264.

[28] Guillermo Vigueras and Juan A Botia. 2007. Tracking causality by visualization of multi-agent interactions using causality graphs. In *International Workshop on Programming Multi-Agent Systems*. Springer, 190–204.

[29] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. 2016. A survey of transfer learning. *Journal of Big data* 3, 1 (2016), 1–40.

[30] Kui Yu, Xindong Wu, and Hao Wang. 2010. Online causal discovery. In *9th IEEE International Conference on Cognitive Informatics (ICCI'10)*. IEEE, 667–671.