Optimal and Stable Decision Strategies in Multi-Agent Problems with Generalized Multi-Arm Bandits

Arvind Raghavan, Elias Bareinboim
Department of Computer Science, Columbia University
{ar4284, eb3304}@columbia.edu

Abstract

It is understood that standard contextual Bandit frameworks don't easily encapsulate scenarios with causal confounding and multi-agent interactions. We introduce a *generalized Multi-Arm Bandit* (g-MAB) formalism that covers a richer class of sequential decision problems, thereby extending the sequential decision theoretic literature from a causal inference perspective. We introduce a taxonomy of g-MABs based on novel graphical criteria, and prove some foundational theoretical results, including the limits of counterfactual optimization in such problems and the conditions which permit the discovery of stationary and optimal solutions. We illustrate the performance of common Bandit algorithms on a set of Prisoner's Dilemma experiments and introduce a policy-search algorithm that empirically outperforms the others.

1 Introduction

The question of how a sequential decision-making agent ought to act has been extensively studied, with reinforcement learning (RL) generally being the the default framework used. However, interactions between multiple agents have proved more complicated, and require approaches from multi-agent RL or traditional game theory. By contrast, decision theory studies the problem in terms of a single agent's decision rules, preferences and utility functions (Von Neumann & Morgenstern 1947; Savage 1954; Jeffrey 1983). Decision theory is fast gaining relevance in the design of robust AI agents (Russell & Norvig 2010 Ch. 2, 16; Pearl 2009 Ch. 4). Decision theoretic frameworks also lend themselves naturally to agent-vs-environment RL algorithms, and even supervised learning (Perdomo et al 2020).

Here, one encounters several issues. First, RL and Bandit frameworks often cannot represent even simple multi-agent problems. Consider a popular application of Autonomous Vehicles (AV) learning to interact. A Markov Decision Process (MDP) cannot efficiently encode the dynamics of each AV inferring the other's general strategy, say by observing recent actions, and adapting its behaviour to cooperate or threaten to punish disruption (Cooper et al 2019). Second, what decision theory a rational agent should follow remains a subject of much debate. The main contenders are *evidential decision theory* (Ahmed 2014) and *causal decision theory* (Joyce 1999), with adversarial vulnerabilities claimed for both (Conitzer 2015; Oesterheld & Conitzer 2021). Third, these concerns gain poignancy in the context of AI which introduces nuances and new challenges regarding agent self-identity, memory and preferences (Conitzer 2019).

Causal inference is a conspicuously under-explored perspective in multi-agent decisions, as noted by (Perdomo et al 2020). A line of work (e.g. Miller et al 2020) studies *strategic prediction* using causal models, but are more interested in understanding rational behaviour in non-stationary supervised learning. The theory of Structural Causal Models (SCM) (Pearl 2009), and the concomitant agent tasks of *seeing*, *doing* and *imagining* (Sec. 2), offer powerful tools and a vocabulary to analyse optimal interactive strategies. This growing field of Causal RL (forthcoming) tackles several challenges, such as using causal models to improve online and offline learning, deciding where and whether to intervene, using interventions to learn causal structure efficiently etc. Here, we specifically address multi-agent interactions.

SCMs offer more degrees of freedom in interactions between agents. We motivate our discussion with the Greedy Casino problem described in (Bareinboim et al 2015), in which ordinary *interventional* optimization fails the agent. Adding more arrows to the associated causal graph in that problem essentially generalizes the Bandit framework, as discussed in Sec. 3. Once generalized, we can then investigate some foundational questions: what decision theory is universally optimal? Is there a limit to counterfactual Bandit randomization procedures? What conditions permit a stationary optimal policy, and permit a Bandit algorithm to discover it? In answering these, our specific contributions are:

- (Sec. 4-6) Introducing a Bandit framework that generalizes agent-environment interactions; establishing a clear hierarchy among decision theories in scenarios that can be represented as a generalized Bandit problem, and conditions in which decision theories are equivalent.
- (Sec. 6) Proving that complex counterfactual optimization is unrealizable under reasonable conditions.
- (Sec 5, 7) Identifying novel graphical criteria which determine when the optimal strategy in a Bandit problem is provably stationary and open-to-discovery by common Bandit algorithms.
- (Sec 7-8) Proposing a new algorithm for Bernoulli Bandits and empirically demonstrating it outperforms com-

mon algorithms in toy Prisoner's Dilemma problems.

2 Preliminaries

Notation. We use capital letters for random variables (W), and small letters for their values (w). Variables are discrete unless stated otherwise. Bolded letters represent sets of random variables or their samples ($\mathbf{W} = \{W_1, ..., W_n\}$). We write $P(w)$ as shorthand for $P(W = w)$. $|W|$ represents the cardinality of the variable's domain and $\triangle_W$, the set of distributions over the domain.

Structural Causal Models (SCM). An SCM, M is a tuple $\langle \mathbf{U}, \mathbf{V}, \mathbf{F}, P(\mathbf{u}) \rangle$ (Pearl 2009). $\mathbf{U}$ and $\mathbf{V}$ are the sets of exogenous and endogenous variables, respectively. $\mathbf{F}$ is a set of structural functions by which endogenous variables are realized. For $f_V \in \mathbf{F}, V \leftarrow f_V(pa_V, u_V)$, where $Pa_V \subseteq \mathbf{V}, U_V \subseteq \mathbf{U}$. Once values of $\mathbf{U}$ are sampled according to $P(\mathbf{u})$, all endogenous variables are realized. M_w and M_{σ_W} refer to the sub-model of M where the value of W has been fixed by an atomic intervention $W \leftarrow w$ or by a soft-intervention $W \leftarrow \sigma_W(\mathbf{Z})$ (Korb et al 2004), respectively.

Causal Graphs. Each SCM M induces a directed (acyclic) graph G. $G_{\overline{W}\underline{X}}$ refers to a sub-graph of G, with incoming arrows to W and outgoing arrows from X removed. G_{σ_W} refers to the sub-graph with additional arrows according to a soft-intervention on W, as a function of other non-descendants. $(\mathbf{X} \perp\!\!\!\perp \mathbf{Z}|\mathbf{W})_G$ means $\mathbf{X}$ and $\mathbf{Z}$ are *d-separated* by $\mathbf{W}$ in G. Given our setting, we use filled-grey nodes to indicate variables that are *available* to the agent at decision-making time. Empty-white nodes represent observable variables which are *hidden* from the agent until after a policy decision is made. Dotted nodes represent exogenous variables, which are typically ommitted unless they are confounders.

Pearl Causal Hierarchy (PCH). (Bareinboim et al 2022; Defn. 2,5,7,8) formulate the 3-layer PCH. *Layer-1* (L1) describes the *observational* distribution of variables $P(\mathbf{v})$ when data is collected in the absence of any interventions. *Layer-2* (L2) expresses the *interventional* quantity $P(\mathbf{z}|do(\mathbf{x}), \mathbf{w})$ for different $\mathbf{z}, \mathbf{x}, \mathbf{w}$ in the system. *Layer-3* (L3) determines *counterfactual* probabilities such as $P(y_x|x', y')$, which is shorthand for $\langle (Y = y)$ had X been set to x, given X was in fact x' and Y was $y' \rangle$. The Causal Hierarchy Theorem (ibid.) states that these three layers are distinct in most SCMs. In RL, L1 describes the agent's "default" or "autopilot" regime, L2 the outcome of "deliberate" agent choices, and L3 the outcomes of hypothetical policies "counter-to-the-fact" of actual policy choice (Forney et al 2017; Pearl & Mackenzie 2018).

3 Greedy Casino Seeks Revenge

(Bareinboim et al 2015) introduce a take on a classic decision problem featuring a Greedy Casino in Las Vegas that employs an army of sellout data scientists to surveil its patrons as they choose between 2 slot machines (the Bandit "arms"). They discover that patron behavior can be predicted by two variables, $D, B \in \{0, 1\}$, which are indicators respectively for whether the patron is drunk and whether the machines are both blinking. If $X \in \{0, 1\}$ represents the patron choice between the 2 machines, they found that X was determined in the SCM as $X \leftarrow f_X(D, B) = D \oplus B$.

Assume $B, D \sim Ber(0.5)$. With this knowledge, they install powerful cameras on their machines and dynamically adjust payout based on whether the machines are blinking and whether the patron is drunk, as shown in Table 1. The wiliness of this scheme is that while the average patron receives $E[Y|X = 0] = E[Y|X = 1] = 0.15$, they don't fall foul of the Nevada law that average payout must exceed 30%. If a gaming commissioner were to subject the machines to a randomized trial, they would indeed find $E[Y|do(X = 1)] = E[Y|do(X = 0)] = 0.30$. The solution here is for Bandit agents to use the *Regret Decision Criterion* (RDC) (ibid; Forney et al 2017), a counterfactual randomization procedure where the agent "pauses" before acting and uses this *intended* action to inform a more deliberate decision. We expound and expand on this in Sec. 6.

	$D = 0$		$D = 1$	
	$B = 0$	$B = 1$	$B = 0$	$B = 1$
$X = 0$	**0.10**	0.50	0.40	**0.20**
$X = 1$	0.50	**0.10**	**0.20**	0.40

Table 1: Payout rates for arms in the Greedy Casino problem; agents' "autopilot" choice under some D, B is in bold.

The graph of the original problem was as in Fig. 1c. Unfortunately, the Greedy Casino eventually wises up to the counterfactual savvy of agents and decides to invest aggressively in its technology. It considers monitoring individual patron histories to update payouts more frequently, as shown in Fig. 1d. For AI agents, it even wants to access the agent's policy directly by reading any open-source code (Tennenholtz 2004), or by running powerfully accurate simulations of agents (Kuhn 2019, Sec. 7), as shown in Fig. 1e.

In short, the Greedy Casino wants to add to its interaction with agents more degrees of freedom which cannot be represented in a standard Bandit problem, impelling us to generalize the canonical framework.

4 Generalizing Decision Problems

From a causal perspective, a regular Multi-Arm Bandit (MAB) can be represented by an SCM where an agent chooses a policy at time t, an action is sampled from this policy, and a reward is drawn from an arm-specific distribution, as shown in Fig. 1a. Contextual MABs allow the reward (and therefore optimal policy) to vary with a context variable, as in Fig. 1b. (Forney et al 2017) generalize this to MAB problems where L1 actions are influenced by unobserved confounders (MAB-UC) as shown in Fig. 1c, requiring L3-informed randomization procedures to discover optimal policies. We further generalize this to a richer class of MABs as defined below.

Definition 1: Generalized Multi-Arm Bandit (g-MAB). A g-MAB is defined as a SCM M where, for each time-step $t \in [T]$, we have the following components:

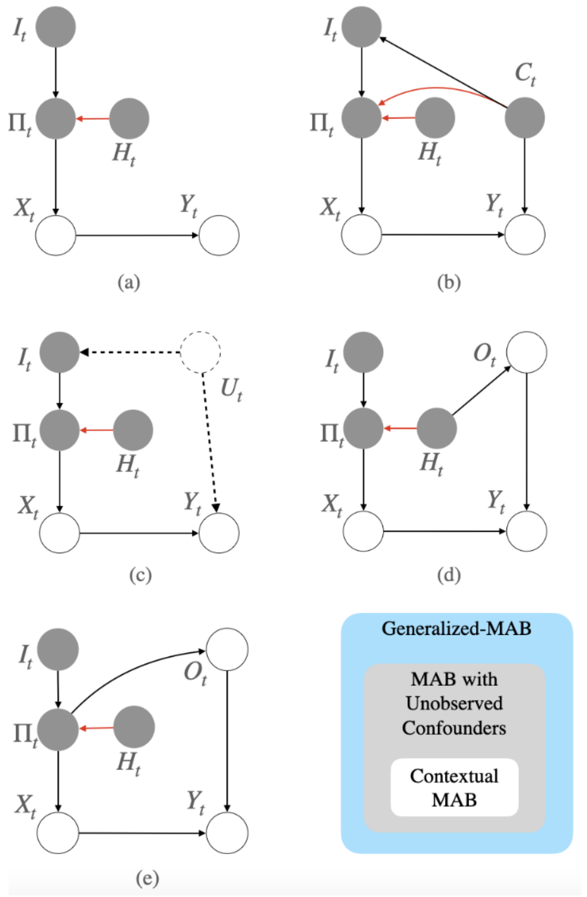
1. *Action-space*: A set of action "arms" the agent can choose from $\{x_1, ..., x_k\}$. Δ_X is the set of all distribution over actions.
2. *Intent*: $I_t \in \Delta_X$ represents the policy the agent *would have chosen* in an L1 regime, without any deliberation.

   I_t is introspectively *available* to the agent before making a policy-decision, an assumption common in cognitive science. This doesn't require knowledge of all the factors influencing one's choice (Tversky & Kahneman 1978).
3. *Policy*: $\Pi_t \in \Delta_X$ represents the actual policy chosen by the agent, from which the action is sampled; obviously, π_t is *available* at decision-making time.

   In the L1 regime of M, $\Pi_t \leftarrow I_t$ (the agent follows its autopilot behaviour)[1]. In the L2/L3 regime, $\Pi_t \leftarrow \sigma_\Pi(I_t, C_t, H_t)$, where agent chooses its policy by soft-intervention using intent, context and history (defined below). Fig. 1 shows black arrows already present in M, and red arrows added in M_{σ_Π}.
4. *Action*: $X_t \in \{x_1, x_2...x_k\}$ represents the arm played by the agent, where $X_t \sim \pi_t$; note, X_t is in general *hidden* from the agent until after the policy is chosen (stochastic sample), unless the policy choice is deterministic.
5. *Context*: C_t is any variable that influences reward, and is *available* to the agent prior to decision-making; context is allowed to influence intent and environment response (defined below), and could be $\emptyset$.
6. *Environment Response*: O_t is any observable variable that is *hidden* from the agent until after a decision is made; response may influence reward, and could be $\emptyset$.
7. *Reward*: Y_t is the Bandit payout.
8. *Unobserved confounders*: $U_t \subseteq \mathbf{U}$ is any variable that influences multiple endogenous variables.
9. *History*: H_t is a special node representing the data structure of available L1, L2 and L3 information for all $\tau \in [t-1]$ (empty for $t = 1$); as such, it has no endogenous or exogenous parents. H_t may affect any node except (Π_t, X_t) in an L1 regime, and also Π_t in L2/L3 regimes.

Essentially, at each time-step t, $\mathbf{u}$ is drawn from $P(\mathbf{u})$, which along with h_t determines all other variables. Agents operate either in the L1 mode (letting autopilot behaviour determine actions) or L2/L3 mode, where they choose policies based on a decision criterion (Defn. 4).

The g-MAB formalism introduces a richer class of problems that permit more causal dependencies than familiar MABs. For instance, Fig. 1d shows a g-MAB where environment response depends on history and Fig. 1e has the policy directly affecting environment response. Fig. 1e also belongs to a category of problems recently studied by (Bell et al 2021) under the name *Newcomblike Decision Processes* (NDP), where rewards depend on (state, action, policy) tuples. G-MABs make such causal relationships explicit, and also include problems where policy is not directly read by the environment, but indirectly *inferred* (e.g. from history as in Fig. 1d). The following sections detail some fundamental theoretical properties of g-MABs.



(a) (b) (c) (d) (e)

Figure 1: SCMs for examples of (a) Regular MAB; (b) Contextual MAB; (c) MAB with Unobserved Confounders; (d) and (e) Generalized MAB; g-MABs include familiar Bandit problems, but permit more causal dependencies. Black arrows appear in M (i.e. L1 regime); additional red arrows appear in M_{σ_Π} (i.e. L2/L3 regime)

.

5 Agent and Environment Capabilities

The added degrees of freedom in a g-MAB allow us to investigate problems which aren't easily represented by regular MABs, such as those in the Sec. 8. Many of these are adversarial in nature, where the environment (perhaps modelling an actual opponent) competes with the agent. For such problems, conceptually we could regard filled-grey nodes as "agent-nodes" and empty-white ones as "opponent-nodes": agent sees grey nodes, agent makes a decision π_t, environment responds with empty nodes, and y_t is determined.

Definition 2: Confoundedness of a g-MAB. A g-MAB with graph G is said to be *confounded* if $(Y_t \not\perp\!\!\!\perp \Pi_t | C_t)_{G_{\underline{\sigma_\Pi}}}$.

This is the familiar notion of *backdoor paths* (Pearl 2009) from Π_t to Y_t in G_{σ_Π}. Note, this is separate from *Semi-Markovianity* (Bareinboim et al 2022). Fig. 1d is *confounded* but *Markovian*. In a sense, this captures the **agent's** capabil-

[1] M does not induce a *positive* observational distribution because $P(i_t, \pi_t) = 0$, for all $i_t \neq \pi_t$.

ity: it is less confident about the effect of its choices, as there may be confounders between policy and reward.

Definition 3: Theory of mind in a g-MAB. A g-MAB with graph G is said to have the *theory of mind* property if $(Y_t \not\perp\!\!\!\perp \Pi_t | X_t, C_t, I_t)_{G_{\sigma_\Pi}}$.

This term is borrowed from *theory of mind* in psychology (Heider 1958; Morton 1980), characterizing the **environment's** adversarial capability: it can access or infer the agent's actual policy. Reward cannot be screened off from agent policy by context, intent or action. Typical MDPs clearly do not have this property as $(r \perp\!\!\!\perp \pi | s, a, s')$. This is also similar in function to the characterization in (Press & Dyson 2012) of opponents with and without *theory of mind* in iterated Prisoner's Dilemma games. In Sec. 8, we offer an agent-vs-environment framing of such games.

We make three comments. (1) Neither *confoundedness* nor *theory of mind* implies the other in a g-MAB. Fig 1c is *confounded* without *theory of mind*, and vice-versa in Fig. 1e. (2) A g-MAB does not need a direct arrow from Π_t to Y_t to have *theory of mind*. Active paths could go, for instance, via H_t, or via colliders at I_t or C_t. (3) The graphical criterion in Defn. 3 empowers designers of agents to use causal inference tools to deny the environment this specific property, such as by identifying or constructing a *d-separating* set, similar to identifying *adjustment-admissible* sets for handling confounding (Pearl 2009).

6 Optimal Strategy and Counterfactual Limits

Usually, we think of a *decision strategy* in a MAB as a mapping from context to policy.[2] In an SCM, $(\mathbf{U} = \mathbf{u})$ represents a realized world among many possible ones, and we need the agent to pick some policy in each world. We ideally want a strategy that optimizes reward in expectation over $\mathbf{U} \sim P(\mathbf{u})$. However, there are structural limits to an agent's capability to do this.

Definition 4: Decision Strategy and Criterion. Given a g-MAB agent at time t, we have

- *Decision strategy:* δ is an implicit mapping from $\mathbf{U} \rightsquigarrow \Delta_X$, which tells us the policy the agent chooses in each possible world $(\mathbf{U} = \mathbf{u})$ at time t.
- A *decision criterion* is an explicit rule that the agent uses to decide on a policy π_t, based on historical data. We define 5 such criteria in Table 2.[3]

We abuse notation slightly and use $\delta(\mathbf{u}), \delta(c_t), \delta(i_t, c_t)$ etc. when it is clear we are mapping from $(\mathbf{u} \mapsto \pi_t), (c_t \mapsto \pi_t), (i_t, c_t \mapsto \pi_t)$ etc. Roughly, the criteria in Table 2 correspond to letting autopilot behaviour determine action, or optimizing using observational/experimental data. While interventional designs such as *randomized controlled trials* (Fisher 1935) are well-established, counterfactual optimizations of the form $\text{argmax}_x E[y_x|x']$ were thought to be unrealizable except for binary treatments (Shpitser & Pearl 2007; Pearl 2009, 341-344). (Bareinboim et al 2015; Forney et al 2017) first provided randomization algorithms where this counterfactual quantity can be optimized for arbitrary treatment cardinality.

This raises interesting questions. First, is there is a universal hierarchy among decision criteria? Depending on an agent's beliefs, available compute or cost of data-collection, it might prefer some criteria over others. Let us notate the expected reward (over $P(\mathbf{u})$) at time t as $Y^\emptyset$, Y^E, Y^C, Y^R, $Y^{R^\dagger}$, when the agent uses the *default, evidential, causal, regret and regret*† decision criteria, respectively.

Theorem 5. Given a g-MAB M, we have

i $Y^R \geq Y^C \geq Y^E$;
ii $Y^R \geq Y^\emptyset$; and
iii $Y^R = Y^C = Y^E \geq Y^\emptyset$, if M is *unconfounded*

Corollary 6. There are g-MABs where $Y^\emptyset \geq Y^C, Y^E$

Thm. 5(i) corroborates the intuition that optimizing using L3 data is strictly more powerful than using lower layers. It also directly addresses criticism of *causal decision theory* (Egan 2007; Ahmed 2014). Provided alleged failure modes can be represented as a g-MAB allowing exogenous intervention (see Hitchcock 2016 for examples; Dawid 2021; Pearl 2022), CDC strictly outperforms EDC. Any modification of EDC that resolves such failings would need to be isomorphic to L2-based interventional randomization.

Thm. 5(iii) allows an agent to use possibly lower-cost, lower-variance optimization with just L1 data, provided the g-MAB is *unconfounded.* Corollary 6 counter-intuitively shows that there are g-MABs where *not* intervening and following default behaviour can be *better* than optimizing using observations or interventions! This aligns with the findings in (Lee & Bareinboim 2018), which considers only deterministic optimal policies. The current setting allows optimal policies to be strictly stochastic. However, counterfactual randomization still fares best, per Thm. 5(ii).

A second interesting question is, given the power of L3-randomization, can we get even better guarantees with more complex counterfactual criteria such as RDC†? Indeed the space of L3 queries is vast (Bareinboim et al 2022, 16-19). Counterfactual optimization is also highly relevant to domains like personalized medicine and fairness (Mueller & Pearl 2022; Plecko & Bareinboim 2022). We may specifically want to customize treatment for individuals who *would have had* a certain outcome, or deliberately pick a sub-optimal criterion which satisfies a counterfactual fairness consideration. However, we prove below that no experiment exists that can solve such higher order counterfactual optimizations under realistic assumptions.

Theorem 7. Given a g-MAB M at time t, where Y_t cannot be fully determined by a soft-intervention $\Pi_t \leftarrow \sigma_\Pi(I_t, C_t, H_t)$, RDC† cannot be realized by any experiment.

Non-determinism of the reward is a weak assumption for most g-MABs, where Y_t is typically sampled from an (arm-specific) distribution. I.e., it has an independent source of

[2] To use graph-compatible notation, we use *policy*, π_t, for a distribution over actions, and *decision strategy* for a mapping to such a distribution. The latter is typically called a *policy* in MDPs.

[3] This distinction makes clear that an agent could theoretically optimize policies for each specific world it encounters, and so needs to devise rules to get as close to this as it can.

Decision Criterion	**PCH Layer**	**Type of data used**	**Output (stochastic)**	**Output (deterministic)**
Default ($\emptyset$)	L1	-	$(\Pi; \mathbf{U} = \mathbf{u})$	$(X; \mathbf{U} = \mathbf{u})$
Evidential (EDC)	L1	Observational	$\arg\max_{\pi} E[Y\|\pi, c]$	$\arg\max_{x} E[Y\|x, c]$
Causal (CDC)	L2	Interventional	$\arg\max_{\pi} E[Y\|do(\pi), c]$	$\arg\max_{x} E[Y\|do(x), c]$
Regret (RDC)	L3	Counterfactual	$\arg\max_{\pi} E[Y_\pi\|\pi', c]$	$\arg\max_{x} E[Y_x\|x', c]$
Regret† (RDC†)	L3	Counterfactual	$\arg\max_{\pi} E[Y_\pi\|\pi', y', c]$	$\arg\max_{x} E[Y_x\|x', y', c]$

Table 2: Examples of criteria agents can use at time t; *default decision criterion* corresponds to not intervening, letting autopilot behaviour pick policies; others involve optimizing policy choice using observational, interventional or counterfactual data.

exogenous noise U_Y. The upshot of Thm. 7 is that RDC sets the theoretical limit of counterfactual experimentation for a Bandit agent. If RDC† cannot be realized, an agent needn't worry about more complex criteria like maximizing $E[Y_x|y_{x'}, x'], E[Y_x|y_{x'}, y', x'], E[Y_x|o_{x'}, y']$ etc in the non-parametric setting. However, under parametric assumptions like exogeneity, monotonicity (Tian & Pearl 2000), or for certain graph families (Zhang et al 2022), we may be able to compute or bound such queries.

We have yet to discuss actual algorithms that can discover these policies. We next cover some results on the stability of optimal strategies.

7 Optimal Strategy and Stability

We have so far indexed variables with a time-step. However, the virtue of a Bandit setup is that the agent doesn't track history as contextual information.

While we can regard the agent's decision strategy as a mapping ($\mathbf{u} \mapsto \pi_t$), the agent only observes $(I_t, C_t = \pi'_t, c_t)$, at decision-making time. Using this, we can consider an optimal strategy to be one that maximizes $\sum_{\pi'_t, c_t} P(\pi'_t, c_t) E[Y_{\pi_t}|\pi'_t, c_t]$.[4] Further, it would be useful to know under what conditions this optimal strategy is stable, in at least two senses.

Theorem 9. If a given g-MAB M is *unconfounded* or does not have *theory of mind*, the optimal decision strategy is constant $\forall t \in \mathbb{N}$ (stationary).

This means we can stably over time handle g-MABs with *confoundedness* or with *theory of mind*, but not always both. As we shall see for the experiments in Sec. 8, the Bandit framework may be insufficient for such cases. However, provided a stationary optimal strategy (whether or not by the condition in Thm. 9), we can effectively drop the time-stamp t for the decision problem.

Definition 10: Nash equilibrium. Given a g-MAB at time t, a decision strategy $(\pi_t|\pi'_t, c_t) = \delta(\pi'_t, c_t)$ is said to be a *Nash equilibrium* at (π'_t, c_t) iff

$$E[Y_{\pi_t, x}|\pi'_t, c_t] = \max_{x'} E[Y_{\pi_t, x'}|\pi'_t, c_t], \forall x \in supp(\pi_t)$$

This notion has received a lot of attention recently, under the names *performative stability* in supervised learning (Perdomo et al 2020) and *ratification* in the NDP framework in (Bell et al 2021). This could be loosely seen as a single-agent sequential counterpart to *Nash* and *Stackelberg equilibria* (Korzhyk et al 2011; Weirich 1999) in multi-agent games, but with opponents abstracted into environment dynamics. Intuitively, this refers to a decision strategy where, given a context, no action outside the support of the policy chosen by the strategy has a greater expected reward than those actions supported by said policy.[5]

Lemma 11. An agent using a value-greedy Bandit algorithm in a g-MAB can only converge to a (stationary) strategy that is a Nash equilibrium.

Theorem 12. Given a g-MAB M at time t which does not have *theory of mind*, any optimal strategy is a Nash equilibrium.

Of course, Thm. 12 is only a possibility result for convergence, not a guarantee. Lemma 11 is well understood outside the RL framework, in traditional game theory: when players engage in "fictitious play", any convergence is only to a Nash equilibrium (Fudenberg & Levine 1998). In decision theory, (Weirich 1999) showed that if all players follow a "self-supporting" strategy (a form of ratification, mentioned earlier), the result is an equilibrium. (Bell et al 2021, Thm. 2) also prove Lemma 11 for a NDP Bandit (e.g. as in Fig 1c). Our extension to the general g-MAB case is straightforward. This result means usual Bandit algorithms are insufficient to handle many simple problems such as in Sec. 8. To the best of our knowledge, discovering optimal policies in g-MABs with *theory of mind* is an open area of research. Our Defn. 10 analyzes such equilibria in terms of simultaneous L2 interventions on both Π_t and X_t, which hopefully informs the design of algorithms to avoid the constraint in Lemma 11.

Common Bandit algorithms like $EXP3$ or *Thompson Sampling* (TS) are value-greedy in that they always seek out arms with higher expected reward, even if it means the expected reward for the overall policy is lowered. Lemma 11

[4] Recall, $(I_t = \pi'_t)$ means that Π_t *would have been* π'_t in the L1/observational regime.

[5] Defn. 10 is not a Nash equilibrium in the strict game theoretic sense, where an opponent's counter-policy must be optimal as well. Here, it is used to describe the simultaneous-move nature of a Bandit agent-vs-environment equilibrium.

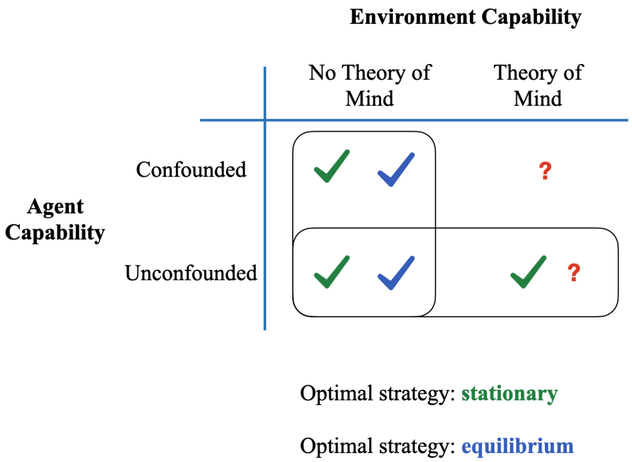


Figure 2: An emerging taxonomy of equilibrium and stationarity of Generalized Multi-Arm Bandit (g-MAB) problems

.

tells us that these algorithms will fail to discover the global optimal strategy if it is not a Nash equilibrium. We need some way of searching over the policy space directly.

We propose in Algo. 1 a policy-search algorithm TPS^C that approximates the optimal solution for g-MABs with Bernoulli actions. We essentially discretize the policy space and optimize over the mid-points of ranges, with agent actions being sampled from one of these policies. We consider Gaussian priors over the parameters of each policy choice, updated with each pull. This is a proof-of-concept which establishes an empirical upper bound on complexity (it would become intractable for larger action spaces). Note that this is more efficient than merely treating intention as a context variable, which we demonstrate empirically in the next section. In Line 7, we seed the priors for $E[Y_{\pi'}|\pi'], \forall \pi'$ by using past observational data $P_{obs}(Y|\pi')$ based on the counterfactual consistency axiom (Pearl 2010), which gives us

$$P(y_x|z, x) = P(y|z, x); \forall z \tag{1}$$

To the best of our knowledge, this is the first algorithm attempting to find an optimal solution to this problem. We next empirically test TPS^C on toy decision scenarios.

8 Experiments

In this section, we investigate three versions of the canonical Prisoner's Dilemma (PD), which cannot be described by a typical Bandit. Multi-agent RL approaches to this problem have long been studied (Sandholm & Crites 1996). (Leibo et al 2017) explore the distinction between cooperative policies and actions in PD-like games, and (Wang et al 2019) discuss how agents respond to opponents with changing strategies in PD. We present a single-agent perspective on this problem as represented by a generalized MAB.

The general setup involves an agent and an opponent who must each decide between two actions: **Cooperate** (a_0) and **Defect** (a_1). The agent's payout table is per Fig. 3a. In the following experiments, we assume the opponent values reciprocity and always strives to act as it expects the agent to.

Algorithm 1: Causal Thomson Policy Sampling (TPS^C) for Bernoulli Bandits

Input: P_{obs}, T
Parameter: k ; (discretization)

```
1:  Let $\{\pi_1, ...\pi_k\}$ be midpoints in discretized domain of $\Pi$.
2:  $\{r_{i,j}, n_{i,j}\} \leftarrow \{0, 1\}, \forall i, j \in [k], i \neq j$
3:  for $\tau = 1, ..., T$ do
4:      $\pi' \leftarrow intention(\tau)$ ; (get intention for trial)
5:      $\mu_{i,j} \leftarrow r_{i,j}/n_{i,j}$
6:      $\hat{\theta}_i \sim \mathcal{N}(\mu_{i,\pi'}, (1/n_{i,\pi'})), \forall i \in [k] \setminus \pi'$
7:      $\hat{\theta}_{\pi'} \leftarrow P_{obs}(Y|\pi')$ ; (consistency axiom)
8:      $i \leftarrow \arg\max_j(\hat{\theta}_j)$
9:      $y \leftarrow pull(\pi_i)$
10:     if $i = \pi'$ then
11:         $P_{obs}(Y|\pi') \leftarrow$ update
12:     else
13:         $r_{i,\pi'} \leftarrow r_{i,\pi'} + y$
14:         $n_{i,\pi'} \leftarrow n_{i,\pi'} + 1$
15:     end if
16: end for
```

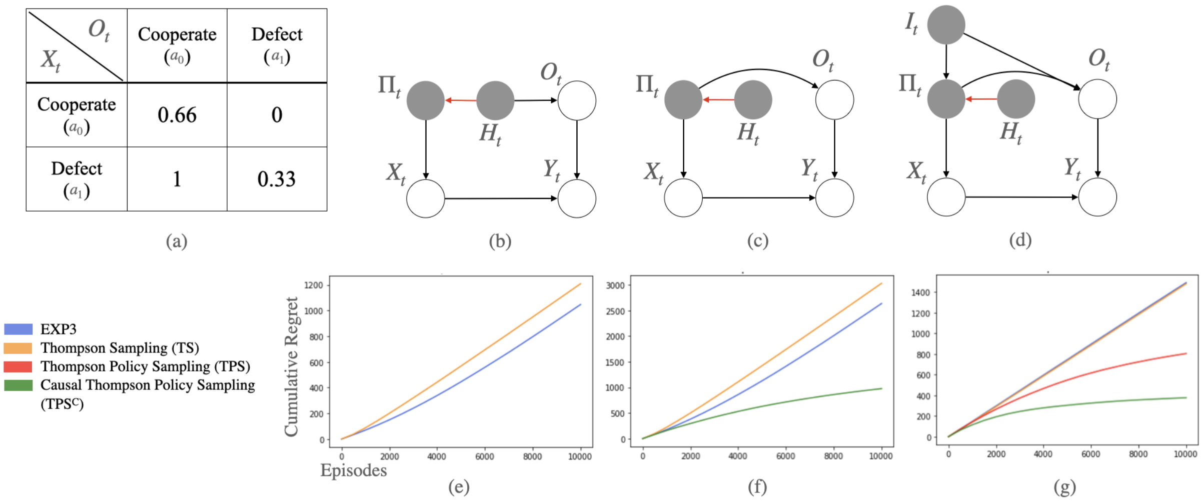
Experiment 1. The opponent decides its action at time t by estimating the agent's current policy based on recent history. It computes an exponential moving average of agent's historical actions to get a distribution over actions, and then samples its action from this policy. The graph of the corresponding g-MAB is in Fig. 3b. This graph has both *confounding* and *theory of mind.*

From Thm. 9, we cannot guarantee a stationary optimal policy. However, we constrain the agent to only change its policy smoothly and slowly. For a small enough threshold speed of change, and an unbiased opponent prior (at t_1), we do have a stationary optimal policy: $\{a_0 : 1.0, a_1 : 0.0\}$ (always-**Cooperate**). It can be verified that this is not a Nash equilibrium, since *given* this policy, the opponent will also **Cooperate**, and **Defect** would have higher expected reward for the agent.

Sure enough, we find value-greedy algorithms $EXP3$ and TS fail to discover the optimal policy, as shown by the cumulative regret being linear in T in Fig. 3e (averaged over 100 runs). In fact, they diverge to the policy of always-**Defect** (which is a Nash equilibrium), as shown in S.I. Relaxing the constraint on rate of change does not help, and in fact would worsen the speed of divergence.

Experiment 2. Consider a modification of Experiment 1 where the opponent feels observing history is not enough, and invests heavily to directly read agent policy at decision-time (e.g. by analyzing the agent's open-source code or running a simulation of the agent at decision-time). The graph now corresponds to Fig. 3c. This g-MAB has *theory of mind* but is *unconfounded*, so we are guaranteed a stationary optimal policy, which happens to remain $\{a_0 : 1.0, a_1 : 0.0\}$ (always-**Cooperate**).

Still, we find that $EXP3$ and TS fail to discover the optimal policy, as would any other value-greedy algorithm,



X_t \ O_t	Cooperate (a_0)	Defect (a_1)
Cooperate (a_0)	0.66	0
Defect (a_1)	1	0.33

Figure 3: (a) Agent payout table for Prisoner's Dilemma; Causal graphs of (b) Experiment 1, (c) Experiment 2, and (c) Experiment 3; Cumulative regret over training lifetime ($T = 10,000$) averaged over 100 runs, shown for (e) Experiment 1, (f) Experiment 2, and (g) Experiment 3; TPS^C shows sub-linear regret, indicating convergence to optimal policy, whereas $EXP3$ and TS appear to fail to converge; TPS^C also outperforms TPS which does not use counterfactual consistency seeding. Here, TPS and TPS^C use a discretization of 10.

since it is not an equilibrium.[6]. However, the TPS^C is able to identify the optimal policy region, as shown by the sub-linear cumulative regret in Fig. 3f. Convergence of TPS^C to the optimal policy is shown in S.I.

Experiment 3. Let us add more detail. Suppose the opponent has a limited surveillance budget. It decides to first read the agent's policy state 10 seconds before decision-making using a cheaper, delayed technology. If it sees the agent intending to **Defect** w.p. $P(a_1) < 0.5$, it uses the intended policy as basis for its own action. If the agent intends to **Defect** w.p. $P(a_1) \geq 0.5$, it switches to real-time monitoring until actual decision-time and samples its action from the agent's realized policy. The graph for this g-MAB is in Fig. 3d, and has both *confounding* and *theory of mind*.

The opponent's methodology happens to yield a stationary optimal strategy: the agent introspects 10 seconds before decision-time. If $P(a_1) < 0.5$ agent chooses $\{a_0 : 0.0, a_1 : 1.0\}$ (always-**Defect**). If $P(a_1) \geq 0.5$ agent chooses $\{a_0 : 1.0, a_1 : 0.0\}$ (always-**Cooperate**). In either case, the sole equilibrium for this problem remains to always-**Defect**.

Once again, $EXP3$ and TP incur significantly more cumulative regret over 10,000 episodes (averaged over 100 runs), while TPS^C demonstrates optimal convergence. In fact, the results in Fig. 3g also clearly shows TPS^C outperforms a version of the same algorithm, TPS, that does not use the counterfactual consistency axiom, and merely treats prior intent as a context variable. Details are in S.I.

Discussion. Prisoner's Dilemma often appears as a useful proxy for real-world scenarios. We used the g-MAB framework to study 3 such PD scenarios which cannot be represented as conventional Bandit problems. We showed how common algorithms can fail to discover even simple optimal strategies. We proposed an algorithm that searches directly over the policy space, and empirically tested its outperformance of $EXP3$ and TS. As mentioned, this is a proof-of-concept, and becomes intractable for large action spaces as discretizing the policy simplex is exponential in $|X_t|$. One possible solution could come from Bandits with continuous action spaces (Agrawal 1995). Discretization of continuous actions is famously ignorant of underlying geometry (Krishnamurthy et al 2019), so emerging methods combine this with smoothing operations over each interval in an efficient way (Majzoubi et al 2020). We leave this to future work.

9 Conclusion

Multi-agent problems are not easy to analyse in regular Bandit frameworks, requiring multi-agent RL or game theoretic formulations. We introduced a *generalized Multi-Arm Bandit* (g-MAB) formalism that can represent a rich class of interactive problems, thereby extending the fields of decision theory and causal RL. We established a universal hierarchy among decision theories in all problems that can be represented as a g-MAB, directly addressing the controversy between *evidential* and *causal* theorists, and proved that higher order counterfactual optimization cannot be realized by any physical experiment. We developed a taxonomy of g-MABs based on novel graphical conditions which guarantee stationary optimal strategies, and under which common Bandit algorithms cannot discover optimal policies. We proposed a policy-search algorithm which for the first time approximates the optimal solution for a general Bernoulli g-MAB, and demonstrated its performance against popular algorithms. We believe the g-MAB framework provides a

[6] To apply a confidence-bound algorithm here, it would need to be modified to allow stochastic policies over arms. UCB-style algorithms deterministically choose an $\arg\max$ arm in each round.

powerful and efficient causal toolkit to complement other perspectives in the design of safe and optimal AI agents.

A References

Agrawal, R. 1995. The Continuum-Armed Bandit Problem. *SIAM Journal on Control and Optimization*, 33(6): 1926–1951.

Ahmed, A. 2014. *Evidence, Decision and Causality*. Cambridge University Press.

Bareinboim, E.; Forney, A.; and Pearl, J. 2015. Bandits with Unobserved Confounders: A Causal Approach. In *NIPS*, 1342–1350.

Bareinboim, E.; Correa, J. D.; Ibeling, D.; and Icard, T. F. 2022. On Pearl's Hierarchy and the Foundations of Causal Inference. *Probabilistic and Causal Inference*.

Bell, J.; Linsefors, L.; Oesterheld, C.; and Skalse, J. 2021. Reinforcement Learning in Newcomblike Environments. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 22146–22157. Curran Associates, Inc.

Conitzer, V. 2015. A Dutch Book Against Sleeping Beauties Who Are Evidential Decision Theorists. *Synthese*, 192(9): 2887–2899.

Conitzer, V. 2019. Designing Preferences, Beliefs, and Identities for Artificial Intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01): 9755–9759.

Cooper, M.; Lee, J. K.; Beck, J.; Fishman, J. D.; Gillett, M.; Papakipos, Z.; Zhang, A.; Ramos, J.; Shah, A.; and Littman, M. L. 2019. Stackelberg Punishment and Bully-Proofing Autonomous Vehicles. In *Social Robotics - 11th International Conference, ICSR 2019, Madrid, Spain, Proceedings*, volume 11876 of *Lecture Notes in Computer Science*, 368–377. Springer.

Dawid, P. 2021. Decision-theoretic foundations for statistical causality. *Journal of Causal Inference*, 9(1): 39–77.

Egan, A. 2007. Some Counterexamples to Causal Decision Theory. *Philosophical Review*, 116(1): 93–114.

Fisher, R. A. 1935. *The Design of Experiments*. Edinburgh: Oliver and Boyd.

Forney, A.; Pearl, J.; and Bareinboim, E. 2017. Counterfactual Data-Fusion for Online Reinforcement Learners. In *Proceedings of the 34th International Conference on Machine Learning*, 1156–1164. PMLR.

Fudenberg, D.; and Levine, D. 1998. *The Theory of Learning in Games*, volume 1. The MIT Press, 1 edition.

Heider, F. 1958. *The psychology of interpersonal relations*. John Wiley & Sons Inc.

Hitchcock, C. 2016. Conditioning, intervening, and decision. *Synthese*, 193(4): 1157–1176.

Jeffrey, R. C. 1983. *The Logic of Decision*. New York, NY, USA: University of Chicago Press, 2nd edition.

Joyce, J. M. 1999. *The Foundations of Causal Decision Theory*. Cambridge Studies in Probability, Induction and Decision Theory. Cambridge University Press.

Korb, K. B.; Hope, L. R.; Nicholson, A. E.; and Axnick, K. 2004. Varieties of Causal Intervention. In Zhang, C.; W. Guesgen, H.; and Yeap, W.-K., eds., *PRICAI 2004: Trends in Artificial Intelligence*, 322–331. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-540-28633-2.

Korzhyk, D.; Yin, Z.; Kiekintveld, C.; Conitzer, V.; and Tambe, M. 2011. Stackelberg vs. Nash in Security Games: An Extended Investigation of Interchangeability, Equivalence, and Uniqueness. *Journal of Artificial Intelligence Research*, 41: 297–327.

Krishnamurthy, A.; Langford, J.; Slivkins, A.; and Zhang, C. 2019. Contextual bandits with continuous actions: Smoothing, zooming, and adapting. In Beygelzimer, A.; and Hsu, D., eds., *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, 2025–2027. PMLR.

Kuhn, S. 2019. Prisoner's Dilemma: The PD with Replicas and Causal Decision Theory. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2019 edition.

Lee, S.; and Bareinboim, E. 2018. Structural Causal Bandits: Where to Intervene? In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Leibo, J. Z.; Zambaldi, V.; Lanctot, M.; Marecki, J.; and Graepel, T. 2017. Multi-Agent Reinforcement Learning in Sequential Social Dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '17, 464–473. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Majzoubi, M.; Zhang, C.; Chari, R.; Krishnamurthy, A.; Langford, J.; and Slivkins, A. 2020. Efficient Contextual Bandits with Continuous Actions. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 349–360. Curran Associates, Inc.

Miller, J.; Milli, S.; and Hardt, M. 2020. Strategic Classification is Causal Modeling in Disguise. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Morton, A. 1980. *Frames of Mind*. Oxford, Oxford University Press.

Mueller, S.; and Pearl, J. 2022. Personalized Decision Making - A Conceptual Introduction. *Technical Report R-513*.

Oesterheld, C.; and Conitzer, V. 2021. Extracting Money from Causal Decision Theorists. *The Philosophical Quarterly*, 71.

Pearl, J. 2009. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition.

Pearl, J. 2010. On the Consistency Rule in Causal Inference: Axiom, Definition, Assumption, or Theorem? In *Epidemiology*, volume 21(6), 872–875.

Pearl, J. 2022. Causation and Decision: On Dawid's "Decision Theoretic Foundation of Statistical Causality". *Journal of Causal Inference*, 2.

Pearl, J.; and Mackenzie, D. 2018. *The Book of Why*. New York: Basic Books. ISBN 978-0-465-09760-9.

Perdomo, J.; Zrnic, T.; Mendler-Dünner, C.; and Hardt, M. 2020. Performative Prediction. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 7599–7609. PMLR.

Plecko, D.; and Bareinboim, E. 2022. Causal Fairness Analysis. *Technical Report R-90*.

Press, W. H.; and Dyson, F. J. 2012. Iterated Prisoner's Dilemma contains strategies that dominate any evolutionary opponent. *Proceedings of the National Academy of Sciences*, 109(26): 10409–10413.

Russell, S.; and Norvig, P. 2010. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 3 edition.

Sandholm, T. W.; and Crites, R. H. 1996. Multiagent reinforcement learning in the Iterated Prisoner's Dilemma. *Biosystems*, 37(1): 147–166.

Savage, L. J. 1954. *The Foundations of Statistics*. Wiley Publications in Statistics.

Shpitser, I.; and Pearl, J. 2007. What Counterfactuals Can Be Tested. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, UAI'07, 352–359. Arlington, Virginia, USA: AUAI Press. ISBN 0974903930.

Tennenholtz, M. 2004. Program equilibrium. *Games and Economic Behavior*, 49(2): 363–373.

Tian, J.; and Pearl, J. 2000. Probabilities of Causation: Bounds and Identification. In *Annals of Mathematics and Artificial Intelligence 28 (1), 287-313*.

Tversky, A.; and Kahneman, D. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157): 1124–1131.

von Neumann, J.; and Morgenstern, O. 1947. *Theory of games and economic behavior*. Princeton University Press.

Wang, W.; Hao, J.; Wang, Y.; and Taylor, M. 2019. Achieving Cooperation through Deep Multiagent Reinforcement Learning in Sequential Prisoner's Dilemmas. In *Proceedings of the First International Conference on Distributed Artificial Intelligence*, DAI '19. New York, NY, USA: Association for Computing Machinery. ISBN 9781450376563.

Weirich, P. 1999. Self-supporting strategies and equilibria in games. *American Philosophical Quarterly*, 36(4): 323–336.

Zhang, J.; Tian, J.; and Bareinboim, E. 2022. Partial Counterfactual Identification from Observational and Experimental Data. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvári, C.; Niu, G.; and Sabato, S., eds., *International Conference on Machine Learning, ICML*, volume 162 of *Proceedings of Machine Learning Research*, 26548–26558. PMLR.

B SUPPLEMENTARY MATERIAL AND PROOFS

Proof of Theorem 6

Let us use the following notation:

- $\pi^E = \delta^{EDC}(c_t)$, to refer to the policy chosen by EDC at time t, given a context c_t
- $\pi^C = \delta^{CDC}(c_t)$, to refer to the policy chosen by CDC at time t, given a context c_t

We can assume ties are broken arbitrarily in for multiple $\arg\max$.

Theorem 6(i). Given a g-MAB M, we have $Y^R \geq Y^C \geq Y^E$.

Proof.

$$Y^R = \sum_{\pi'_t, c_t} \underset{\pi}{max} E[Y_t | do(\Pi_t = \pi), \pi'_t, c_t].P(\pi'_t, c_t) \qquad \text{definition} \tag{2}$$

$$= \sum_{i_t, c_t} \underset{\pi}{max} E[Y_t | do(\Pi_t = \pi), i_t, c_t].P(i_t, c_t) \qquad (\Pi_t = I_t) \text{ in } M \tag{3}$$

$$\geq \sum_{i_t, c_t} E[Y_t | do(\Pi_t = \pi^C), i_t, c_t].P(i_t, c_t) \qquad \text{true for any } \pi_t \neq \arg\max \tag{4}$$

$$= \sum_{i_t, c_t} E[Y_t | do(\Pi_t = \pi^C), i_t, c_t].P(i_t | c_t).P(c_t) \tag{5}$$

$$= \sum_{i_t, c_t} E[Y_t | do(\Pi_t = \pi^C), i_t, c_t].P(i_t | do(\Pi_t = \pi^C), c_t).P(c_t) \qquad \text{Rule 3: } (I_t \perp\!\!\!\perp \Pi_t | C_t)_{G_{\overline{\sigma_{\Pi}(C_t)}}} \tag{6}$$

$$= \sum_{i_t, c_t} E[Y_t, i_t | do(\Pi_t = \pi^C), c_t].P(c_t) \tag{7}$$

$$= \sum_{c_t} E[Y_t | do(\Pi_t = \pi^C), c_t].P(c_t) \tag{8}$$

$$= Y^C \qquad \text{definition} \tag{9}$$

$$= \sum_{c_t} \underset{\pi}{max} E[Y_t | do(\Pi_t = \pi), c_t].P(c_t) \qquad \text{definition} \tag{10}$$

$$\geq \sum_{c_t} E[Y_t | do(\Pi_t = \pi^E), c_t].P(c_t) \qquad \text{true for any } \pi_t \neq \arg\max \tag{11}$$

$$= Y^E \qquad \text{definition} \tag{12}$$

□

Eqn. (6) uses Rule 3 from *do-calculus* (Pearl 2009). With incoming arrows to Π_t removed, the only outgoing arrow can be to O_t, Y_t or X_t. It can be easily verified that the *d-separation* condition holds for any configuration of the g-MAB due to colliders at O_t, Y_t and (possibly) X_t.

NB:

- On Eqn. (11): in general, $Y^E = \sum_{c_t} E[Y_t | do(\Pi_t = \pi^E), c_t].P(c_t) \neq \sum_{c_t} \underset{\pi}{max} E[Y_t | (\Pi_t = \pi), c_t].P(c_t)$. See Lemma 13.
- EDC is peculiar in that it *optimizes* over L1/observational data, but *enacts* its policy as an intervention. In general, there is no guarantee that the policy which has worked best in autopilot mode will provide the same expected reward under an intervention, as it did under default behaviour.
- In this an subsequent proofs we use the summation sign with Π_t and I_t, for succinctness of notation. These would need to be integral signs in the general case.

Theorem 6(ii). Given a g-MAB M, we have $Y^R \geq Y^{\emptyset}$.

Proof.

$$\begin{aligned}
Y^R &= \sum_{\pi'_t, c_t} \underset{\pi}{max} E[Y_t | do(\Pi_t = \pi), \pi'_t, c_t].P(\pi'_t, c_t) && \text{definition} && (13)\\
&\geq \sum_{\pi'_t, c_t} E[Y_t | do(\Pi_t = \pi'_t), \pi'_t, c_t].P(\pi'_t, c_t) && \text{true for any } \pi_t \neq \arg\max && (14)\\
&= \sum_{\pi'_t, c_t} E[Y_t | \pi'_t, c_t].P(\pi'_t, c_t) && \text{Consistency axiom: Eqn. (1)} && (15)\\
&= Y^{\emptyset} && \text{definition} && (16)
\end{aligned}$$

□

Lemma 13. If a g-MAB M is *unconfounded*, $Y^E = \sum_{c_t} \underset{\pi}{max} E[Y_t|(\Pi_t = \pi), c_t].P(c_t)$.

Proof.

$$\begin{aligned}
(\pi^E | c_t) &= \arg\max_{\pi} E[Y_t | (\Pi_t = \pi), c_t] && \text{definition} && (17)\\
Y^E &= \sum_{c_t} E[Y_t | do(\Pi_t = \pi^E), c_t].P(c_t) && \text{definition} && (18)\\
&= \sum_{c_t} E[Y_t | (\Pi_t = \pi^E), c_t].P(c_t) && \text{Rule 2: } (Y_t \perp\!\!\!\perp \Pi_t | C_t)_{G_{\underline{\sigma_\Pi}}} && (19)\\
&= \sum_{c_t} \underset{\pi}{max} E[Y_t | (\Pi_t = \pi), c_t].P(c_t) && \text{by Eqn. (17)} && (20)
\end{aligned}$$

□

The condition for rule 2 in Eqn. (19) is the definition of *unconfoundedness* of the g-MAB. As stated at the end of the proof of Thm. 6(i), this condition is not true in general.

Theorem 6(iii). Given a g-MAB M, we have $Y^R = Y^C = Y^E \geq Y^{\emptyset}$, if M is *unconfounded*.

Proof.

$$\begin{aligned}
Y^R &= \sum_{\pi'_t, c_t} \underset{\pi}{max} E[Y_t | do(\Pi_t = \pi), \pi'_t, c_t].P(\pi'_t, c_t) && \text{definition} && (21)\\
&= \sum_{i_t, c_t} \underset{\pi}{max} E[Y_t | do(\Pi_t = \pi), i_t, c_t].P(i_t, c_t) && (\Pi_t = I_t) \text{ in } M && (22)\\
&= \sum_{i_t, c_t} \underset{\pi}{max} E[Y_t | do(\Pi_t = \pi), c_t].P(i_t, c_t) && \text{Rule 1: } (Y_t \perp\!\!\!\perp I_t | C_t)_{G_{\overline{\sigma_\Pi}}} && (23)\\
&= \sum_{c_t} \underset{\pi}{max} E[Y_t | do(\Pi_t = \pi), c_t].P(c_t) && && (24)\\
&= Y^C && \text{definition} && (25)\\
&= \sum_{c_t} \underset{\pi}{max} E[Y_t | (\Pi_t = \pi), c_t].P(c_t) && \text{Rule 2: } (Y_t \perp\!\!\!\perp \Pi_t | C_t)_{G_{\underline{\sigma_\Pi}}} && (26)\\
&= Y^E && \text{Lemma 13} && (27)\\
&\geq \sum_{c_t, \pi_t} E[Y_t | (\Pi_t = \pi), c_t].P(c_t, \pi_t) && \text{true for any } \pi_t \neq \arg\max && (28)\\
&= Y^{\emptyset} && \text{definition} && (29)
\end{aligned}$$

□

The conditions for the rules applied in Eqns. (23) and (26) as a consequence of the g-MAB being *unconfounded*. Removing outgoing arrows from Π_t, there are no back-door paths which cannot be blocked by C_t.

Proof of Corollary 7

Let us use the following notation:

- $\pi^E = \delta^{EDC}(c_t)$, to refer to the policy chosen by EDC at time t, given a context c_t
- $\pi^C = \delta^{CDC}(c_t)$, to refer to the policy chosen by CDC at time t, given a context c_t

Corollary 7. There are g-MABs where, $Y^{\emptyset} \geq Y^C, Y^E$.

Proof. We show this by example. Consider the g-MAB M with a graph as shown in Fig. 4 and specified as follows. W.L.O.G, we only allow deterministic policies for this g-MAB.

$\underline{M}$
$U_t \sim \text{Bernoulli}(0.5)$
$I_t \leftarrow U_t$
$\Pi_t \leftarrow I_t$
$X_t \leftarrow \Pi_t$
$Y_t \leftarrow \neg U_t \oplus X_t$

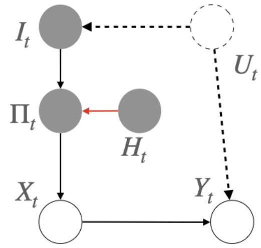


Figure 4: Graph for an example g-MAB illustrating Corollary 7
.

Under the *default decision criterion* (no intervention; autopilot behaviour), we have

$$\begin{aligned} Y^{\emptyset} &= \sum_{u_t} E[Y_t|u_t].P(u_t) && (30)\\ &= 0.5(E[Y_t|U_t = 1] + E[Y_t|U_t = 0]) && (31)\\ &= 0.5(E[\neg U_t \oplus X_t|U_t = 1, X_t = 1] + E[\neg U_t \oplus X_t|U_t = 0, X_t = 0]) && (32)\\ &= 1 && (33)\end{aligned}$$

Under the *evidential decision criterion* (intervention; optimization using L1 data), we have

$$\begin{aligned} \pi^E &= \arg\max_{\pi} E[Y_t|(\Pi_t = \pi)] && \text{definition} && (34)\\ &= \{0, 1\} && \text{both actions give reward=1 in L1} && (35)\\ & && && (36)\\ Y^E &= \sum_{u_t} E[Y_t|do(\Pi_t = \pi^E), u_t].P(u_t) && \text{definition} && (37)\\ &= 0.5(E[Y_t|do(X_t = 1), U_t = 1] + E[Y_t|do(X_t = 1), U_t = 0]) && && (38)\\ &= 0.5(E[\neg U_t \oplus X_t|U_t = 1, do(X_t = 1)] + E[\neg U_t \oplus X_t|U_t = 0, do(X_t = 1)]) && && (39)\\ &= 0.5(1 + 0) && && (40)\\ &= 0.5 && && (41)\end{aligned}$$

Under the *causal decision criterion* (intervention; optimization using L2 data), we have

$$\begin{aligned}
\pi^C &= \arg\max_{\pi} E[Y_t|do(\Pi_t = \pi)] && \text{definition} && (42)\\
&= \{0, 1\} && \text{both actions give reward=0.5 in L1} && (43)\\
& && && (44)\\
Y^C &= \sum_{u_t} E[Y_t|do(\Pi_t = \pi^C), u_t].P(u_t) && \text{definition} && (45)\\
&= 0.5(E[Y_t|do(X_t = 1), U_t = 1] + E[Y_t|do(X_t = 1), U_t = 0]) && && (46)\\
&= 0.5 && && (47)
\end{aligned}$$

Here, we have an example of a g-MAB where $Y^{\emptyset} > Y^E, Y^C$.

□

Proof of Theorem 8

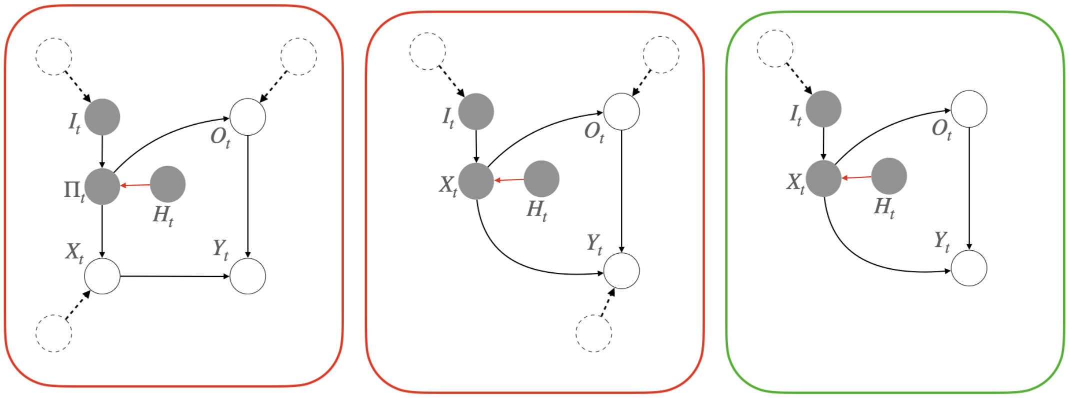


Figure 5: Examples of g-MABs that don't permit RDC† in red, and an example that does in green; most SCMs assume each node has an exogenous source of variation, but some toy game theoretic problems may involve a reward fully determined by agent and opponent action

.

Lemma 14. Given a g-MAB M where Y_t cannot be fully determined by a soft-intervention $\Pi_t \leftarrow \sigma_\Pi(I_t, C_t, H_t)$, $\exists U_Y \in \mathbf{U} \cup H_t$ s.t. there is a directed path $U_Y \rightarrow$ ... $\rightarrow Y_t$ which does not pass through Π_t or C_t.

Proof. It is easy to see that, if we cannot determine Y_t solely by the soft intervention described, then among the arguments for a deterministic function for Y_t, there must be some source of variation that is not fully captured by (Π_t, C_t). If every exogenous node (inc. history) that has a directed path to Y_t always reaches Y_t only via Π_t or C_t, then given (Π_t, C_t), Y_t is d-separated from any exogenous source of noise or intervention (i.e., is fully determined).

Finally, no exogenous node can have a directed paths to Y_t via I_t but **not** via Π_t. Recall that I_t has only one outgoing arrow, to Π_t, so any direct path into I_t goes further onward solely via Π_t.

□

In typical g-MABs, Lemma 14 is satisfied by there simply being a U_Y with a sole outgoing arrow to Y_t. But this definition also caters for g-MABs where Y_t is fully determined given X_t and an environment response O_t, in which case U_Y would be the independent source of variation for O_t, or the sampling variation of X_t. It also refers to unobserved confounders that do have a directed path to Π_t or C_t, because these also have a separate directed path to Y_t. We use the label U_Y W.L.O.G.

Lemma 15. Given a g-MAB M where Y_t cannot be fully determined by a soft-intervention $\Pi_t \leftarrow \sigma_\Pi(I_t, C_t, H_t)$, let $\mathbf{U_Y}$ be the set of all U_Y that satisfy the condition in Lemma 13. Then, Y_t can be fully determined by $\mathbf{U_Y}$ and (Π_t, C_t) in M and in M_{σ_Π}.

Proof. This follows from Lemma 14.

In M, $(Y_t \perp\!\!\!\perp \mathbf{U} \setminus \mathbf{U_Y} | \Pi_t, C_t)$. I.e., it is independent from any other source of interventional or exogenous variation, by definition of $\mathbf{U_Y}$. In M_{σ_Π}, we can only add a (possibly) new arrow from C_t or H_t to Π_t. This would not affect the condition $(Y_t \perp\!\!\!\perp \mathbf{U} \setminus \mathbf{U_Y} | \Pi_t, C_t)$ since $Y_t - C_t - \Pi_t$ cannot be a collider. □

Lemma 16. Given a g-MABs M, in general if the agent cannot estimate $P(y_{\pi_t} | \pi'_t, y'_t, c_t)$ for $y \neq y'$, RDC† cannot be computed.

Proof. W.L.O.G., consider the case where Y_t is binary. In order to compute RDC†, given some conditioning set (π'_t, y'_t, c_t), the agent needs to compute $\underset{\pi_t}{\arg\max} E[Y_{\pi_t} | \text{condition}] = \underset{\pi_t}{\arg\max} P(Y_{\pi_t} = 1 | \text{condition})$.

The mapping from $\langle \pi_t \mapsto P(Y_{\pi_t} = 1 | \text{condition}) \rangle$ could be any arbitrary, possibly discontinuous, function. As long as $P(Y_{\pi_t} = 1 | \text{condition}) < 1$, we can define arbitrary g-MABs where the mappings are such that we need to be able to estimate $P(Y_{\pi_t} = 1 | \text{condition})$ for any π_t to be able to compute the $\arg\max$. This includes g-MABs where the conditioning event is $(\pi'_t, c_t, Y_t = 0)$.

(Note that the actual computation of the $\arg\max$ may need to involve discretization or PAC-learning, if the space of Π_t is continuous. But this is unrelated to the Lemma statement.)

□

Theorem 8. Given a g-MAB M at time t, where Y_t cannot be fully determined by a soft-intervention $\Pi_t \leftarrow \sigma_\Pi(I_t, C_t, H_t)$, RDC† cannot be realized by any experiment.

Proof. For clarity of exposition let us drop the time subscript t. By Lemma 16, we have that, in general, an agent would need to be able to compute $P(y_\pi|\pi', y', c)$ for $(y \neq y')$ in order to realize RDC†. Note that Lemma 15 places no restriction on what type of g-MAB needs this requirement, in the non-parametric setting.

Let $\mathbf{U_Y}$ be the exogenous variables (noise, and possibly history at time t) defined in Lemma 14. From Lemma 15, we have that Y is determined by the mapping $\Pi \times C \times \mathbf{U_Y} \rightsquigarrow Y$. Let us note that in M (L1 regime), ($I = \Pi$), so we have that Y is determined by the mapping $I \times C \times \mathbf{U_Y} \rightsquigarrow Y$ in M.

Let us consider the event whose probability is the query $P(y_\pi|\pi', y', c)$. The event is short-hand for

i. $I, C, \mathbf{U_Y}$ are such that $Y = y'$ in M (i.e. L1 regime); **and**
ii. $\Pi, C, \mathbf{U_Y}$ are such that $Y = y$ in M_{σ_Π} (i.e. L2/L3 regime); **given**
iii. I, C, Π

Since the only unknown in the deterministic function of Y in both M_{σ_Π} and M is $\mathbf{U_Y}$, this query represents the probability mass on the values in the domain of $\mathbf{U_Y}$ s.t. the event conditions i-iii are satisfied.

Consider first the scenario where the directed paths from $\mathbf{U_Y}$ to Y that don't pass through Π or C (the paths described in Lemma 14) are not all mediated by X or O. I.e., this is the scenario that Y directly receives an un-mediated exogenous noise. Given that $\mathbf{U_Y}$ cannot be directly measured, the only way to gauge whether the un-mediated component of $\mathbf{U_Y}$ is in the part of the domain compatible with a certain value $(Y = y)$ is to *actually observe* $(Y = y)$. In other words, the only way to gauge whether the un-mediated component of $\mathbf{U_Y}$ is in the part of the domain compatible with event condition i (respectively, ii) is to *actually observe* $Y = y'$ (respectively, $Y = y$).

The crux of this proof is that, at time t, in order to observe $Y = y'$ (respectively, $Y = y$) to satisfy event condition i (respectively, ii), the physical process of L1/observation (respectively, L2/L3 experimentation) that produced this outcome has already been realized. Even if we were to repeat the learning under event condition iii (perhaps even rewinding to the $t-1$ check-point in simulation), and we observe $Y = y$ (respectively, $Y = y'$), we will only know the un-mediated component of $\mathbf{U_Y}$ is compatible with event condition ii (respectively i). Since it is not possible for any experiment to know that the un-mediated component of $\mathbf{U_Y}$ simultaneously satisfies event condition i and ii, given iii, the query cannot be estimated. By Lemma 14, RDC† cannot be realized by any experiment.

For completeness, let us consider the scenario that there is no un-mediated noise directly affecting Y, and that $\mathbf{U_Y}$ is always mediated through X and/or O. The same argument in the previous paragraph can be extended by replacing "$Y = y$" with "$(X = x, O = o)$ such that $Y = y$", and "$Y = y'$" with "$(X = x, O = o)$ s.t. $Y = y'$".

□

Definition 16: Interceptibility. If a node W is said to be *interceptible* in a g-MAB M if it satisfies:

- $\exists \mathbf{Z} \subseteq An(W)$ s.t. $An(W) \setminus De(\mathbf{Z}) = An(\mathbf{Z}) \setminus \mathbf{Z}$; and
- $\mathbf{Z}$ is *available* at decision-making time

$\mathbf{Z}$ is called its *interception set*. Essentially, W is *interceptible* if it can be fully determined as a function of variables which are *available* at decision-making time. As a special case, Π_t is *interceptible* with *interception set* I_t because this is available to the agent via introspection. Also, by definition, all nodes that are themselves *available* at decision-making time are *interceptible*, with *interception sets* as themselves.

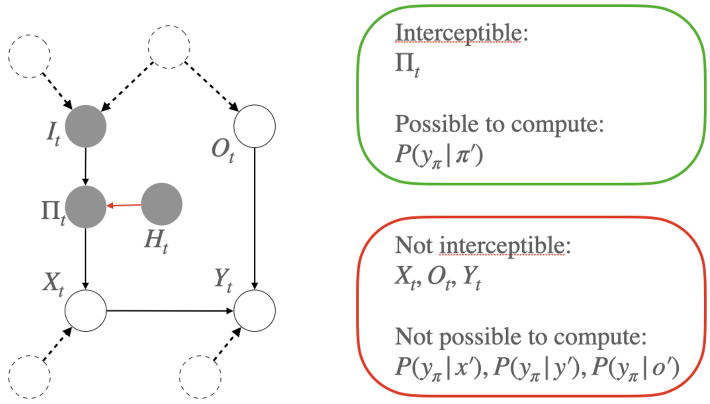


Figure 6: Examples of nodes which are *interceptible* and which are not; *interceptible* nodes can be used in an optimization criterion

.

Proof of Theorem 9

Let us use the notation:

- $\pi_t^O = \delta_t^O(\pi', c)$ to refer to an optimal policy at time t, given an arbitrary (π', c)

Assumption 17. The g-MAB M described in this theorem is such that:

i. Exogenous variables $\mathbf{U}$ have the same distribution, at all t.
ii. Structural equation f_V remains the same for $V \in \mathbf{V}$, at all t.

NB: This does not stop history, H_t, from affecting variables, including exogenous ones. The SCM can be re-parameterized to have H_t affect the relevant endogenous variables.

Theorem 9. If a given g-MAB M is *unconfounded* or does not have *theory of mind*, the optimal decision strategy is constant $\forall t \in \mathbb{N}$ (stationary).

Proof.

$$E_t[Y_{\pi_t^O}|\pi', c] = \max_\pi E_t[Y_\pi|\pi', c] \qquad \text{definition} \qquad (48)$$

Case 1. M is *unconfounded*

We have that $(Y_\tau \perp\!\!\!\perp \Pi_\tau | C_\tau)_{G_{\underline{\sigma_\Pi}}}$ for $\tau = t, t'$. This means, given $(C_\tau = c)$, any active path from Π_τ to Y_τ is only via a directed causal path.

I.e.,

$$\text{Given } (C_\tau = c),\ do(\Pi_\tau = \pi) \text{ has the same causal effect on } Y_\tau, \forall \tau = t, t' \qquad (49)$$

$$E_t[Y_{\pi_t^O}|c] = \max_\pi E_{t'}[Y_\pi|c], \forall t, t' \qquad \text{by Line (49)} \qquad (50)$$

$$E_t[Y_{\pi_t^O}|\pi', c] = \max_\pi E_{t'}[Y_\pi|\pi', c], \forall t, t' \qquad \text{Rule 1: } (Y_\tau \perp\!\!\!\perp I_\tau|C_\tau)_{G_{\underline{\sigma_\Pi}}} \qquad (51)$$

Rule 1 applies in Eqn. (51) because any back-door path from Y_τ to I_τ not via C_τ would introduce confounding that we are assuming does not exist.

Case 2. M does not have *theory of mind*

We have that $(Y_\tau \perp\!\!\!\perp \Pi_\tau | X_\tau, I_\tau, C_\tau)_{G_{\sigma_\Pi}}$ for $\tau = t, t'$. This means, given $(I_\tau, C_\tau = \pi', c)$, any active path from Π_τ to Y_τ is only via a directed causal path through X_τ.

I.e.,

$$\text{Given } (I_\tau, C_\tau = \pi', c),\ do(\Pi_\tau = \pi) \text{ has the same causal effect on } Y_\tau, \forall \tau = t, t' \tag{52}$$

$$E_t[Y_{\pi_t^O} | \pi', c] = \max_\pi E_{t'}[Y_\pi | \pi', c], \forall t, t' \qquad \text{by Line (52)} \tag{53}$$

Eqns. (51) and (53) conclude the proof showing that if π_t^O is the optimal policy at t, it is also the optimal policy at $t' \neq t$, for some arbitrary observed (π', c). □

Proof of Lemma 11

Given a g-MAB, M, let us assume that a Bandit algorithm converges to a terminal strategy $\delta^*(i_t, c_t) = (\pi^*|i_t, c_t)$.

Assumption 18. The value-greedy Bandit algorithm described in this Lemma meets the following criteria:

i. The number of times each arm is explored over the learning sequence is infinite in the limit, as $t \to \infty$
ii. $\lim_{t\to\infty} P(E[Y_{\pi_t,x}|\pi'_t, c_t] \leq \max_{x'} E[Y_{\pi_t,x'}|\pi'_t, c_t] - \delta) = 0, \forall x \in supp(\pi_t)$ and $\delta > 0$

Assumption 18(i) limits the consideration to algorithms that allow for infinite exploration, such as $EXP3$ and Thompson Sampling (TS) (NB: these algorithms still allow for convergence to a deterministic policy in the limit). However, this assumption excludes from consideration confidence-bound based algorithms such as the standard version of UCB, where exploration of sub-optimal arms stops after a finite number of time-steps. UCB, in its standard form, also does not allow the discovery of optimal policies which are strictly stochastic.

Assumption 18(ii) states that the algorithm behaves greedily w.r.t. the expected reward of each arm, regardless of the policy which chose the arm. This is a standard feature of model-free Bandit algorithms, which update their preference weights for each arm based on the reward sampled under arm-pulls, without tracking the underlying policy. The condition states that, as the algorithm's policy converges to δ^*, the probability of it choosing actions which are sub-optimal (given the current π_t) compared to the actions that are not chosen tends to 0. It can be easily verified that $EXP3$ and TS obey this feature.

Lemma 11. An agent using a value-greedy Bandit algorithm in a g-MAB can only converge to a (stationary) strategy that is a Nash equilibrium.

Proof. As (Bell et al 2021, Thm. 2) observes, this simply follows from Assumption 18(ii). The arguments extend to any SCM.

Let x' be a "sub-optimal" arm given π^*, meaning that $\exists\delta > 0$ such that

$$E[Y_{\pi^*,x'}|\pi'_t, c_t] \leq \max_x E[Y_{\pi^*,x}|\pi'_t, c_t] - \delta \tag{54}$$

Since $\pi_t \to \pi^*$, for a large enough t, we have that

$$E[Y_{\pi_t,x'}|\pi'_t, c_t] \leq \max_x E[Y_{\pi_t,x}|\pi'_t, c_t] - \frac{\delta}{2} \tag{55}$$

By Assumption 18(ii), we have the probability of picking such an arm x' tends to 0 in the limit, thus giving us

$$\lim_{t\to\infty} \pi_t(x') = \pi^*(x') = 0, \text{given } (i_t, c_t) \tag{56}$$

$(\pi^*|i_t, c_t)$ always places zero weight on such "sub-optimal" actions. In other words, the terminal strategy always satisfies the equilibrium condition in Defn. 10.

□

Proof of Theorem 12

Let us use the notation:

- $\pi^O = \delta^O(\pi'_t, c_t)$ to refer to an optimal policy at time t, given (π'_t, c_t)

Theorem 12. Given a g-MAB M at time t which does not have *theory of mind*, any optimal strategy is a Nash equilibrium.

Proof.

$$
\begin{aligned}
E[Y_{\pi^O}|\pi'_t, c_t] &= \max_{\pi} E[Y_\pi|\pi'_t, c_t] && \text{definition} && (57)\\
E[Y_{\pi^O}|\pi'_t, c_t] &\leq \max_{x'} E[Y_{\pi^O, x'}|\pi'_t, c_t] && && (58)\\
&= \max_{x'} E[Y_{x'}|\pi'_t, c_t] && \text{Rule 3: } (Y_t \perp\!\!\!\perp \Pi_t | I_t, C_t)_{G_{\overline{X_t \Pi_t(I_t, C_t)}}} && (59)
\end{aligned}
$$

Rule 3 applies here because we are guaranteed that M does not have the *theory of mind* property, which means $(Y_t \perp\!\!\!\perp \Pi_t | X_t, I_t, C_t)_{G_{\sigma_\Pi}}$. Let us define π^G to be a greedy policy that deterministically chooses an $\arg\max$ arm which maximizes the quantity in Eqn. (53).

$$
\begin{aligned}
E[Y_{\pi^O}|\pi'_t, c_t] &\leq E[Y_{\pi^G}|\pi'_t, c_t] && \text{Eqn. (52), (53)} && (60)\\
E[Y_{\pi^O}|\pi'_t, c_t] &= E[Y_{\pi^G}|\pi'_t, c_t] && \text{Eqn. (51)} && (61)\\
&= \max_{x'} E[Y_{x'}|\pi'_t, c_t] && \text{definition} && (62)\\
&= \max_{x'} E[Y_{\pi^O, x'}|\pi'_t, c_t] && \text{Rule 3: } (Y_t \perp\!\!\!\perp \Pi_t | I_t, C_t)_{G_{\overline{X_t \Pi_t(I_t, C_t)}}} && (63)\\
\sum_{x \in supp(\pi^O)} \pi^O(x).E[Y_{\pi^O, x}|\pi'_t, c_t] &= \max_{x'} E[Y_{\pi^O, x'}|\pi'_t, c_t] && \text{definition} && (64)\\
E[Y_{\pi^O, x}|\pi'_t, c_t] &= \max_{x'} E[Y_{\pi^O, x'}|\pi'_t, c_t] && \forall x \in supp(\pi^O) && (65)
\end{aligned}
$$

□

The last step of the proof is simply because each term in the weighted average on the LHS of Eqn. (58) is individually upper-bounded by the RHS, and is non-negative. Therefore each term must be equal to the RHS.

The last line gives the condition for π^O to be a Nash equilibrium at arbitrary (π'_t, c_t).

Details of (non-causal) Thompson Policy Sampling (TPS)

This version of Thompson Sampling also discretizes the Bernoulli policy space and converges to the optimal policy range. However, it does not make use of the counterfactual consistency axiom (refer to Sec. 7 for an explanation), whereas TPS^C is able to use past observational data to seed the policy-arm priors when the consistency conditions are satisfied (Lines 7 and 11 in Algo. 1).

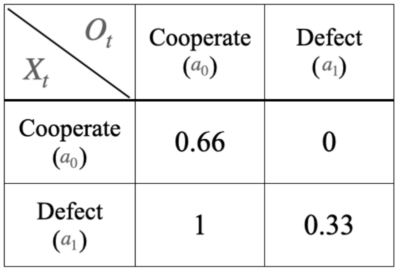
Algorithm 2: Thomson Policy Sampling (TPS) for Bernoulli Bandits

```
**Input**: $T$
**Parameter**: $k$ ; (discretization)
Let $\{\pi_1, ...\pi_k\}$ be midpoints in discretized domain of $\Pi$.
$\{r_{i,j}, n_{i,j}\} \leftarrow \{0, 1\}, \forall i, j \in [k], i \neq j$
for $\tau = 1, ..., T$ do
    $\pi' \leftarrow intention(\tau)$ ; (get intention for trial)
    $\mu_{i,j} \leftarrow r_{i,j}/n_{i,j}$
    $\hat{\theta}_i \sim \mathcal{N}(\mu_{i,\pi'}, (1/n_{i,\pi'})), \forall i \in [k]$
    $i \leftarrow \arg\max_j(\hat{\theta}_j)$
    $y \leftarrow pull(\pi_i)$
    $r_{i,\pi'} \leftarrow r_{i,\pi'} + y$
    $n_{i,\pi'} \leftarrow n_{i,\pi'} + 1$
end for
```

Details of Experimental Set-Up

Common details across experiments:

- Number of episodes per experiment, $T = 10,000$
- Each experiment was repeated 100 times, and the results were averaged to get representative charts
- Agent reward for each episode as per payout table in Fig. 7



O_t X_t	Cooperate (a_0)	Defect (a_1)
Cooperate (a_0)	0.66	0
Defect (a_1)	1	0.33

Figure 7: Payout table for Prisoner's Dilemma

.

Experiment 1

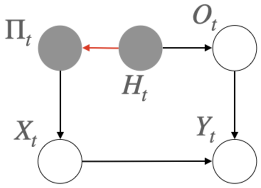


Figure 8: Graph for the g-MAB in Experiment 1

.

Optimal policy	$\pi^O(a_0) = 1.0; \pi^O(a_1) = 0.0$
Nash equilibrium	$\pi^E(a_0) = 0.0; \pi^E(a_1) = 1.0$
Is optimal policy an equilibrium?	No

Table 3: Optimal and equilibrium policy for Experiment 1

Agent policy:

- Chosen according to algorithms $EXP3$ and TS
- Agent is constrained to only update its policy at a small speed (i.e. its Bernoulli probability of picking a_1 at episode t must be within a range of the probability at $t - 1$)

Opponent policy at episode t:

- Chosen as an exponential moving average of all past agent actions in h_t: $p_t = \alpha.x_{t-1} + (1 - \alpha).p_{t-1}, \forall t > 1$, where p_t is the opponent's probability of choosing a_1 (defect) at episode t
- The opponent starts off unbiased ($p_1 = 0.5$)
- We set the weight parameter, $\alpha = 0.7$, to better approximate the latest agent policy (recall, the agent can only change its policy with a limited speed)

Algorithms and convergence:

- $EXP3$ was run with learning rate $\eta = 0.01$; increasing eta only made the divergence faster

- TS was run with Gaussian priors on arms, with μ as empirical mean, and σ^2 as $(100/\text{count})$; using Gaussian priors allowed us to smooth out the priors with higher variance, as a way to ensure the agent contraint on speed of update.
- Both algorithms are value-greedy in the sense of Defn. 17. By Lemma 11, they converge to the Nash equilibrium and diverge from the optimal policy, as shown in Fig. 9.

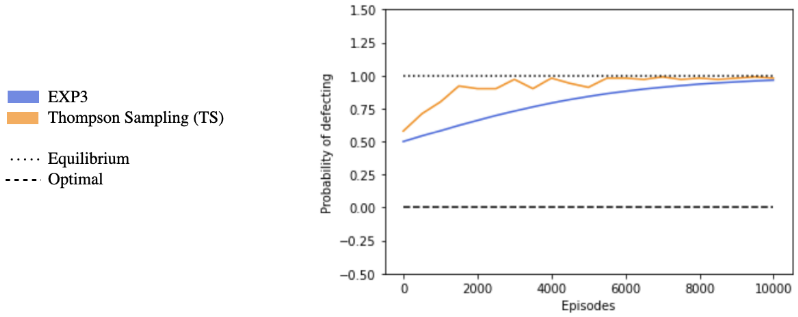


Figure 9: Results for $EXP3$ and TS in Experiment 1

.

Experiment 2

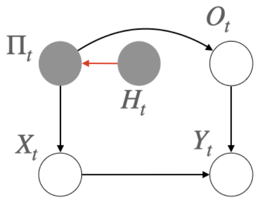


Figure 10: Graph for the g-MAB in Experiment 2

.

Optimal policy	$\pi^O(a_0) = 1.0; \pi^O(a_1) = 0.0$
Nash equilibrium	$\pi^E(a_0) = 0.0; \pi^E(a_1) = 1.0$
Is optimal policy an equilibrium?	No

Table 4: Optimal and equilibrium policy for Experiment 2

Agent policy: chosen according to algorithms $EXP3$, TS and TPS^C.

Opponent policy at episode t: equal to agent policy π_t.

Algorithms and convergence:

- $EXP3, TS$ were run as per Experiment 1
- TPS^C was run as per Algo. 1, with a discretization parameter, $k = 10$.
- $EXP3, TS$ are value-greedy and converge to the Nash equilibrium and diverge from the optimal policy.
- TPS^C is not value-greedy and discovers the optimal policy range (of the discretized intervals), as shown in Fig. 11.

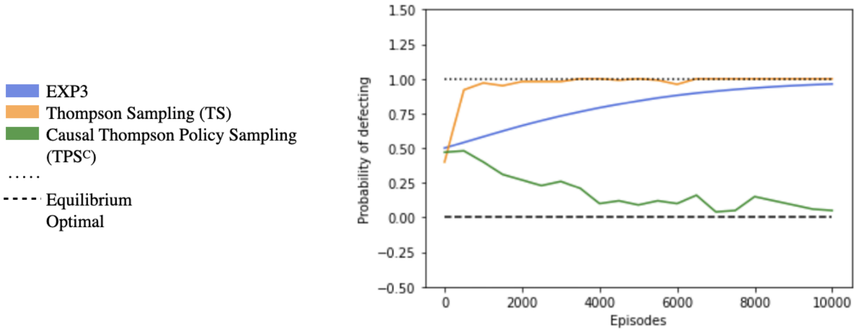


Figure 11: Results for $EXP3, TS$ and TPS^C in Experiment 2

.

Experiment 3

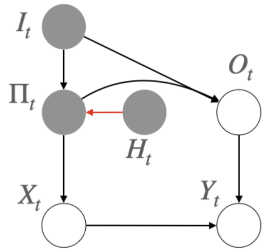


Figure 12: Graph for the g-MAB in Experiment 3

.

	$i_t(a_1) < 0.5$	$i_t(a_1) \geq 0.5$
Optimal policy	$\pi^O(a_0) = 0.0; \pi^O(a_1) = 1.0$	$\pi^O(a_0) = 1.0; \pi^O(a_1) = 0.0$
Nash equilibrium	$\pi^E(a_0) = 0.0; \pi^E(a_1) = 1.0$	$\pi^E(a_0) = 0.0; \pi^E(a_1) = 1.0$
Is optimal policy an equilibrium?	Yes	No

Table 5: Optimal and equilibrium policy for Experiment 3

Agent policy: chosen according to algorithms $EXP3$, TS, TPS and TPS^C.

Opponent policy at episode t:

- If agent's intended probability of defecting $i_t(a_1)$ is under 0.5, then equal to agent's intended policy, i_t
- If agent's intended probability of defecting $i_t(a_1)$ is at least 0.5, then equal to agent actual policy π_t

Algorithms and convergence:

- $EXP3, TS$ were run as per Experiment 1
- TPS^C, TPS were run as per Algo. 1 and 2, respectively, with a discretization parameter, $k = 10$.
- $EXP3, TS$ are value-greedy and converge to the Nash equilibrium; this coincides with the optimal policy when agent's intended probability of defecting is under 0.5, but diverges from the optimal policy when the intended probability of defecting is at least 0.5.
- TPS^C, TPS are not value-greedy and approximate the optimal policy in both cases, as shown in Fig. 13. However, because they are sensitive to original intent, they converge slower to their stationary policy than $EXP3, TS$.
- TPS^C outperforms TPS in terms of regret, which does not make use of past observational data by exploiting the counterfactual consistency axiom. In Fig. 13, we see TPS^C converge more smoothly than TPS to the optimal policy.

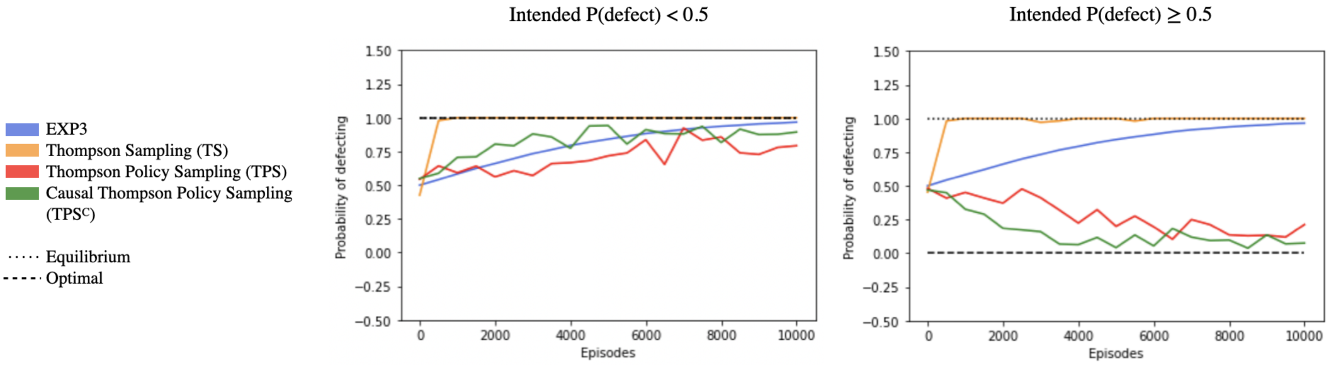


Figure 13: Results for $EXP3, TS, TPS$ and TPS^C in Experiment 3

.