Adversarial Causal Deception Scenarios: Preliminary Modeling and Policy Formation

Milo Fritzen[1], Andrew Forney[1], Adrienne Raglin[2], Sunny Basak[2], Peter Khooshabeh[2],

[1]Loyola Marymount University, Los Angeles, CA 90045, USA
[2]DEVCOM Army Research Laboratory

Abstract

As autonomous systems become increasingly integrated into society, they may be presented with falsehoods and adversarial deception that can harmfully skew their perception of the state of their environment. Still tasked with making decisions in these contexts, it is thus important for systems to be aware of and plan around potential deception for optimal decision-making. This is inherently a causal problem given that deception often masquerades inaction as action, like a phishing attempt pushing urgency by spoofing a need that is not real. Agents successfully navigating these *adversarial causal deception scenarios* must understand what *acts* can truly change the state, and where *misinformation* is merely advertising that it has changed. This paper provides a causal framework for considering the portions of the state that are vulnerable to action and misinformation (an Adversarial Causal Decision Network (ACDN)), and outlines a planning process (Adversarial Causal Expectimax Search (ACES)) to avoid adversarial deception attempts in pursuit of the agent's purpose.

Introduction

The study of multi-agent systems wherein actors are competing to satisfy their contrasting objectives without complete knowledge of their environment or adversaries has a rich history in game theory (Osborne and Rubinstein 1994), as does the design of artificial agents created to optimally play in these scenarios (e.g., by modeling the environments as POMDPs (Kaelbling, Littman, and Cassandra 1998) or teaching agents to perform through reinforcement learning (Sutton and Barto 2018)). Consequently, many techniques have attempted to empower the reasoning capabilities of agents by modeling the latent environment space (Russell and Norvig 2021), but this task is complicated by several perturbations that the current work explores, including the ability for agents to deceive one another into mistaking parts of the true state and wherein each agents' objectives/motives may be unknown to one another. Take, for instance, the spread of misinformation or propaganda through social media (Allcott and Gentzkow 2017): in even a single thread between two opposing political parties, the true state can be mistaken by the public due to information asymmetry between politicians and voters, misinformation campaigns, bots and bad-faith actors attempting to control the narrative, etc. all the while those consuming of these threads also question the motives of those contributing (e.g., debating if a politician's stance is ideologically aligned with theirs vs. self-serving in some hidden way).

A surprising number of other real-world scenarios follow a recipe similar to political misinformation and how consumers navigate dilemmas of trust:

1. *Cybersecurity:* phishing attempts often involve bad-actors misrepresenting roles, identities, and urgency in order to obtain credentials or extort funds (Wright, Johnson, and Kitchens 2023).
2. *Organizational Politics:* promotions, tenure, and merit assignments may incentivize self-promotion or reputation warfare to achieve objectives (Gross et al. 2021).
3. *Jury Deliberation:* in which there is ample reasoning under uncertainty of presented evidence and testimony, portions of which may be false (Vrij and Hartwig 2021; Frank et al. 2004).
4. *AI Safety and Alignment:* in which agents may employ deception as an emergent behavior from a misspecified reinforcement signal (Park et al. 2024; Arnold, Kasenberg, and Scheutz 2017; Banovic et al. 2023), or must navigate relationships with deceptive users.

This work attempts to unite the themes present in these example scenarios through a unifying modeling framework inspired by their common components: notably, each involves the following traits:

- Multi-agent systems wherein each agent may be acting to satisfy their own, potentially contrasting, objectives.
- A partially-observed state whose knowledge is relevant for maximizing these potentially contrasting objectives.
- The ability of agents to choose from different actions that observe, manipulate, and misrepresent the state.
- A causal structure to the state's components as these scenarios involve motifs of attribution and motivational explanation between the observations and actors alike.

We thus deem (and will later formalize) these scenarios as Adversarial Causal Deception Scenarios (ACDS) and begin by choosing one motivating example as a simplified proxy environment to study in this initial exploration.

Consider the social deception game Mafia, a game in which some players are randomly chosen to be mafia members (the "evil" team), and others are the townspeople (the "good" team). Many variants of this game also include other roles, such as the Jester, whose goal is to be voted out by the group (and thus has different priorities from the other players), or the Detective, who can learn other players' roles. Importantly, all roles are hidden from the public player base until the end of the game. Each round, the mafia members secretly eliminate one player, and the Detective secretly investigates one player, learning if they are a mafia member or not (in some variants, this information can also be tainted). Then, the remaining players may choose to eliminate someone from the game by a majority vote. The game ends when all the mafia members have been eliminated, they outnumber the non-mafia members, or another player's win condition is met, such as the Jester being voted out. If a player were to claim that they were the Detective, it would not directly alter the game state, as no roles are changed and no eliminations occur, yet the information revealed can still influence the decisions of other players.

An associative reasoning agent may correlate players who confidently claim powerful roles with actually being those roles, and may therefore choose not to vote to eliminate them when they otherwise would have. This becomes especially problematic in scenarios where a Mafia player deliberately lies about their role to avoid suspicion or manipulate the group's voting behavior. However, a player on the good team may also benevolently lie that they are the Detective so as to allow the true Detective time to gather information. An agent empowered with causal reasoning could recognize that the act of claiming a role does not causally influence a player's true role, but is instead *motivated by* certain roles that will possess certain priorities. This approach is preferable because it can account for the true causal relationship between observed statements and hidden roles, plan against adversarial attempts at targeted disinformation, strategically use its own abilities to maximal effect, and compare observed performance of other agents to hypothetical optimal performance of agents in each role.

This short-paper's main contributions do not fully solve, but rather motivate and seed, approaches to ACDSs by:

1. Formalizing a novel Adversarial Causal Decision Network (ACDN) to model the environment, opposing agents' priorities, and where / how agents may investigate, affect, and misrepresent the hidden state.
2. Demonstrate how ACDNs lead to a novel form of Adversarial Causal Expectimax Search (ACES) for a preliminary form of policy extraction.
3. Provide recommendations for how the application of the above can be situated in more sophisticated reasoning environments alongside a discussion of limitations.

Background

Modeling ACDSs requires tools that can represent the hidden state, as well as how and where agents can observe and manipulate it, in accordance with their potentially varying objectives. Prior work provides a non-multi-agent approach to this by way of Causal Decision Networks that we will build on to meet the demands of ACDSs.

Definition 1. (*Causal Decision Network (CDN)*) A CDN (Odell et al. 2026) is a 4-tuple $C = \langle M, I, A, Y \rangle$: an enhancement to traditional Structural Causal Models (SCMs, (Pearl 2009)) that are used for specifying not only how variables in the causal system interact, but also which are amenable to observation and/or intervention; a CDN consists of:

1. **SCM**, $M = \langle U, V, F, P(u) \rangle$, a traditional SCM encoding the cause–effect relations between system variables.
2. **Investigations**, I, a set of variables in the SCM amenable to observation, i.e., whose value the agent can choose to expose. Represented graphically as triangles pointing to variables in $V \cup U$ that the investigation exposes.
3. **Acts (AKA interventions)**, A, a set of variables in the SCM that can be intervened upon, i.e., whose value can be forced to a desired one despite the "natural" functions deciding its value. Represented graphically as squares pointing to variables in $V \cup U$ that the intervention affects (implemented using the causal do-operator).
4. **Utility-scored Outcome Variables**, Y, whose values are scored by some utility function f deciding the quality of the outcome, conditional upon the state, i.e., $f(Y \mid S)$. Higher utility implies more desirable. Represented graphically as any variable in $V \cup U$ pointing to a diamond (i.e., the parents of any diamond are those scored by the utility function).

Using a CDN, a causal form of expectimax search can be employed to determine the best sequence of investigations and acts that maximize the expected utility of outcomes:

Definition 2. (*Causal Expectimax Search (CES)*) (Odell et al. 2026) A causal expectimax search strategy for solving a CDNs explores a type of causal expectimax tree that plans for all possible sequences of investigations and acts to select the optimal action that maximizes expected utility according to the *agent state*, $s = \langle e, a, T, D \rangle$ (for collected evidence, e, chosen acts a, remaining time budget T [with each action costing some c time], and remaining decisions D):

1. **Max Nodes** representing state s and the points at which to make choices. Their value $V(s)$ is the maximum of any chance-node child value $Q(s, a)$ such that
$$V(s) = \max_a Q(s, a).$$
2. **Chance Nodes** representing action a chosen from state s determining the probability-weighted possible transitions / next states s' from a given max node state s with value $Q(s, a) = \sum_{s'} P(s' \mid s) * V(s')$.
3. **Transition probabilities** $P(s' \mid s)$ that are computed using the CDN and the current agent state s such that:
   (a) Investigation transitions, $\mathrm{inv}(H)$, are computed via the counterfactual $P(H_a \mid e)$, where e is all current observed evidence in S and a all current acts.
   (b) Act transitions, forcing some variable H to obtain value h, are assumed to be executed with certainty via the causal do-operator, $do(H = h)$ (Pearl 2012).

4. **Terminal nodes**, scored via the expected utility:

$$EU[Y_a \mid e] = \sum_{y \in Y} P(y_a \mid e) * f(y)$$

of the outcome variables specified in the CDN, Y.

5. **CES Investigation Transitions:** For investigations $\text{inv}(H)$, adds observed value h that is not presently part of the agent state, i.e., $H \notin s$, with transition $Tr(s, \text{inv}(H)) = \langle (e \cup \{H = h\}), a, T - c_{inv(H)}, D - \{\text{inv}(H)\}\rangle$.
6. **CES Act Transitions:** For intervention $\text{do}(H = h)$, must erase any previous observations or interventions in the descendants of H, denoted $\text{desc}(H)$, since acts resample variables along the causal path from H, re-enabling them to be later investigated or intervened upon, with transition $Tr(s, \text{do}(H = h)) = \langle (e - H) - \text{desc}(H), (a \cup \{H = h\}) - \text{desc}(H), T - c_{do(H_t = h_t)}, (D - \{\text{do}(H = h), \text{inv}(H)\}) \cup \text{og\_desc}(H)\rangle$, where $\text{og\_desc}(H)$ denotes the original set of investigations and acts licensed from the CDN for only the descendants of H.
7. **CES Null Choice Transitions:** To advance CES closer to a terminal state ($D = \emptyset$) without a forced choice to lower utility, CES defines $Tr(s, d = \text{do}(X = \emptyset)) = Tr(s, d = \text{inv}(X = \emptyset)) = \langle e, a, T, D - d\rangle$.

CES builds what is essentially a causally-empowered expectimax tree that plans the optimal sequence of investigations and acts that maximize the agent's expected utility from the starting state, which allows for policy extraction ("solving" the CDN, similar to how policies are extracted from sufficiently specified POMDPs). However, there are missing pieces from CDNs if they are to be used to address ACDSs, including the accounting of multiple agents and deception.

Adversarial Causal Decision Networks

To model the Mafia game, there are certain aspects of the state that only some agents can see (e.g., the Detective in exposing teams, the mafia in knowing who each other are), states that are desirable to some agents but not all (e.g., the Jester being voted off being uniquely desirable to that player), and the ability to apply deception (e.g., in some variants of the game, the ability to leave a journal that other players can read upon being killed, which may contain false info). Since CDNs are missing these important facets, there is a need to encode where and which agents can affect the flow of information and how their competing priorities lead to optimal play. We thus define Adversarial Causal Decision Networks to extend traditional CDNs by including multiple acting agents rather than just one, and to introduce a new type of action that represents misinformation.

Definition 3. (*Adversarial Causal Decision Network (ACDN)*) An ACDN is a CDN (Def. 1) extended to model:

1. **Actors**, $\mathcal{A} = \{\mathcal{A}^{(1)}, \mathcal{A}^{(2)}, ..., \mathcal{A}^{(k)}\}$ representing each actor / agent in the ACDS. Each previously modeled action and utility node now belongs to a particular agent to account for the different abilities and priorities of each. Agent actions and priorities are now subscripted to distinguish them, e.g., $inv(H)_{\mathcal{A}^1}$ indicates that agent $\mathcal{A}^{(1)}$ investigated variable H.
2. **Misinformation**, Mis, a set of variables in the underlying CDN's SCM for which the agent may misrepresent the state via action $mis(H)_{\mathcal{A}^{(i)}} = h'$, the effect of which is that if a separate agent $A^{(k \neq i)}$ later tries to investigate that same state variable H, i.e., $inv(H)_{\mathcal{A}^k}$, they will observe the fabricated state value $H = h'$ regardless of whether or not h' is the true state value. Represented as heptagons in ACDN graphical representations.

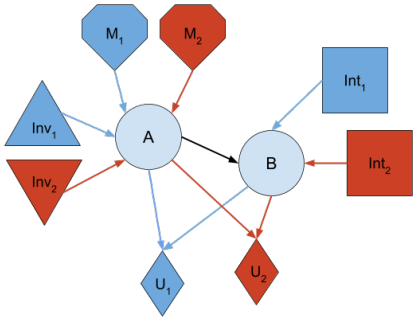

Figure 1: Simple ACDN with two state nodes, A, B and symmetrical actions.

Fig. 1 demonstrates a simple ACDN with symmetrical actions and priorities that may be used to represent a simple deception game that is 0-sum between 2 adversaries, $\mathcal{A}^{(1)}, \mathcal{A}^{(2)}$. Notably, the diamonds (U) represent the state variables/values the agents prioritize (e.g., if node $A \in \{0, 1\}$, $\mathcal{A}^{(1)}$ may desire $A = 0$ and vice versa for $\mathcal{A}^{(2)}$), with triangle actions (Inv) representing investigation actions possible at node A. The agent may perform these actions to learn the value of the connected node. The heptagon actions (M) are misinformation actions. These alter the value that investigations will return, allowing agents to deceive one another. Lastly, the squares (Int) represent interventions the players can make, changing the value of the attached node.

In such an ACDS, assuming turn-taking for actions much like what happens in the day/night phases of games like Mafia, what follows is an intricate dance of deciding where, when, and in what order to act, especially planning against an adversary who can do the same. For instance, because neither agent may *change* the state of node A, but its state may interact with the optimal choice for how to change node B, the optimal starting move may be to either learn the true state of A first or to misinform an adversary about it. A procedure is thus needed to determine the optimal course of how such a deception-capable agent would behave.

Adversarial Causal Expectimax Search

Given an ACDN, the order in which an agent chooses to investigate, intervene, or misinform its adversaries matters. Optimal agents will thus need to plan around where an opponent is potentially working to thwart them. To account for this, we amend CES (Def. 2) to perform the newly-required record-keeping for offline policy extraction of ACDNs.

Definition 4. *(Adversarial Causal Expectimax Search (ACES)).* ACES expands CES through the following:

1. The modeling agent must now track each agent state for each actor separately, i.e., $s_{\mathcal{A}^{(i)}} = \langle e_{\mathcal{A}^{(i)}}, a_{\mathcal{A}^{(i)}}, T_{\mathcal{A}^{(i)}}, D_{\mathcal{A}^{(i)}} - d \rangle \; \forall \; \mathcal{A}^{(i)} \in \mathcal{A}$
2. Misinformation actions now alter each *other* agent state's evidence e except for the misinforming agent such that $Tr(X_{t,\mathcal{A}^{(k \neq i)}}, mis_{\mathcal{A}^{(i)}}(H_t = h'_t)) = \langle (e_{\mathcal{A}^{(i)}} \cup mis(H_t = h'_t)), a_{\mathcal{A}^{(i)}}, T_{\mathcal{A}^{(i)}}, D_{\mathcal{A}^{(i)}} \rangle$, indicating that actor $\mathcal{A}^{(i)}$ will be misinformed about variable H_t, seeing it as value h'_t if $inv_{\mathcal{A}^{(i)}}(H_t)$ is chosen in the future.
3. Terminal states can now be scored from the maximizing agent's perspective in a number of different strategies, including:
   - *Maximizer:* maximizing agent makes choices that maximize its ACDN's expected utility nodes, ignoring other agents'.
   - *Spoiler:* the maximizing agent makes choices that *minimize* its adversaries' expected utility nodes, ignoring its own.
   - *Differ:* the maximizing agent makes choices that maximize the difference between its expected utility and its adversaries'.

   For simplicity, we assume adversarial agents adopt the Maximizer strategy from the perspective of their utility function.

Note that this modified scoring mechanism for terminal states results from agents potentially having different priorities that make the environment non-constant-sum, and thus, traditional minimax / expectimax value functions no longer hold in ACDSs. Like CES, terminal states are reached if (1) some scenario-specific state condition is met, (2) all agents run out of actions to choose, or (3) all agents' time budgets are exhausted. Fig. 2 shows an example format of an ACES tree that should explore all possible interventions, investigations, and misinformation actions in a 2-agent game with upside-right triangles representing the maximizing agent's actions, and upside-down triangles their opponent's. The planning proceeds from the root representing the current state and alternating turns for opponents to act (assuming a constant-time-cost for each action choice).

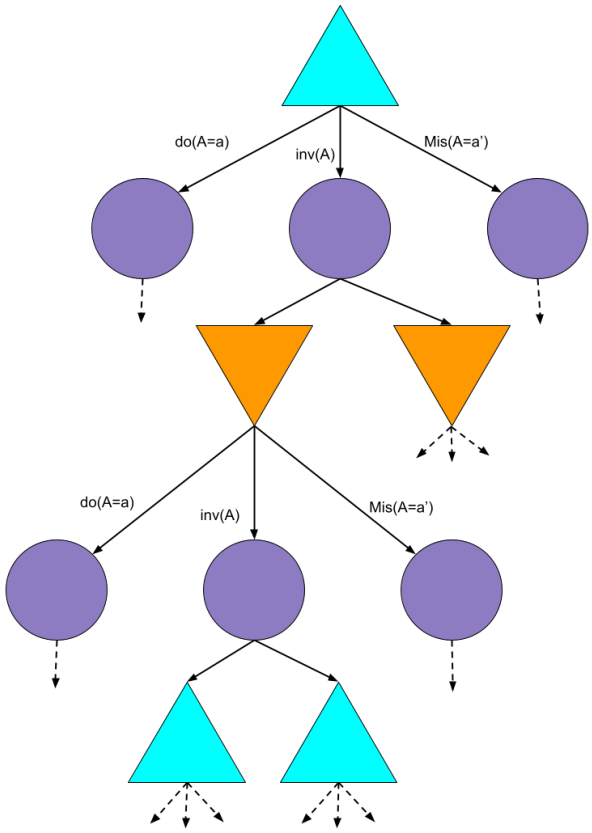

Figure 2: An Adversarial Causal Expectimax Search tree for the misinformation game.

Discussion

The main benefits of using ACES instead of Minimax, Expectimax, or Reinforcement Learning are that it enables the agent to consider moves made by opponents with goals other than minimizing its score, even in scenarios where information is unknown or unreliable, and with any structural knowledge of the causal system exploited. However, the feasibility of the theoretical ACES tree creation and exploration starts to falter at scale.

Mainly, the complexity of the ACES tree grows exponentially with larger games that have additional actions and actors. This, in turn, becomes a problem for the computational efficiency of performing ACES to find an optimal move in a non-trivial game. On the other hand, because terminal states are scored by expected value, a depth-limited variant may work to provide an imperfect real-time heuristic to guide agent choices. In terms of application, the ability to craft an ACDN conforming to intricate game rules in a game like Mafia is complex and requires additional consideration; it is more probable that an ACDN would be situated in a hierarchical planning agent to decide upon general strategy (i.e., to bluff or not) rather than lower-level actions.

Future studies may investigate fitting ACDN-equipped agents into proxy social deception games like a minified version of Mafia, or even pair these with large language models to improve their reasoning capabilities. Comparisons to traditional, associative intelligence solutions like reinforcement learners or base LLMs could demonstrate the added benefits of causal modeling.

References

Allcott, H.; and Gentzkow, M. 2017. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, 31(2): 211–236.

Arnold, T.; Kasenberg, D.; and Scheutz, M. 2017. Value alignment or misalignment-what will keep systems accountable? In *AAAI Workshops*, 81–88.

Banovic, N.; Yang, Z.; Ramesh, A.; and Liu, A. 2023. Being trustworthy is not enough: How untrustworthy artificial intelligence (AI) can deceive the end-users and gain their trust. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1): 1–17.

Frank, M. G.; Feeley, T. H.; Paolantonio, N.; and Servoss, T. J. 2004. Individual and small group accuracy in judging truthful and deceptive communication. *Group Decision and Negotiation*, 13(1): 45–59.

Gross, C.; Debus, M. E.; Ingold, P. V.; and Kleinmann, M. 2021. Too much self-promotion! How self-promotion climate relates to employees' supervisor-focused self-promotion effectiveness and their work group's performance. *Journal of Organizational Behavior*, 42(8): 1042–1059.

Kaelbling, L. P.; Littman, M. L.; and Cassandra, A. R. 1998. Planning and Acting in Partially Observable Stochastic Domains. *Artificial Intelligence*, 101(1–2): 99–134.

Odell, A. M.; Forney, A.; Raglin, A.; Basak, A.; and Khooshabeh, P. 2026. Constrained Causal Decision Dilemmas. In Degen, H.; and Ntoa, S., eds., *HCI International 2025 – Late Breaking Papers, Part XV*, Lecture Notes in Computer Science, 240–265. Gothenburg, Sweden: Springer Cham.

Osborne, M. J.; and Rubinstein, A. 1994. *A Course in Game Theory*. Cambridge, MA: MIT Press. ISBN 9780262150415.

Park, P. S.; Goldstein, S.; O'Gara, A.; Chen, M.; and Hendrycks, D. 2024. AI deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5).

Pearl, J. 2009. *Causality*. Cambridge university press.

Pearl, J. 2012. The do-calculus revisited. *arXiv preprint arXiv:1210.4852*.

Russell, S.; and Norvig, P. 2021. *Artificial Intelligence: A Modern Approach*. Pearson, 4 edition. ISBN 9780134610993.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 2 edition. ISBN 9780262039246.

Vrij, A.; and Hartwig, M. 2021. Deception and lie detection in the courtroom: The effect of defendants wearing medical face masks. *Journal of applied research in memory and cognition*, 10(3): 392–399.

Wright, R. T.; Johnson, S. L.; and Kitchens, B. 2023. Phishing susceptibility in context: A multilevel information processing perspective on deception detection. *MIS Quarterly*, 47(2): 803–832.