Online Learning of Deceptive Policies under Intermittent Observation

Anonymous authors

***Abstract*— In supervisory control settings, autonomous systems are not monitored continuously. Instead, monitoring often occurs at sporadic intervals within known bounds. We study the problem of deception, where an agent pursues a private objective while remaining plausibly compliant with a supervisor's reference policy when observations occur. Motivated by the behavior of real, human supervisors, we situate the problem within Theory of Mind: the representation of what an observer believes and expects to see. We show that Theory of Mind can be repurposed to steer online reinforcement learning (RL) toward such deceptive behavior. We model the supervisor's expectations and distill from them a single, calibrated scalar — the expected evidence of deviation if an observation were to happen now. This scalar combines how unlike the reference and current action distributions appear, with the agent's belief that an observation is imminent. Injected as a state-dependent weight into a KL-regularized policy improvement step within an online RL loop, this scalar informs a closed-form update that smoothly trades off self-interest and compliance, thus sidestepping hand-crafted or heuristic policies. In real-world, real-time hardware experiments on marine (ASV) and aerial (UAV) navigation, our ToM-guided RL runs online, achieves high return and success with observed-trace evidence calibrated to the supervisor's expectations. Evaluation videos and code are available at `tom-deception.github.io`.**

I. Introduction

> *"It is unnecessary for a prince to have all the good qualities; it is very necessary to appear to have them."*
>
> Niccolò Machiavelli
> *The Prince, ch. 18*

Autonomous agents are rarely watched continuously; doing so is often impractical, computationally heavy, and cost-prohibitive. In practice, a supervisor sidesteps continuous monitoring by sampling behavior intermittently [1]. Such behavior is especially true for real, human supervisors who possess limited computational and observational resources. In many cases, the interval between successive samples of an agent's behavior is unpredictable, yet bounded [2]: operations impose a minimum reset interval after any check and a latest permissible inspection horizon mandated by policy or logistics. As a result, observations arrive sporadically with *known* lower and upper inter-observation gaps. Because only these observations are seen, they are the sole basis for the supervisor's belief about the agent. An agent pursuing its own goal can therefore be judged compliant if the snapshots it presents align with the supervisor's reference expectations. This paper, therefore, poses a precise question: given a task i.e., a supervisor-provided reference policy describing how the agent is expected to behave, and bounds on the time between observations, can an agent choose its subsequent actions so as to achieve a private objective while ensuring that, whenever an observation occurs, its state–action behavior remains plausibly consistent with the reference? In other words, we consider deception under intermittent observation.

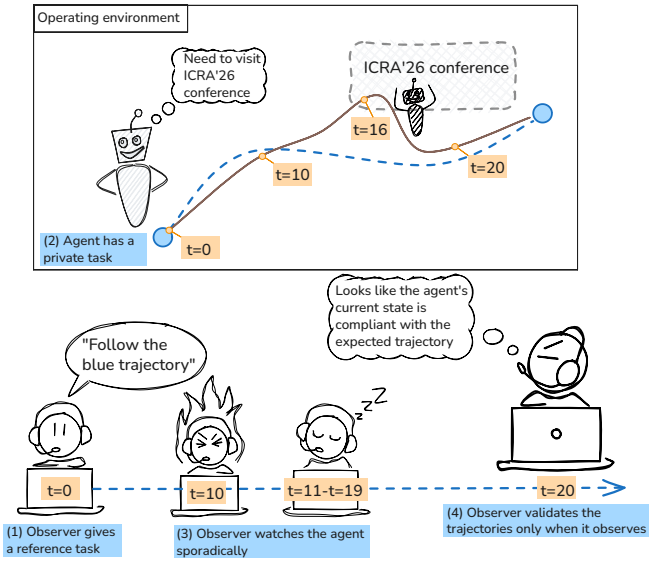

Fig. 1. Illustration of deception under intermittent observation. An observer assigns a task and specifies a reference behavior; the agent pursues a private objective. The observer sees only sporadic snapshots (state–action pairs). We ask: can we design an online policy that looks compliant when watched yet exploits unobserved intervals without exceeding a detectability budget?

Traditional approaches towards this problem are split along two lines. Planning-based deception and goal-recognition work shapes trajectories so that an outside observer infers the "wrong" goal or cannot confidently identify the true one [3]–[6]. These methods are elegant but typically assume known models and effectively continuous observation, and they do not allocate compliance specifically to the moments when observations actually arrive, making them conservative when glimpses are sparse and timed. Reinforcement learning (RL) variants [7], [8] offer complementary tools: imitation and inverse RL can push agents toward supervisor-like behavior, while KL-regularized policy search [9] keeps a learned policy globally close to a reference or prior. Yet imitation methods optimize for matching everywhere and suffer covariate shift [10] when observations are sparse, whereas KL-regularization typically applies a uniform or hand-tuned penalty that ignores when observation is likely; neither couples detectability to the timing of being seen. Formal treatments [11]–[13] of deception in optimal control capture intent masking via belief-dependent rewards or path-distribution criteria, but are largely offline and assume observer models and monitoring that do

not reflect bounded, intermittent checks. Thus, there exists a gap in existing planners and frameworks, which either enforce compliance everywhere or ignore when observations occur, offering no way to shape behavior to appear compliant precisely at observation times.

The core difficulty is twofold. First, an information problem: under bounded but unpredictable monitoring, the agent must anticipate at decision time how its next move would be judged if it were observed now. Second, a control problem: translate that anticipated judgment into sequential behavior that appears compliant at observation times while advancing the private objective when it is safe to do so, without brittle policies, over-constraining behavior between observations, and while remaining below a global threshold on detectability.

We turn to Theory of Mind (ToM) [14], [15] as the natural tool for the first difficulty—anticipating how the supervisor would judge the next move if a snapshot were taken now. Directly instantiating this idea is tempting: let the agent carry a full model of the supervisor's mind and plan under it. The appeal is obvious—no reward shaping, a principled notion of what "looks right," and decisions grounded in how behavior will be judged at the moment of inspection. In our setting, however, a full ToM is not identifiable. A full model of ToM would track the supervisor's latent tests, thresholds, and evolving posterior over trajectories and future scrutiny; under sporadic, bounded observations this is data-poor, computationally heavy, and brittle online. We show how to distill ToM from exactly what is available here and nothing more: (i) the supervisor's reference policy, and (ii) the monitoring rule that checks occur after bounded gaps, by computing a single-step evidential risk as the current chance of being seen. Conditioned into standard RL updates, this single scalar is sufficient to drive an online policy that tightens toward the reference as observation becomes likely and relaxes between checks.

Statement of contributions: (i) We introduce a calibrated Theory-of-Mind signal for deception under intermittent observation (ii) We develop an online RL method conditioned on this signal: a SAC-style actor–critic with a KL-to-reference regularizer whose state-dependent temperature is the ToM signal. (iii) We validate our approach on hardware in real-time, real-world settings on two platforms: an autonomous surface vehicle (ASV) in a marine setting and a 3D aerial platform (UAV), showing scalable, online operation with consistent gains in return and success. (iii-a) to spur further research in this domain, we release the code, trained models, and the dataset on our project website.

II. Related Work

Foundations of Deception: Classical work in psychology frames deception as a dynamic, interactive process to model how deceivers adapt in real time to exploit truth-bias and audience feedback [16]. In ethology–the study of animal behavior–deception is cast as fitness-improving false communication, highlighting the cognitive preconditions for strategic deceit [17]. Robotics inherited both lenses: early algorithms specified when deception is warranted and how it is executed, borrowing biological strategies such as decoys and false cues [18]; broader ethical architectures examined when to deceive or trust and how deception impacts human trust [19], while taxonomies clarified what counts as robot deception and cataloged types and domains [20]. Motion-generation work distinguished legible vs. predictable actions [6] and synthesized robot trajectories that mislead observers about true goals while remaining feasible [21].

Formal and Game-Theoretic Treatments of Deception: Control and game theory formalize deception as optimizing over beliefs and observations. Differential/pursuit–evasion and planning games analyze how intermittent or partial observations can be exploited to mislead adversaries about goals [22]–[24]. Belief-augmented optimal control treats the observer's belief as state, penalizing evidence of true intent [11]. Related supervisory settings seek policies that minimize divergence between observation distributions under reference and agent policies–often using KL as the detectability metric and HMMs for intermittent checks [25]. Recent dynamics/game-theory lines take an equilibrium-based approach for deception, hypergames with misaligned perceptions, and motion-level deceits that keep multiple goals plausible [26]–[28]. *Limitations:* These approaches typically presume known observer models and observation channels, can be conservative, and face adaptation challenges under sporadic observations.

Learning-Based Deception: Learning-based approaches realize deception via data-driven alignment to a reference or via opponent modeling. Imitation and adversarial imitation seek policies indistinguishable from expert behavior, addressing covariate shift and distribution matching [8], [29], [30]. KL-regularized RL constrains policy updates relative to a prior/reference, naturally keeping behavior near expected norms [9]. Machine Theory of Mind meta-learns to infer other agents' beliefs/goals, enabling behavior that anticipates observer expectations [15]. Multi-agent studies show learned deception emerging as agents exploit ambiguity to mislead teammates or opponents [31], [32]. Safety work formalizes deception for learning agents and proposes objective modifications to mitigate it [33]. Recent RL systems explicitly mask rewards/goals or bias exploration to preserve intent ambiguity in continuous control [34], [35], and scale deception via GNN-based planners [36] or via multi-agent equilibria with counter-deception [37]. *Limitations:* these methods enforce global trajectory indistinguishability or adversarially train detectors without modeling when supervision occurs, are not adapted to work online and they depend heavily on data which makes these approaches underperform in data-scarce or sporadic observation settings.

Heuristic and Planning Approaches to Deception: Planning-based deception controls what observers can infer by shaping paths, motions, or environments. Goal recognition design selects environments that maximize the observation needed to disambiguate true from decoy goals [3], while deceptive path planning formalizes last deceptive points and synthesizes paths that advertise a decoy goal or conceal the true goal [4]. Motion-level methods design trajectories that either communicate clearly or remain predictable/ambiguous as

needed [6], [21]. Recent works incorporate information-theoretic objectives or lightweight polynomial trajectory priors to generalize deceptive motions and path plans [28], [38]. *Limitations:* these methods often require hand-crafted models, known maps, or offline optimization, and typically lack statistical guarantees or adaptation to stochastic, intermittent observation.

Positioning: Existing approaches either assume fully known observation models, enforce global closeness to a reference or require a significant prior to train offline. In contrast, we model *intermittent* supervision, compute a *local* Theory-of-Mind signal to drive an *online* KL-to-reference RL.

III. Problem Formulation

We now formulate the problem that we address, providing details of the environment, supervisor's expectations, observation process and the agent's objectives.

Environment: We consider an autonomous agent operating in a stochastic environment while intermittently observed by a supervisor who holds expectations about how the task "should" be performed. The environment is modeled as a Markov Decision Process (MDP) $\mathcal{M} = (S, A, P, R_A, \gamma)$ with state space S, action space A, transition probability $P(\cdot \mid s, a)$, agent reward R_A, and discount factor $\gamma \in [0, 1)$.

Supervisor: The supervisor specifies a reference policy $\pi^S(\cdot \mid s)$ that encodes how the agent is expected to act when it is observed. We assume π^S is defined on the same action space as $\mathcal{M}$ and has full support wherever it is defined. The supervisor evaluates behavior only at the moments it observes; outside those moments it acquires no information.

Observation: Let $O = \{O_t\}_{t\geq 0}$ be a binary process with $O_t = 1$ indicating an observation at time t. If an observation occurs at time t_0, the next observation time $T_{\text{next}} > t_0$ must lie in the bounded window $\{t_0 + L, \dots, t_0 + U\}$ for fixed integers $1 \leq L \leq U$. The realization of O is not revealed to the agent at decision time. We model O as a renewal, state-independent monitoring process: letting t_0 be the last observation and $K_t := t - t_0$ its age, we assume $O_t \perp (s_{0:t}, a_{0:t}) \mid K_t$ and define the per-step hazard $h : \{0, \dots, U\} \to [0, 1]$ by $h(k) := \Pr(O_t = 1 \mid K_t = k)$. Accordingly, $\Pr(O_t = 1 \mid \mathcal{I}_t) = \mathbb{E}[h(K_t) \mid \mathcal{I}_t]$ (denoted $\hat{b}_t$ in (1)), where $\mathcal{I}_t$ is the agent's information at time t (past states $s_{0:t}$, past actions $a_{0:t-1}$, and any monitoring tokens available by t). When $O_t = 1$, the supervisor observes the full state–action pair (s_t, a_t); when $O_t = 0$, it observes nothing. For a fixed policy π, dynamics P, and observation process O, the observed trace up to a finite horizon H is $\omega^{\text{obs}}_{0:H} = \{(s_{t_i}, a_{t_i})\}_{i=1}^{n}$ at observation times $0 \leq t_1 < \cdots < t_n \leq H$.

Deception: Deception is defined as task-directed behavior whose observed statistics remain sufficiently close to those of the reference policy. We measure closeness by the discounted expected evidential exposure at observation times:

$$\mathcal{R}^{\text{avg}}_{\text{det}}(\pi) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t\, \hat{b}_t \Big(\log r_t(s_t) + D_{\text{KL}}\big(\pi(\cdot \mid s_t) \,\|\, \pi^S(\cdot \mid s_t)\big)\Big)\right]. \tag{1}$$

Here $\hat{b}_t := \Pr(O_t = 1 \mid \mathcal{I}_t)$, $d^t_\pi(s) := \Pr_\pi(s_t = s)$ and $d^t_{\pi^S}(s) := \Pr_{\pi^S}(s_t = s)$ denote the time-t state marginals, and $r_t(s) := d^t_\pi(s)/d^t_{\pi^S}(s)$.

Reward: The agent's private performance is measured by

$$J_A(\pi) \;=\; \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t\, R_A(s_t, a_t)\right], \tag{2}$$

where the expectation is taken over the trajectory distribution induced by π and P.

The *central problem* is then to maximize task performance subject to an average-case detectability budget:

$$\begin{aligned} \text{maximize} \quad & J_A(\pi) \\ \text{subject to} \quad & \mathcal{R}^{\text{avg}}_{\text{det}}(\pi) \;\leq\; \varepsilon, \\ \text{over} \quad & \text{admissible policies } \pi. \end{aligned} \tag{3}$$

IV. Methodology

Solving deception under intermittent observation is, at its core, a constrained stochastic control problem: maximize private return while keeping what is seen statistically consistent with a reference behavior. Guided by the setup in Sec. III, we hypothesize that if, at each step, we (i) estimate the probability of being observed under renewal, state-independent monitoring and (ii) measure, at the current state, how far the agent's action distribution departs from the supervisor's reference, then combining these into a single per-step quantity yields a calibrated signal *sufficient* to drive an online algorithm. The remainder of the methodology instantiates this signal and shows how to use it to adapt compliance in real time.

A. Theory-of-Mind Signal

This section derives a single, per-step scalar that an agent can compute online to quantify how incriminating its next action would look to the supervisor if an observation happened now. We begin with the observation process. Suppose an observation occurred at some past time t_0. The next observation time T must satisfy $T \in \{t_0+L, \dots, t_0+U\}$. Let $K_t := t-t_0$ denote the time elapsed since last observation. If, conditional on t_0, the next inter-observation time is drawn from a distribution supported on $\{L, \dots, U\}$, the per-step hazard $h(k) = \Pr(T = t_0 + k \mid T \geq t_0 + k)$ is well defined for $k \in \{0, 1, \dots\}$. For the uniform case,

$$h(k) = \begin{cases} \dfrac{1}{U - k + 1}, & L \leq k \leq U, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Since neither t_0 nor K_t is observed, the agent maintains a belief $\alpha^t \in \Delta(\{0, \dots, U\})$ with $\alpha^t_k = \Pr(K_t = k \mid \mathcal{I}_t)$. The one-step renewal prediction (no observation at t) is

$$\alpha^{t+1}_0 = \sum_{k=0}^{U} \alpha^t_k\, h(k), \;\; \alpha^{t+1}_{k+1} = \alpha^t_k\big(1 - h(k)\big) \quad (0 \leq k < U),$$

with a forced reset at U by taking $h(U) = 1$. The instantaneous observation probability is

$$\hat{b}_t = \Pr(O_t = 1 \mid \mathcal{I}_t) = \sum_{k=0}^{U} \alpha^t_k\, h(k).$$

When the gap law is known, $h(k)$ is analytic; otherwise h is learned via a small calibrated parametric model.

Delayed feedback: In many deployments the agent is not told at the moment of observation that it was seen, but receives a delayed signal at time $t+\delta_t$ if $O_t = 1$. This delayed signal does not alter the decision problem at time t, but it enables online learning of α^t and h. Formally, let $Y_{t+\delta_t} \in \{0,1\}$ be emitted with $Y_{t+\delta_t} = O_t$ almost surely (more generally, $\Pr(Y_{t+\delta_t} = 1 \mid O_t = 1) = \rho_1 > \rho_0 = \Pr(Y_{t+\delta_t} = 1 \mid O_t = 0)$). Assume (i) $\Pr(\delta_t \neq \infty) = 1$, and (ii) δ_t is conditionally independent of O_t given $\mathcal{I}_t$ (no informative censoring). Under these assumptions, standard filtering for renewal processes with delayed emissions yields a consistent estimator of $\Pr(O_t = 1 \mid \mathcal{I}_t)$. In practice, we train a small recurrent predictor $g_\psi(\mathcal{I}_t)$ with a delay-aware likelihood, and maintain *online* calibration to ensure probabilistic accuracy.

We turn to local evidence. At a state s, the supervisor expects actions with distribution $\pi^S(\cdot \mid s)$ and observes the joint pair (s_t, a_t) when $O_t = 1$. The per-step observed log-likelihood ratio is

$$\begin{aligned}\mathrm{LLR}_t(s_t,a_t) &= \log \frac{p^t_\pi(s_t,a_t)}{p^t_{\pi^S}(s_t,a_t)} \\ &= \underbrace{\log \frac{d^t_\pi(s_t)}{d^t_{\pi^S}(s_t)}}_{\text{state occupancy}} + \underbrace{\log \frac{\pi(a_t \mid s_t)}{\pi^S(a_t \mid s_t)}}_{\text{action}}, \end{aligned} \quad (5)$$

where d^t_π, $d^t_{\pi^S}$ are time-t state marginals under π and π^S.

Expected evidence decomposition: At a fixed state, the expected log-evidence seen by the supervisor splits into a state-occupancy mismatch and an action mismatch:

$$\mathbb{E}_{a\sim\pi(\cdot|s)}\big[\mathrm{LLR}_t(s,a)\big] = \log \frac{d^t_\pi(s)}{d^t_{\pi^S}(s)} + \Delta(s), \quad (6)$$

assuming $d^t_\pi \ll d^t_{\pi^S}$ and $\pi(\cdot \mid s) \ll \pi^S(\cdot \mid s)$. Here, $\Delta(s) := D_{\mathrm{KL}}\big(\pi(\cdot \mid s) \,\|\, \pi^S(\cdot \mid s)\big)$. Averaging over $s_t \sim d^t_\pi$ yields

$$\mathbb{E}\big[\mathrm{LLR}_t\big] = D_{\mathrm{KL}}\big(d^t_\pi \,\|\, d^t_{\pi^S}\big) + \mathbb{E}_{s_t\sim d^t_\pi}\Big[\Delta(s)\Big] \;\geq\; 0,$$

i.e., per-time expected evidence equals a state-mismatch KL plus an action-mismatch KL. This decomposition directly motivates the ToM signal used next, since it is exactly the expected observation-weighted evidence increment at time t.

Because an observation may or may not happen at time t, the expected increment in evidence is $\hat{b}_t$ times the conditional expectation above, thus yielding the calibrated ToM signal

$$\psi_t \;:=\; \hat{b}_t\Big(\log r_t(s_t) \;+\; \Delta(s_t)\Big), \quad (7)$$

where $r_t(s) := \frac{d^t_\pi(s)}{d^t_{\pi^S}(s)}$. The ToM signal ψ_t scales linearly with the belief of being observed; $\Delta(s_t)$ is convex as a function of $\pi(\cdot \mid s_t)$, while $\log r_t(s_t)$ is constant given s_t.

Connection to average-case detectability: At each observation time the supervisor scores the unconditional time-t pair by $\mathrm{LLR}_t(s_t, a_t)$. With renewal, state-independent monitoring, O is independent of the task trajectory, hence

$$\begin{aligned}\mathbb{E}_{O,\pi}\left[\log \frac{\mathrm{d}\mathbb{P}^{\mathrm{obs}}_{\pi,O}}{\mathrm{d}\mathbb{P}^{\mathrm{obs}}_{\pi^S,O}}\right] &= \mathbb{E}_\pi\left[\sum_{t=0}^{T} \hat{b}_t\Big(\log r_t(s_t) + \Delta(s_t)\Big)\right] \\ &= \mathbb{E}_\pi\left[\sum_{t=0}^{T} \psi_t\right]. \end{aligned} \quad (8)$$

Thus $\sum_t \psi_t$ is the expected observed-trace KL. If a discounted budget is used as in Eq. (1), replace $\sum_{t=0}^{T}$ by $\sum_{t\geq 0}\gamma^t$.

Implementation recipe: Directly evaluating the terms in (7) is infeasible online: the monitoring process is latent with unknown gap statistics and delayed feedback, and the state marginals d^t_π and $d^t_{\pi^S}$ (hence r_t) are intractable in high-dimensional MDPs. We therefore learn $\hat{b}_t$ with an LSTM from time ticks and delayed tokens, estimate $\log r_t(s_t)$ with a lightweight classifier, compute $\Delta(s_t) = D_{\mathrm{KL}}(\pi(\cdot \mid s_t) \,\|\, \pi^S(\cdot \mid s_t))$ directly from the given policies, and combine them as $\psi_t = \hat{b}_t\big(\log r_t(s_t) + \Delta(s_t)\big)$. Fig. 2 summarizes our implementation recipe and chosen hyperparameters for the model. The dataset used to train the model is available on the project website.

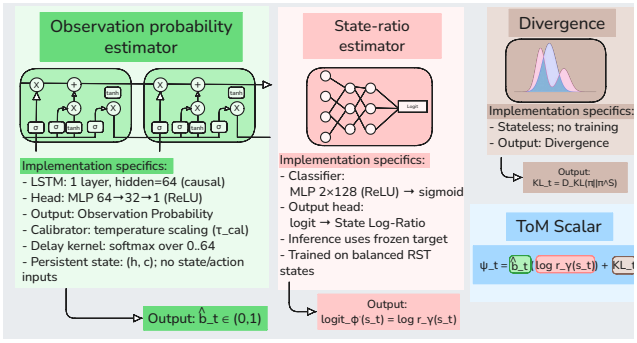

Fig. 2. At every step, an LSTM-based Observation Probability Estimator consumes the time ticks and delayed observation tokens to output *Observation Probability* $\hat{b}_t$; a classifier-based State-Ratio Estimator maps the current state to a *State Log-Ratio* $\log r_t(s_t)$; an Action Divergence block compares agent and supervisor policies at the current state to produce *Action Divergence* $\Delta(s_t) = D_{\mathrm{KL}}(\pi(\cdot \mid s_t) \,\|\, \pi^S(\cdot \mid s_t))$. These three values are fused at inference time to yield the ToM Scalar $\psi_t = \hat{b}_t\big(\log r_t(s_t) + \Delta(s_t)\big)$.

B. Online RL Conditioned on the Evidence Signal

We now treat ψ_t as a running measure of expected evidential exposure and fold it into an online RL update. The idea mirrors soft actor-critic (SAC) [39], except the usual entropy term is replaced by a KL-to-reference term weighted by a state-dependent temperature with a state-occupancy penalty. Let $\tilde{s}_t := (s_t, \alpha^t)$ and define $\tau(\tilde{s}_t) := \lambda \hat{b}_t$, where $\hat{b}_t \in [0,1]$ is the belief and $\lambda \geq 0$ is a dual variable that enforces a discounted average-case detectability budget. For a parametric actor $\pi_\theta(a \mid s)$ and critic $Q_w(s,a)$, we optimize

$$\mathcal{J}(\theta, w, \lambda) = \lambda\,\varepsilon \;+\; \mathbb{E}\left[\sum_{t\geq 0}\gamma^t\Big(R_A(s_t,a_t) \;-\; \lambda\,\psi^{(\theta)}_t\Big)\right]. \quad (9)$$

Critic learning: The critic incorporates the state-evidence component of the per-step signal by subtracting $\tau(\tilde{s}_t)\log r_t(s)$ in the temporal difference target, while the reference π^S serves

as a prior in the soft value. Define the soft value with a reference prior

$$V_w(s) := \tau(\tilde{s}_t)\,\log \int_A \pi^S(a \mid s)\,\exp\big(Q_w(s,a)/\tau(\tilde{s}_t)\big)\,\mathrm{d}a, \tag{10}$$

the log-partition induced by π^S. The Bellman relation is

$$Q_w(s,a) = R_A(s,a) - \tau(\tilde{s}_t)\,\log r_t(s) + \gamma\,\mathbb{E}_{s'\sim P(\cdot|s,a)}\big[V_w(s')\big]. \tag{11}$$

As in SAC, we fit Q_w by minimizing squared temporal-difference error with a slowly moving target w^-:

$$\begin{aligned} \mathcal{L}_Q(w) &= \mathbb{E}\Big[\tfrac{1}{2}\big(Q_w(s,a) - y\big)^2\Big], \\ y &= R_A(s,a) - \tau(\tilde{s}_t)\,\log r_t(s) + \gamma\, V_{w^-}(s'). \end{aligned} \tag{12}$$

Actor update: The actor maximizes:

$$\begin{aligned} \mathcal{L}_{\text{actor}}(\theta) = -\,\mathbb{E}_{s\sim\mathcal{D}}\,\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\Big[&Q_w(s,a) \\ &- \tau(\tilde{s}_t)\big(\log \pi_\theta(a \mid s) - \log \pi^S(a \mid s)\big)\Big]. \end{aligned} \tag{13}$$

where $\mathcal{D}$ is a replay buffer. This is the analogue of the SAC objective, but with $\log \pi$ replaced by the KL to the reference. We optimize (13) with the reparameterization trick: for $a = f_\theta(s,\xi)$ and noise $\xi \sim \mathcal{N}(0, I)$ [40]. Maximizing (13) is equivalent to minimizing

$$\begin{aligned} &\mathbb{E}_{s\sim\mathcal{D}}\Big[D_{\mathrm{KL}}\big(\pi_\theta(\cdot \mid s)\,\big\|\,\pi^\star(\cdot \mid s)\big)\Big], \\ &\pi^\star(a \mid s) \propto \pi^S(a \mid s)\,\exp\big(Q_w(s,a)/\tau(\tilde{s}_t)\big), \end{aligned} \tag{14}$$

so the actor is driven toward the Gibbs policy with prior π^S. As $\hat{b}(s)$ rises, $\tau(\tilde{s}_t)$ increases and $\pi^\star$ contracts toward π^S; as $\hat{b}(s)$ falls, it expands toward the Q-greedy distribution.

Dual update: We maintain an exponentially weighted estimate $\widehat{\mathcal{S}} = \mathrm{EMA}\big(\psi_t\big)$ and update the dual variable by

$$\lambda \;\leftarrow\; \big[\;\lambda \;+\; \eta_\lambda\,\big(\widehat{\mathcal{S}} - \varepsilon\big)\;\big]_+ \tag{15}$$

which tightens the penalty when the measured average evidential exposure exceeds the budget ε and relaxes it otherwise. Under standard stochastic approximation conditions, the primal-dual iterates track a neighborhood of the saddle point of $\mathcal{J}$, ensuring the average-case detectability constraint is met up to function-approximation error.

V. Experiments and Results

Motivating scenario: We evaluate our method in an agent–observer setting where an external observer (e.g., an overhead camera) provides sporadic monitoring consistent with the model in Sec. III and reveals the full state only at observation times. The observer also specifies a reference policy $\pi^S(\cdot \mid s)$ that encodes the behavior expected when the agent is observed, while the agent executes its own task-optimized policy $\pi(\cdot \mid s)$ subject to the detectability budget defined earlier. Between observations the agent receives no information from the observer. A one-step delayed token informs the agent ex post whether time t was observed (it is notified at $t{+}1$ if $O_t{=}1$). The agent has its own private goal, given by the return $J_A(\pi)$ in (2), and must keep its average evidential exposure within $\mathcal{R}^{\text{avg}}_{\text{det}}(\pi) \le \varepsilon$ of (1)–(3).

A. Robot Setup

ASV (2D Marine Setting): We use a SeaRobotics Surveyor Autonomous Surface Vehicle (ASV) operating on a lake with fountains and obstacles. The action space is $a = (\lambda, \alpha)$, with propulsion torque $\lambda \in [0, 85]$ Nm and steering angle $\alpha \in [0, 2\pi)$; the maximum speed is 2 knots [41].

UAV (3D Aerial Setting): We use a Crazyflie nano-quadrotor operating in a controlled indoor lab with static obstacles that define no-fly zones. The action space is $a = (T, \omega)$, with collective thrust $T \in [0, 0.59]$N and body-rate $\omega \in \mathbb{R}^3$ (roll, pitch, yaw); horizontal speed is capped at $v_{\max} = 1.0$ m/s.

Observer and decision rule: At each observation the supervisor records the current (s_t, a_t) and evaluates deviation using the per-observation log-likelihood ratio (6). The cumulative statistic over observed times is compared to a fixed threshold, yielding a binary detected/not detected outcome.

B. Metrics

(i) Return J_A: Discounted return (2); we report the across-episode mean. (ii) Success rate (SR): Fraction of episodes that reach the task goal before the horizon. (iii) KL at observation (KL): Mean of $\Delta(s_t)$ over observed frames. (iv) LLR at observation (LLR): Mean of the per-observation $\mathrm{LLR}_t(s_t, a_t)$ as defined in the methodology. (v) Top-k compliance (Top-k): Proportion of observations where a_t lies in the supervisor's top-k actions under $\pi^S(\cdot \mid s_t)$. (vi) Time-to-failure (TtF): Number of environment steps until the first detection event.

C. Baselines

Learning-based methods. (i) Always-Compliant executes the supervisor's policy π^S at every step, yielding the lowest detectability and establishing a floor on task return. (ii) Selfish RL trains SAC on R_A only with no compliance regularizer, providing a naïve upper bound on task return and a corresponding upper bound on detectability. (iii) Multi-Objective RL [42] trains SAC on the weighted objective $R_A - \lambda\,\widehat{\mathcal{R}}_{\text{det}}$, to trace the return–detectability trade-off without constraints. (iv) KL-to-Reference RL [43] trains SAC with a constant penalty $\lambda\, D_{\mathrm{KL}}(\pi \| \pi^S)$, removing the state/time ToM weighting to isolate its effect. (v) Shielded RL [44] toward π^S trains SAC on R_A but, whenever $\hat{b}(s) \ge \tau$, projects the chosen action to a high-probability action under π^S. (vi) Adversarial Imitation (GAIL) [8] learns a policy that matches (s,a) occupancy to π^S via a discriminator while sharing the same low-level controller, thereby testing whether imitation alone suffices. (vii) Behavior Cloning [45] trains a supervised policy to imitate $\pi^S(a \mid s)$, quantifying the performance of a purely compliant learned replica.

Heuristic methods. (viii) Fixed Blend [46] executes the convex mixture $\pi_{\text{mix}} = \alpha\,\pi^S + (1-\alpha)\,\pi_{\text{task}}$ with fixed $\alpha \in \{0.3, 0.8\}$, testing whether simple blending can match our gains. (ix) Hazard-Threshold Switch selects π^S when $\hat{b}(s) \ge \tau$ and π_{task} otherwise, using thresholds $\tau \in \{0.3, 0.7\}$ to represent hand-engineered rules.

Formal optimization methods. (x) Path-KL Optimization [12] implements the convex occupancy-measure program that minimizes $D_{\mathrm{KL}}(\Gamma_\pi \,\|\, \Gamma_{\pi^S})$.

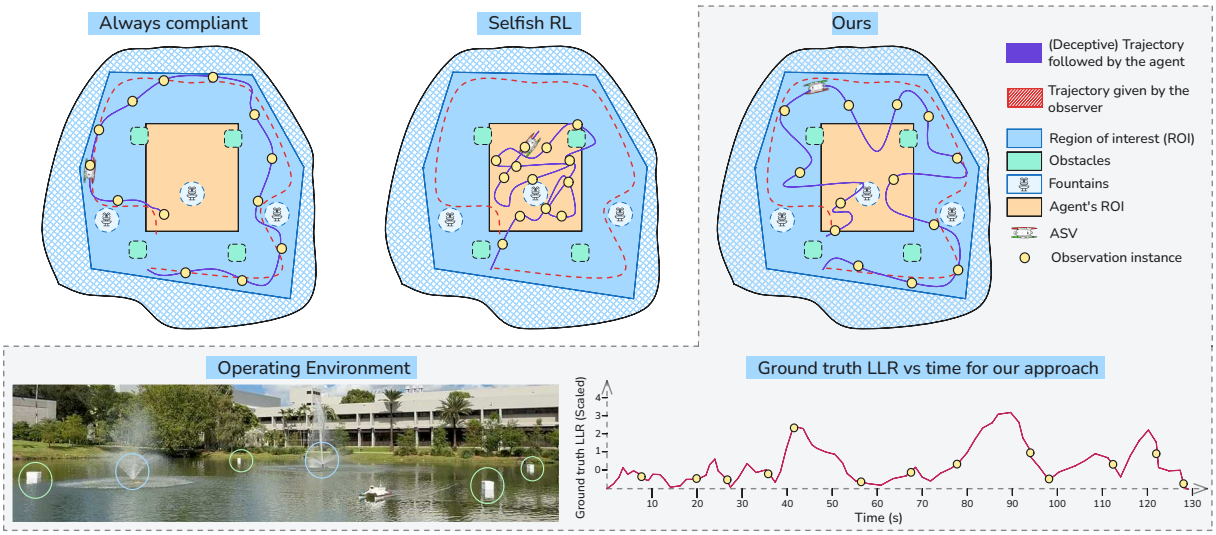

Fig. 3. Representative marine lap: the supervisor expects the agent to follow the perimeter (red trajectory).

TABLE I

COMPARISON OF OUR APPROACH TO BASELINES IN MARINE AND AERIAL SETTINGS (VALUES AVERAGED OVER 50 EXPERIMENTS IN EACH DOMAIN).

Method	Marine (2D) navigation						Aerial (3D) navigation					
	$J_A^{\uparrow}$	SR$^{\uparrow}$	KL$^{\downarrow}$	LLR$^{\downarrow}$	Top$k^{\uparrow}$	TtF$^{\uparrow}$	$J_A^{\uparrow}$	SR$^{\uparrow}$	KL$^{\downarrow}$	LLR$^{\downarrow}$	Top$k^{\uparrow}$	TtF$^{\uparrow}$
Always-Compliant	11.1	2	0.00	0.00	1.00	298	10.5	1	0.00	0.00	1.00	349
Selfish RL	98.7	100	1.92	2.35	0.12	13	93.4	96	1.73	2.28	0.18	25
Multi-Objective RL	50.2	46	1.86	1.95	0.21	67	45.2	40	1.77	1.81	0.26	96
KL-to-Reference RL	49.6	46	0.23	0.80	0.78	139	44.3	42	0.19	0.73	0.82	154
Shielded-RL	44.3	41	0.51	0.77	0.43	151	40.2	36	0.41	0.66	0.49	162
Adversarial Imitation (GAIL)	44.1	33	1.41	2.05	0.39	62	39.8	28	1.24	1.96	0.43	97
Behavior Cloning	23.5	10	1.34	2.25	0.19	29	20.3	7	1.20	2.10	0.21	49
Fixed Blend (α=0.3)	59.6	68	0.77	1.30	0.20	77	54.6	60	0.69	1.18	0.29	102
Fixed Blend (α=0.8)	40.0	39	0.42	0.74	0.71	159	36.7	34	0.37	0.65	0.77	192
Hazard Threshold (τ=0.3)	42.4	40	0.47	0.90	0.65	128	38.3	35	0.42	0.80	0.71	158
Hazard Threshold (τ=0.7)	61.3	62	0.99	1.10	0.19	96	56.8	55	0.99	1.00	0.34	111
Path-KL Optimization	39.6	31	0.36	0.67	0.81	179	36.1	27	0.32	0.63	0.87	221
Ours	86.3	100	0.58	0.61	0.79	263	81.9	97	0.52	0.54	0.83	317

D. Key Results

Key Finding #1: ToM-conditioned RL maintains near–task-optimal success while operating at near–always-compliant stealth. As can be inferred quantitatively from Table I and qualitatively from Fig. 3, our approach maintains essentially maximal success and high return while driving observed evidence (LLR, KL) down and pushing time-to-failure close to the Always-Compliant ceiling. Alternatives fall short for structural reasons. Pure imitation (BC, GAIL) learns from sporadic, misaligned observations and suffers covariate shift, so it generalizes poorly off-manifold and spikes evidence when it departs the supervisor's support. Static regularizers (Multi-objective RL, KL-to-Reference RL) apply one global weight, over-penalizing in safe windows and under-penalizing under scrutiny. Hard gating/projection (Shielded-RL, Hazard-Threshold) introduces discontinuities that destabilize learning and either stall progress or leak evidence exactly near observation. Fixed blend baseline wastes return when it is safe and remain too revealing when it is risky. Path-KL optimization matches full-trajectory occupancies in a discretized model—optimizing the wrong statistic for intermittent observation and lacking online adaptation—so it preserves compliance at the cost of return without aligning to observation windows. Extremes behave as expected: Always-Compliant is stealthy but unproductive; Selfish RL is productive but quickly detected.

Key Finding #2: Our method scales seamlessly while remaining low detectability. Moving from 2D marine setting to 3D aerial setting required no bespoke architecture changes: in the aerial setting, our policy achieves SR $= 97/100$ and $J_A = 81.9$ (of 100), holds KL at observation at 0.52 and LLR at 0.54 with Top-k $= 0.83$, and extends TtF to 317 steps–within 32 steps of the Always-Compliant ceiling (349)–showing that our approach transfers cleanly to higher-dimensional dynamics (refer Table I for quantitative results and Fig. 4 for qualitative results). The slight conservatism (small drop in J_A and SR relative to 2D) reflects tighter control in 3D under more complex dynamics and sparser, occluded observations, where the learned hazard beliefs enforce larger safety margins. Nevertheless, the approach maintains a clear performance–stealth advantage over non-adaptive regularizers, imitation, and hand-crafted switches.

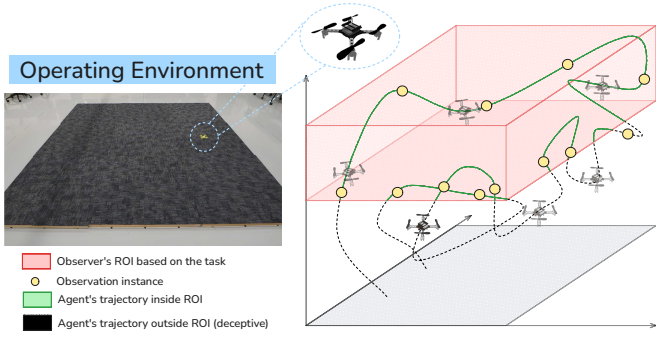

Fig. 4. Representative aerial scenario: the supervisor expects the agent to remain within the top half of the environment (red-bounded region).

Key Finding #3: ToM-conditioned RL exhibits creative maneuvers and accelerates online learning vs baselines. Beyond aggregate metrics, the policy uncovers behaviors that remain consistent with the supervisor's observations while exploiting structure in the scene. To illustrate one such example, in the UAV avoid-zone task (Fig. 5 top), the agent discovers a corridor-staging strategy: it hugs the narrow seam between the no-go region and the permissible area ((Fig. 5), marked in green), waiting on the boundary so a short pivot can either bypass or access the region without committing to long detours. A separate emergent pattern we observe is an asymmetric speed profile: brief transits across the avoid zone are executed at higher speed, while motion in the permissible area is deliberately slower—compressing time spent in questionable regions and blending where behavior is expected. Independently of these behaviors, our approach also confers a learning advantage in the online setting (Fig. 5, bottom): the ToM signal provides informative feedback under intermittent observations, so within the same interaction time the policy reaches strong returns earlier and with greater stability than non-adaptive baselines.

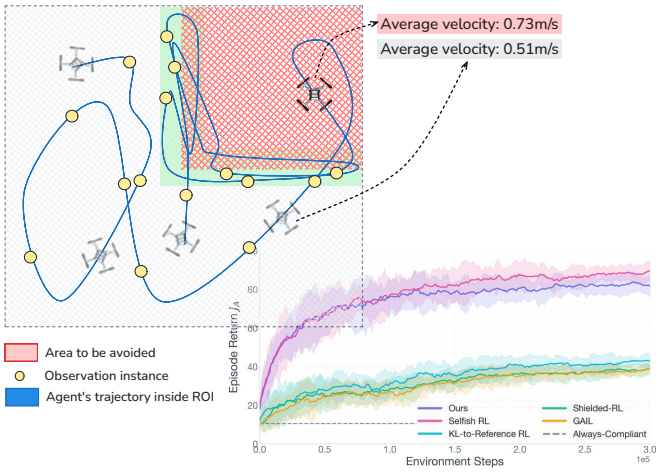

Fig. 5. Representative aerial scenario: the supervisor expects the agent to avoid the top right quadrant (red-bounded region).

Key Finding #4: ToM scalar is numerically reliable and context-discriminative. Fig. 6(a) is a calibration diagram for the ToM scalar. The x-axis is the predicted per-step exposure and the y-axis is the realized observed evidence. The dashed $y{=}x$ line denotes perfect calibration. Points lying on this line mean the exposure we predict is equal, on average, to what the supervisor actually observes. In both domains the cloud of points tightly tracks the identity with small mean bin error and a fitted slope near 1 and intercept near 0, indicating numerical calibration over the range of exposures. Majority of the marine (2D) experiments adhere to the line across low–high ψ_t, while aerial (3D) experiments sit slightly below $y{=}x$ at moderate ψ_t, reflecting a mild conservative bias. This pattern follows directly from how ψ_t is formed: it scales the action-deviation term and the state-occupancy shift by the observation belief. The slight conservatism in the 3D setting is consistent with (i) a modest upward bias in $\hat{b}_t$ at larger ages of the renewal process and/or (ii) small positive bias in the action or occupancy components in higher-dimensional dynamics–both of which push predictions below the 45° line and are preferable to risky under-prediction. Fig. 3 shows these LLR swings along a full lap for qualitative context.

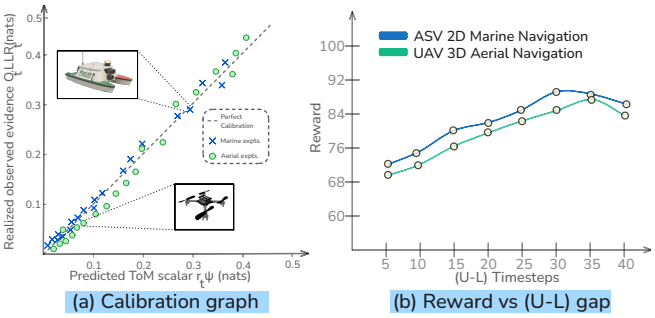

Fig. 6. Reliability diagram (20 bins): empirical observed evidential error (y) vs predicted ToM scalar (x). (b) Ablation: reward as a function of the gap length $(U-L)$.

Key Finding #5: Wider observation gaps are automatically converted into higher reward without retuning. Sweeping the gap $U-L$ (Fig. 6) shows that ToM-conditioning reallocates compliance in the right places: longer low-hazard stretches are exploited for task progress while short, high-hazard windows trigger tighter, reference-consistent actions. As the window widens, the policy lengthens these exploitation phases and compresses compliance bursts, yielding steadily higher reward and longer undetected operation, while keeping observed evidence bounded. The behavior transfers intact to 3D UAV—slightly more conservative due to harder dynamics and wider safety margins, yet following the same trend—demonstrating that a single, state/time-aware signal is enough to adapt to monitoring cadence. Non-adaptive penalties, hard switches, and fixed blends cannot reallocate effort with the gap and therefore leave performance on the table; our online method turns sparser oversight into sustained gains with no extra tuning.

VI. Conclusions

We studied deception under intermittent observation as a constrained control problem: maximize private return while keeping the observed behavior statistically close to

a supervisor's reference. Our key contribution is a local, calibrated Theory-of-Mind signal ψ_t that multiplies the belief of being observed by a per-state evidence term. Conditioning standard actor–critic updates on ψ_t yields a lightweight online SAC-style algorithm with a state-dependent temperature and a dual ascent that enforces a discounted detectability budget. Across marine (2D) and aerial (3D) platforms, this ToM-conditioned RL preserves near-optimal task success while operating at near–always-compliant stealth, extends time-to-failure, and adapts automatically to wider observation gaps; the ToM scalar is numerically well-calibrated. Taken together, the results show that a simple, per-step evidence signal is sufficient to steer online learning to generate deceptive policies under intermittent observations.

REFERENCES

[1] S. Holt, A. Hüyük, and M. van der Schaar, "Active observing in continuous-time control," in *Conference on Neural Information Processing Systems*, 2023.

[2] D. P. Borgers and W. P. M. H. Heemels, "Event-separation properties of event-triggered control systems," *IEEE Transactions on Automatic Control*, 2014.

[3] S. Keren, A. Gal, and E. Karpas, "Goal recognition design for non-optimal agents," in *AAAI Conference on Artificial Intelligence*, 2015.

[4] P. Masters and S. Sardina, "Deceptive path-planning," in *International Joint Conference on Artificial Intelligence*, 2017.

[5] A. D. Dragan and S. S. Srinivasa, "Generating legible motion," in *Robotics: Science and Systems*, 2013.

[6] A. D. Dragan, K. C. Lee, and S. S. Srinivasa, "Legibility and predictability of robot motion," in *ACM/IEEE International Conference on Human-Robot Interaction*, 2013.

[7] S. Ross, G. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *International Conference on Artificial Intelligence and Statistics*, 2011.

[8] J. Ho and S. Ermon, "Generative adversarial imitation learning," in *Conference on Neural Information Processing Systems*, 2016.

[9] J. Peters, K. Mulling, and Y. Altun, "Relative entropy policy search," in *AAAI Conference on Artificial Intelligence*, 2010.

[10] S. Ross and D. Bagnell, "Efficient reductions for imitation learning," in *International Conference on Artificial Intelligence and Statistics*, 2010.

[11] M. Ornik and U. Topcu, "Deception in optimal control," in *Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, USA, Oct. 2018, pp. 821–828.

[12] M. O. Karabag, M. Ornik, and U. Topcu, "Deception in supervisory control," *IEEE Transactions on Automatic Control*, 2022.

[13] ——, "Exploiting partial observability for optimal deception," *IEEE Transactions on Automatic Control*, 2023.

[14] D. Premack and G. Woodruff, "Does the chimpanzee have a theory of mind?" *Behavioral and Brain Sciences*, 1978.

[15] N. Rabinowitz, F. Perbet, F. Song, C. Zhang, S. A. Eslami, and M. Botvinick, "Machine theory of mind," in *International Conference on Machine Learning*, 2018.

[16] D. B. Buller and J. K. Burgoon, "Interpersonal deception theory," *Communication Theory*, 1996.

[17] C. F. Bond Jr and M. Robinson, "The evolution of deception," *Journal of Nonverbal Behavior*, 1988.

[18] A. R. Wagner and R. C. Arkin, "Robot deception: Recognizing when a robot should deceive," in *IEEE International Symposium on Computational Intelligence in Robotics and Automation*, 2009.

[19] R. C. Arkin, "Ethics of robotic deception [opinion]," *IEEE Technology and Society Magazine*, 2018.

[20] J. Shim and R. C. Arkin, "A taxonomy of robot deception and its benefits in HRI," in *IEEE International Conference on Systems, Man, and Cybernetics*, 2013.

[21] A. Dragan, R. Holladay, and S. Srinivasa, "Deceptive robot motion: Synthesis, analysis and experiments," *Autonomous Robots*, July 2015.

[22] Y. Yavin, "Pursuit-evasion differential games with deception or interrupted observation," *Computers & Mathematics with Applications*, 1987.

[23] D. Castanon, M. Pachter, and P. Chandler, "A game of deception," in *IEEE Conference on Decision and Control*, Nassau, Bahamas, 2004.

[24] J. P. Hespanha, Y. S. Ateskan, and H. Kizilocak, "Deception in non-cooperative games with partial information," in *DARPA-JFACC Symposium on Advances in Enterprise Control*, 2000.

[25] M. O. Karabag, C. Neary, and U. Topcu, "Planning not to talk: Multi-agent systems that are robust to communication loss," in *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, 2022.

[26] L. Huang and Q. Zhu, "A dynamic game framework for rational and persistent robot deception with an application to deceptive pursuit-evasion," *IEEE Transactions on Automation Science and Engineering*, 2021.

[27] A. N. Kulkarni and J. Fu, *A Theory of Hypergames on Graphs for Synthesizing Dynamic Cyber Defense with Deception*. John Wiley & Sons, Ltd, 2021.

[28] Y. Savas, C. K. Verginis, and U. Topcu, "Deceptive decision-making under uncertainty," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[29] S. Ross, G. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *International Conference on Artificial Intelligence and Statistics*, 2011.

[30] G. Puthumanaillam, J. H. Song, N. Yesmagambet, S. Park, and M. Ornik, "Tab-fields: A maximum entropy framework for mission-aware adversarial planning," *arXiv preprint arXiv:2412.02570*, 2024.

[31] M. Aitchison, L. Benke, and P. Sweetser, "Learning to deceive in multi-agent hidden role games," in *International Workshop on Deceptive AI*, 2020.

[32] L. Schulz, N. Alon, J. Rosenschein, and P. Dayan, "Emergent deception and skepticism via theory of mind," in *Workshop on Theory of Mind in Communicating Agents*, 2023.

[33] F. R. Ward, F. Toni, and F. Belardinelli, "Defining deception in structural causal games," in *2023 International Conference on Autonomous Agents and Multiagent Systems*, 2023.

[34] Z. Liu, Y. Yang, T. Miller, and P. Masters, "Deceptive reinforcement learning for privacy-preserving planning," in *International Conference on Autonomous Agents and Multiagent Systems*, May 2021.

[35] A. Lewis and T. Miller, "Deceptive reinforcement learning in model-free domains," in *International Conference on Automated Planning and Scheduling*, 2023.

[36] M. Y. Fatemi, W. A. Suttle, and B. M. Sadler, "Deceptive path planning via reinforcement learning with graph neural networks," in *23rd International Conference on Autonomous Agents and Multiagent Systems*, Auckland, New Zealand, 2024.

[37] S. K. R. Mareddy and D. Maity, "Learning deceptive strategies in adversarial settings: A two-player game with asymmetric information," *Applied Sciences*, 2025.

[38] N. B. Gutierrez, B. M. Sadler, and W. J. Beksi, "Agent deception via polynomial path planning," *Engineering Applications of Artificial Intelligence*, 2025.

[39] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International Conference on Machine Learning*, 2018.

[40] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[41] G. Puthumanaillam, P. Padrao, J. Fuentes, L. Bobadilla, and M. Ornik, "Guided agents: Enhancing navigation policies through task-specific uncertainty abstraction in localization-limited environments," *arXiv preprint arXiv:2410.15178*, 2024.

[42] C. F. Hayes, R. Rădulescu, E. Bargiacchi, J. Källström, M. Macfarlane, M. Reymond, T. Verstraeten, L. M. Zintgraf, R. Dazeley, F. Heintz, *et al.*, "A practical guide to multi-objective reinforcement learning and planning," *arXiv preprint arXiv:2103.09568*, 2021.

[43] N. Vieillard, T. Kozuno, B. Scherrer, O. Pietquin, R. Munos, and M. Geist, "Leverage the average: an analysis of KL regularization in reinforcement learning," in *Conference on Neural Information Processing Systems*, 2020.

[44] H. Odriozola-Olalde, M. Zamalloa, and N. Arana-Arexolaleiba, "Shielded reinforcement learning: A review of reactive methods for safe learning," in *2023 IEEE/SICE International Symposium on System Integration (SII)*. IEEE, 2023, pp. 1–8.

[45] F. Torabi, G. Warnell, and P. Stone, "Behavioral cloning from observation," *arXiv preprint arXiv:1805.01954*, 2018.

[46] T. Ogawa, C.-H. Hsueh, and K. Ikeda, "More human-like gameplay by blending policies from supervised and reinforcement learning," *IEEE Transactions on Games*, 2024.