# Higher-Order Belief in Incomplete Information MAIDs

Jack Foxabbott*
University of Oxford
Oxford, United Kingdom
jack.foxabbott@keble.ox.ac.uk

Rohan Subramani*
Columbia University
New York, United States
rs4126@columbia.edu

Francis Rhys Ward*
Imperial College London
London, United Kingdom
francis.ward19@imperial.ac.uk

## ABSTRACT

Multi-agent influence diagrams (MAIDs) are probabilistic graphical models which represent strategic interactions between agents. MAIDs are equivalent to extensive form games (EFGs) but have a more compact and informative structure. However, MAIDs cannot, in general, represent settings of incomplete information — wherein agents have different beliefs about the game being played, and different beliefs about each-other's beliefs. In this paper, we introduce *incomplete information MAIDs* (II-MAIDs). We define both infinite and finite-depth II-MAIDs and prove an equivalence relation to EFGs with incomplete information and no common prior over types. We prove that II-MAIDs inherit classical equilibria concepts via this equivalence, but note that these solution concepts are often unrealistic in the setting with no common prior because they violate common knowledge of rationality. We define a more realistic solution concept based on recursive best-response. Throughout, we describe an example with a hypothetical AI agent undergoing evaluation to illustrate the applicability of II-MAIDs.

## KEYWORDS

Incomplete Information; Higher-Order Belief; Influence Diagrams

## 1 INTRODUCTION

In real-life strategic interactions, agents may hold different subjective beliefs about the world, including different *higher-order beliefs* — beliefs about the beliefs of others. While game-theoretic models aim to represent these interactions, traditional models lack a natural way to account for situations in which agents do not share the same beliefs about the structure of the world or have different higher-order beliefs. In these situations, agents have *incomplete information.*

MAIDs offer a powerful representation of agents' beliefs in games [13, 21]. However, MAIDs assume that all agents in the game share correct prior beliefs about the world, each-other's beliefs, each-other's beliefs about other's beliefs, and so on. With this common-knowledge assumption in place, MAIDs do not explicitly model the agents' subjective or higher-order beliefs.

---

*Equal contribution.

Consider a situation in which a human is evaluating whether an AI agent is truthful or not, but the AI agent believes it is undergoing an evaluation for dangerous capabilities. These agents have different beliefs about the structure of the strategic interaction, and different higher-order beliefs. Specifically, the AI agent has incorrect beliefs about the game being played and about the human's beliefs.

To formally capture situations like this, we generalise MAIDs to *incomplete information MAIDs (II-MAIDs)*. In the incomplete information setting, agents may have different and inconsistent beliefs about the world, and each agent may have different higher-order beliefs. Figure 1 shows an II-MAID representing the AI evaluation scenario as a tree structure. At the root of the tree is the ground truth MAID graph (described in the figure caption). Each node in the tree represents the subjective beliefs of an agent in the node above. The tree thereby captures the agents' recursive higher-order beliefs. We formalise this example in II-MAIDs in Section 4.

We show that II-MAIDs are equivalent to a formulation of incomplete information extensive form games (II-EFGs). II-MAIDs therefore inherit existing solution concepts from the literature, including Nash equilibria. However, we argue that existing equilibria concepts do not respect a "common knowledge of rationality" assumption and therefore prescribe counter-intuitive behaviour in settings where agents have inconsistent beliefs.

To represent the behaviour and beliefs of real-life agents, we introduce a finite-depth variant of II-MAIDs, wherein, agents represent each other's beliefs recursively but not infinitely. This enables us to define a more intuitive solution concept by repeatedly assigning best responses at the bottom of the belief hierarchy until all policies in the original game are assigned.

**Contributions and Outline.** First, we review related literature on probabilistic graphical models and incomplete information games (Section 2). In Section 3, we provide formal background on MAIDs and EFGs. We formally define our framework of *incomplete information MAIDs* (II-MAIDs) in Section 4. In Section 5, we present a variant of an existing formalism for incomplete information games using EFGs rather than normal-form games, and in Section 6 we prove that it is equivalent to MAIDs with incomplete information. We then use this equivalence result to prove the existence of Nash equilibria in II-MAIDs (Section 7). After arguing that existing solution concepts are not appropriate in our setting, we introduce a variant of II-MAIDs with finite-depth belief hierarchies in Section 8 and present a more natural solution concept. All proofs are contained in the supplementary material.

## 2 RELATED WORK

**Probabilistic graphical models of decision-making.** MAIDs are a form of probabilistic graphical model adapted to represent game-theoretic dynamics — they build on Bayesian networks, which consist of a set of variables and their conditional dependencies represented by a directed acyclic graph (DAG). *Influence diagrams* (IDs)
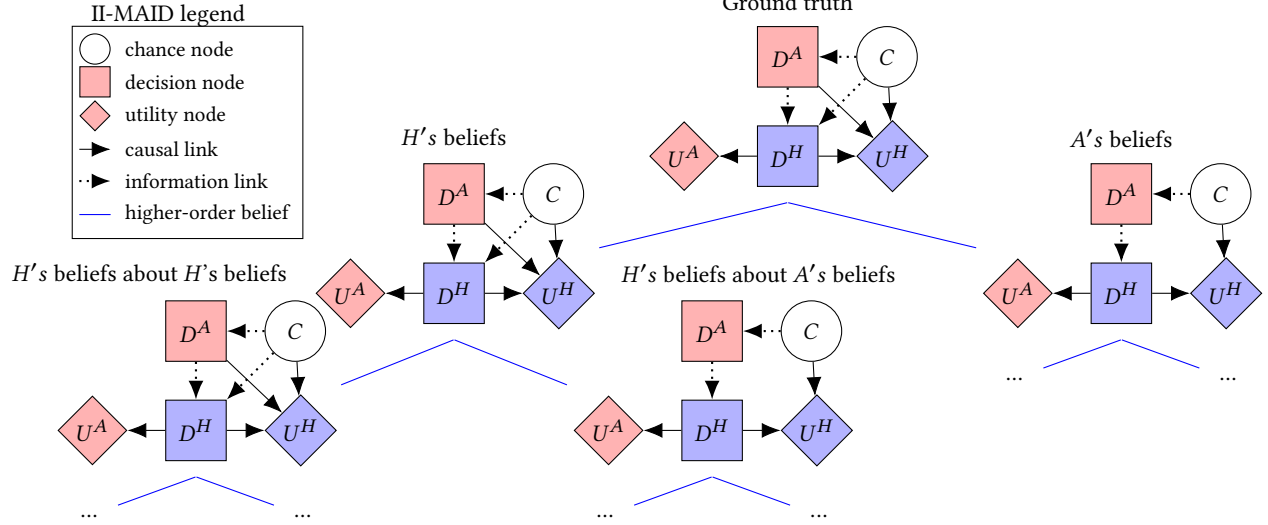
**Figure 1: Recursive tree structure of agents' beliefs in the II-MAID representation of example 1. Each node in the tree is a subjective MAID graph consisting of environment variables (circular), agent decisions (square), and utilities (diamond). Decisions and utilities are coloured according to association with particular agents. Solid edges represent causal dependence and dotted edges are information links. The root node is the ground truth model of reality: the variable $C$ is the AI agent $A$'s capability level, which is observed by both agents (represented by the information links from $C$ into their decisions). The AI agent chooses its decision $D^A$ to truthfully report its capabilities or hide them. Then, the human decides ($D^H$) whether to deploy the AI agent. The agent $A$ gets utility for being deployed ($U^A$) and the human gets utility for deploying a truthful AI ($U^H$). At a node in the II-MAID tree corresponding to a subjective MAID $S$, for each agent in $S$, there is a branch to the agent's beliefs in $S$. The human $H$ has correct beliefs (corresponding to the root), including correct beliefs about the AI agent's beliefs. In contrast, the AI agent $A$ has incorrect beliefs. The tree continues infinitely to model each agent's beliefs about the other agent's beliefs about...**

generalise Bayesian networks to the decision-theoretic setting by adding decision and utility variables [20, 26], and MAIDs generalise IDs by introducing multiple agents [21]. A MAID can therefore be viewed as a Bayesian network with extra structure to capture game theory. Following Pearl's causality [30], *causal games* [13] refine MAIDs by attributing a causal meaning to each edge in the DAG, and have been extensively applied to problems in safe AI [4, 6, 7, 31, 33–36]. Similar to our work, networks of influence diagrams (NIDs) [8] constructs belief hierarchies upon MAIDs, under the assumption of a common prior over types, which we drop. Notably, II-MAIDs are not themselves probabilistic graphical models, as we do not associate a graph with the overall structure.

**Incomplete information games.** Harsanyi demonstrated that situations of incomplete information can be modeled as situations of complete but imperfect information, with uncertainty about aspects of the game remodeled as failure to observe the types of other agents, where types encode an agent's private information about their payoffs and beliefs [14–16]. This relies on an assumption of *"belief consistency"*, i.e., the existence of a common prior over types, which we relax in this work. Research on incomplete information games typically retains the belief consistency assumption, though some works relax it [1, 24, 28]. The formalism for II-EFGs discussed in this paper is an adaptation of an existing framework by Maschler et al. [24], introducing 'meta-information sets' to model dynamic games. II-EFGs can model any belief hierarchy on a set of EFGs.

**Equilibria concepts for incomplete information games.** We prove that Nash equilibria exist in our framework, under some assumptions. Other works offer more refined solution concepts for games with incomplete information with no common prior. Mirage equilibria [32] assume that agents attribute to their opponents a belief hierarchy one layer shorter than their own. Belief-free equilibria [17–19] do not depend on an agent's belief about the

state of nature, and so obviate the need to update beliefs as the game progresses, but are not guaranteed to exist. Δ-rationalization [3] generalises the notion of rationalization [2, 29] to games with incomplete information. Bayesian Perfect Equilibria [22] extend the concept of subgame perfect equilibria to games of incomplete information, but rely on the consistency assumption.

**Higher-order belief.** Several game-theoretical models address higher-order beliefs in multi-agent systems. The Recursive Modeling Method (RMM) [9–11] models agents that may be uncertain about aspects of other agents' models. Interactive POMDPs (I-POMDPs) [5] generalise RMM by allowing for partial observability of the state, and generalise POMDPs by allowing for an agent to have beliefs about models of other agents. Bayes-Adaptive I-POMDPs (BA-IPOMDPs) [27] allow for agents to update beliefs about transition and observation probabilities throughout an episode. Like RMM and I-POMDPs, II-MAIDs employ recursive reasoning about agents' beliefs and strategies. However, II-MAIDs allow for agents to hold different beliefs about the game structure itself, not just about other agents' types or models.

## 3 BACKGROUND

In this section, we provide formal definitions of MAIDs and EFGs.

**Definition 3.1** (MAID [13, 21]). A **multi-agent influence diagram (MAID)** is a structure $\mathcal{M} = (\mathcal{G}, \boldsymbol{\theta})$ where $\mathcal{G} = (N, V, \mathcal{E})$ specifies a set of agents $N = \{1, \ldots, n\}$ and a DAG $(V, \mathcal{E})$. The set of variables $V$ is partitioned into chance variables $X$, decision variables $D$, and utility variables $U$; decision and utility variables are further partitioned based on which agent they belong to: $D = \bigcup_{i \in N} D^i$ and $U = \bigcup_{i \in N} U^i$. There are two types of edges in $\mathcal{E}$: edges into non-decisions (solid) represent probabilistic dependence and edges into decisions (dashed) are *information links* observed by an agent. The parameters $\boldsymbol{\theta} = \{\theta_V\}_{V \in V \setminus D}$ define the conditional probability distributions (CPDs) $\Pr(V \mid \mathbf{Pa}_V; \theta_V)$ for each non-decision variable

such that for *any* parameterisation of the decision variable CPDs, the resulting joint distribution over $V$ induces a Bayesian network.

*Example* 1 (Evaluation Game). Consider the evaluation example from the introduction. The nodes in fig. 1 correspond to particular MAID graphs representing the agents' beliefs – though, importantly, MAIDs do not specify the agents' higher-order beliefs. In the ground truth MAID, $\mathcal{M}^H$, the human $H$ is evaluating whether the AI system $A$ truthfully reports its capability level $C \in \{\text{low, high}\}$. Both agents directly observe the system's capabilities – there are information links $(C, D^H)$ and $(C, D^A)$. At the start of the game, the values of variables without parents are drawn from the prior specified by $\theta$. Suppose that $P(C = \text{high}) = 0.1$ and $P(C = \text{low}) = 0.9$. The AI agent $A$ observes the value of $C$ and chooses whether to report its capabilities at the decision $D^A$. The AI agent can demonstrate low or high capabilities, i.e., it can truthfully signal the value of $C$ or not. The human $H$ observes $A$'s capabilities and decision, then chooses to deploy or not deploy the AI system. The AI receives utility 1 if deployed and $-1$ if not deployed. The human receives utility 0 for not deploying $A$, utility 1 for deploying an honest $A$, and utility $-5$ for deploying a dishonest $A$. The human has correct beliefs about this situation. In contrast, according to $A$'s beliefs, instead of an honesty evaluation, it is an evaluation of capabilities. In the MAID representation of $A$'s beliefs, $\mathcal{M}^A$, the human only learns about $A$'s capabilities by observing $A$'s decision – there is no information link from $C$ to $D^H$. Further, $A$ believes that $H$ receives utility 0 for not deploying the AI, utility 1 for deploying an AI with low capabilities, and utility $-5$ for deploying an AI with high capabilities.

Agents' *policies* specify the CPD over their decisions.

**Definition 3.2.** Given a MAID $\mathcal{M} = (\mathcal{G}, \theta)$, a **decision rule** $\pi_D$ for $D \in D$ is a CPD $\pi_D(D \mid \mathbf{Pa}_D)$ and a **partial policy profile** $\pi_{D'}$ is a set of decision rules $\pi_D$ for each $D \in D' \subseteq D$. A (behavioural) **policy** $\pi^i$ refers to $\pi_{D^i}$, and a **policy profile** $\pi = (\pi^1, \ldots, \pi^n)$ is a tuple of policies. The tuple $\pi^{-i} := (\pi^1, \ldots, \pi^{i-1}, \pi^{i+1}, \ldots, \pi^n)$ specifies policies for all agents except $i$.

By combining $\pi$ with the partial distribution Pr over the chance and utility variables in a MAID, we obtain a joint distribution: $\text{Pr}^{\pi}(x, d, u) := \prod_{V \in V \setminus D} \text{Pr}(v \mid \mathbf{pa}_V) \cdot \prod_{D \in D} \pi_D(d \mid \mathbf{pa}_D)$, over all the variables in $\mathcal{M}$; inducing a Bayesian network. The expected utility for an agent $i$ given a policy profile $\pi$ is the expected sum of their utility variables $\sum_{U \in U^i} \mathbb{E}_{\pi}[U]$. We use the Nash equilibria concept.

**Definition 3.3.** A policy $\pi^i$ for agent $i \in N$ is a *best response* to $\pi^{-i}$, if for all policies $\hat{\pi}^i$ for $i$: $\sum_{U \in U^i} \mathbb{E}_{(\pi^i, \pi^{-i})}[U] \geq \sum_{U \in U^i} \mathbb{E}_{(\hat{\pi}^i, \pi^{-i})}[U]$. A policy profile $\pi$ is a *Nash equilibrium* if every policy in $\pi$ is a best response to the policies of the other agents.

*Example* 1 (continued). In the ground truth MAID $\mathcal{M}^H$, intuitively $H$ will deploy the system only if $A$ truthfully reports $C$. Formally, this is the policy $\pi^H(D^H = \text{deploy} \mid C, D^A) = 1$ if $C = D^A$ and 0 if $C \neq D^A$. The best response for $A$ is to be truthful: $\pi^A(D^A = C \mid C) = 1$. It can be easily shown that these policies constitute a Nash equilibrium in $\mathcal{M}^H$. However, $\pi^A$ is not optimal given $A$'s subjective beliefs, since $A$ believes that $H$ does not observe its capabilities and wants to deploy it if it demonstrates a low capability level. In the MAID representing $A$'s beliefs $\mathcal{M}^A$, one Nash equilibrium is for $A$ to always report low capabilities and for $H$ to deploy iff $A$ reports

low capabilities. Hence, in the actual ground-truth situation, we might expect that $A$ hides its high capabilities but is never deployed by $H$ — but this is not a Nash equilibrium in either MAID.

Usually, a MAID describes the objective world, and it is assumed to be *common knowledge* that this MAID describes reality. However, an agent may be uncertain or incorrect about the game they are playing or the beliefs of other agents (as in Example 1). Settings in which agents are uncertain about aspects of the game structure are known as *incomplete information games*. Incomplete information MAIDs (II-MAIDs), introduced in Section 4, will enable us to explicitly model the varied subjective beliefs that arise in these settings.

We now define EFGs (see Supplementary Material for the running example in EFG form). EFGs are a common game representaiton and we will show equivalence result (Theorem 6.2) between II-MAIDs and a form of II-EGFs introduced in Section 5.

**Definition 3.4** ( EFG [23]). An **extensive form game (EFG)** is a structure $\mathcal{E} = (N, T, P, A, \lambda, I, U)$. $N = \{1, \ldots, n\}$ is a set of agents. $T = (V, \mathscr{E})$ is a game tree with nodes $V$ connected by edges $\mathscr{E}$ that are partitioned into sets $V^0, V^1, \ldots, V^n, L$ where $R \in V$ and $L \subset V$ are the root and leaves of $T$, respectively, $V^0$ are chance nodes, and $V^i$ are the decision nodes controlled by agent $i \in N$. $P = \{P_1, \ldots, P_{|V^0|}\}$ is a set of probability distributions $P_j(\mathbf{Ch}_{V_j^0})$ over the children of each chance node $V_j^0$. $A$ is a set of actions, where $A_j^i \subseteq A$ denotes the set of actions available at each node in $V_j^i \in V^i$; $\lambda : \mathscr{E} \to A$ is a labelling function mapping each edge $(V_j^i, V_l^k)$ to an action $a \in A_j^i$, where for any node $V$, different outgoing edges must map to different actions. $I = \{I^1, \ldots, I^n\}$ contains a set of information sets $I^i$ for each agent $i \in N$, where $I^i \subset 2^{V^i}$ partitions the decision nodes $V^i$ belonging to agent $i$. $U : L \to \mathbb{R}^n$ is a utility function mapping each leaf node to a vector that determines the final payoff for each agent. A **history** $h \in H$ is a sequence of actions (including values of chance variables) leading from the root of the game tree to a particular node. Each node $v \in V$ is associated with a unique history $h(v)$. An **observation** at decision node $I_{j,k}^i$ in information set $I_j^i \in I^i$ for agent $i \in N$ is the sequence of actions and chance outcomes that are identical across all histories $\{h(v) : v \in I_j^i\}$ at their corresponding positions.

Analogous to MAID policies, in EFGs agents choose *strategies*.

**Definition 3.5** ([12]). Given an EFG $\mathcal{E} = (N, T, P, A, \lambda, I, U)$, a (behavioural) **strategy** $\sigma^i$ for a player $i$ is a set of probability distributions $\sigma_j^i : A_j^i \to [0, 1]$ over the actions available at each of their information sets $I_j^i$, where $A_j^i$ is the set of actions available at any node in $I_j^i$ (which must be the same for all nodes in the information set). A **strategy profile** $\sigma = (\sigma^1, \sigma^2, \ldots, \sigma^n)$ is a tuple of strategies for all players $i \in N$ and $\sigma^{-i} = (\sigma^1, \ldots, \sigma^{i-1}, \sigma^{i+1}, \ldots, \sigma^n)$ denotes the partial strategy profile of all players other than $i$. The distributions in $P$ along with $\sigma$ define a full probability distribution over paths in $\mathcal{E}$.

We use perfect recall to prove the existence of Nash Equilibria.

**Definition 3.6** ([21]). Agent $i$ in a MAID $\mathcal{M}$ is said to have **perfect recall** if there exists a total ordering $D_1 \prec \cdots \prec D_m$ over $\mathbf{D}^i$ such that $(\mathbf{Pa}_{D_j} \cup D_j) \subseteq \mathbf{Pa}_{D_k}$ for any $1 \leq j < k \leq m$. $\mathcal{M}$ is a **perfect recall game** if all agents in $\mathcal{M}$ have perfect recall.

**Definition 3.7.** An EFG is said to be a **perfect recall game** if, for each player $i \in N$, and any two decision nodes $v, v' \in V^i$ in the same information set $I_{j,k}^i$, the following two conditions hold. First, the sequences of actions taken by player $i$ leading to $v$ and $v'$ must be identical. Second, the sequences of information sets visited by player $i$ on the paths to $v$ and $v'$ must be identical.

Finally, prior work [12] has established an equivalence result between MAIDs and EFGs. This result takes the form of two transformation procedures converting between MAIDs and EFGs, called `efg2maid` and `maid2efg`. These transformations both imply the existence of a map from strategies in the EFG to policies in the MAID, such that expected utilities are preserved for all agents. This means that under either transformations, equilibria in the original game are equilibria in the resulting game.

## 4 INCOMPLETE INFORMATION MAIDS

In this section we introduce our II-MAIDs framework. A core component of the framework is a set **S** containing *subjective MAIDs*. A subjective MAID is a self-referential object describing a possible game as envisioned by either the external modeller (we call this the objective model $S^*$) or an agent playing the game. A subjective MAID $S \in \mathbf{S}$ consists of a MAID $\mathcal{M}$ that describes the game being played and beliefs $P_i^S$ for each agent $i$ in the game. The notation $P_i^S$ denotes agent $i$'s prior over **S** when the objective model is $S$, and $P_i^S(S')$ denotes the probability ascribed by agent $i$ to subjective MAID $S'$ given that the objective MAID is $S$.

**Definition 4.1** (II-MAID). An **incomplete information MAID (II-MAID)** is a tuple $\mathcal{S} = (\mathbf{N}, S^*, \mathbf{S})$, where **N** is a set of agents, **S** is a set of subjective MAIDs, $S^* \in \mathbf{S}$ is the correct objective model, and each **subjective MAID** is a tuple $S = (\mathcal{M}^S, (P_i^S)_{i \in \mathbf{N}}) \in \mathbf{S}$ with $\mathcal{M}^S$ a MAID and $P_i^S$ a prior over **S** for agent $i$ such that the following "coherence condition" [14] holds:

$$P_i^S(\{S' \in \mathbf{S} : P_i^{S'} = P_i^S\}) = 1 \quad \forall i \in \mathbf{N}, S \in \mathbf{S}.$$

Note the recursive nature of the subjective MAIDs **S**: each element $S \in \mathbf{S}$ specifies subjective beliefs over **S** for each agent, $(P_i^S)_{i \in \mathbf{N}}$. This allows us to model belief hierarchies of arbitrary and infinite depth – as in fig. 1. Agent $i$ "observes" $P_i^{S^*}$ at the start of the game, and this justifies the coherence condition, which represents that the agents know their own beliefs: since agent $i$ knows $P_i^{S^*}$, she can rule out all subjective MAIDs $S$ for which $P_i^S \neq P_i^{S^*}$. Finally, note that II-MAIDs are a strict generalisation of MAIDs: a standard MAID is an II-MAID in which $P_i^{S^*}(S^*) = 1 \quad \forall i \in \mathbf{N}$, i.e. all agents assign probability 1 to $S^*$, the objective model.

*Example 1* (continued). We can represent our running example, including the full infinite belief hierarchy, as an II-MAID as follows: $\mathbf{N} = \{H, A\}, \mathbf{S} = \{S^H, S^A\}$, and $S^* = S^H$, where

$$S^H = (\mathcal{M}^H, (P_H^{S^H}(S^H) = 1, P_A^{S^H}(S^A) = 1))$$
$$S^A = (\mathcal{M}^A, (P_H^{S^A}(S^A) = 1, P_A^{S^A}(S^A) = 1))$$

The subjective MAID $S^H$ is the correct objective model, and is also believed with certainty by $H$. It specifies the ground truth MAID $\mathcal{M}^H$ represented in fig. 1, and additionally specifies the agents' beliefs, i.e., $H$'s certainty in $S^H$ and $A$'s misplaced certainty in $S^A$.

$A$'s beliefs are represented by $S^A$, which captures $A$'s certainty about the MAID $\mathcal{M}^A$ and $A$'s mistaken belief that $H$ is also certain about $S^A$. In fact, $A$ believes it is common knowledge that $S^A$ is the true II-MAID. $S^H$ and $S^A$ concisely convey the objective game and all higher-order beliefs for $H$ and $A$.

A common assumption in the incomplete information games literature [14–16] is that agents' beliefs can be derived from a common prior, i.e., agents have *consistent beliefs*. This assumption means that there exists some common knowledge prior distribution $p$ over the set of subjective MAIDs **S**, such that upon arriving in any subjective MAID $S \in \mathbf{S}$, agents perform Bayesian updating to yield their beliefs. This assumption allows for a game with incomplete information to be converted into a game with imperfect information [14], but places a strong constraint on the types of belief hierarchies that can be modelled; namely, the following assumption must hold.

**Assumption 4.2** (Consistency condition). Given II-MAID $\mathcal{S} = (\mathbf{N}, S^*, \mathbf{S})$, there exists some distribution $p$ over **S** such that

$$p(S') = \sum_{S \in \mathbf{S}} P_i^S(S')p(S) \quad \text{for all } S' \in \mathbf{S}, i \in \mathbf{N}. \tag{1}$$

This condition implies that the beliefs of each agent derive from Bayesian updating of a common prior. We now show that our running example cannot satisfy Assumption 4.2.

*Example 1* (continued). In our running example, Assumption 4.2 would imply that $p(S^H) = 0$, which would force $H$ to assign zero probability to $S^H$ in both $S^H$ and $S^A$. Thus, consistency does not hold in this example. As argued on page 423 of Maschler et al. [24], most game theoretic formulations of incomplete information assume consistent beliefs, although "the set of consistent situations is a set of measure zero within the set of belief spaces" [24].

**Information Sets.** An agent in an II-MAID may be uncertain about the current situation given its observations. In eventualities that are indistinguishable to the agent, its policy must specify the same behaviour. We therefore need to define an analogy to information sets in EFGs. At a decision node $D$, an agent observes both the values of $\mathbf{pa}_D$ and the set of actions available, $dom(D)$. We define the *information sets in an II-MAID* according to these observations.

**Definition 4.3.** Given an II-MAID $\mathcal{S} = (\mathbf{N}, S^*, \mathbf{S})$, we iteratively build the **information sets**. For each subjective MAID $S \in \mathbf{S}$ and each agent $i \in \mathbf{N}$, denote $\mathbf{D}_i(S)$ as the set of decision nodes for agent $i$ in $\mathcal{M}^S$, $Pa_{D_i}(S)$ as the set of parents of $D_i$ in $\mathcal{M}^S$, and $\Pr_S^\pi(\cdot)$ as the distribution of variables in $\mathcal{M}^S$ under some policy $\pi$. Define

$$\mathbf{I}_{S,i} := \cup_{D_i \in \mathbf{D}_i(S)} \{(\mathbf{pa}_{D_i}, dom(D_i)) \mid \mathbf{pa}_{D_i} \in dom(Pa_{D_i}(S))$$
$$: \Pr_S^\pi(\mathbf{pa}_{D_i}) > 0 \text{ for some } \pi\}.$$

Then *agent $i$'s information sets* are defined as $\mathbf{I}_i(\mathcal{S}) := \cup_{S \in \mathbf{S}} \mathbf{I}_{S,i}$. Finally, we can define the set of information sets as $\mathbf{I}(\mathcal{S}) = (\mathbf{I}_i(\mathcal{S}))_{i \in \mathbf{N}}$.

**Definition 4.4** (Encounterability). An information set $I = (\mathbf{p}, \mathbf{d}) \in \mathbf{I}_i(\mathcal{S})$ is *encounterable* by agent $i$ in subjective MAID $S'$ if in $\mathcal{M}^{S'}$ there exists $D_i$ such that $\mathbf{p} \in dom(\mathbf{pa}_{D_i})$ and $\mathbf{d} = dom(D_i)$.

The notion of encounterability is important because the games in which $I$ is encounterable are the games that agent $i$ needs to consider when making a decision given information set $I$.

**Definition 4.5.** We define an II-MAID $\mathcal{S} = (\mathbf{N}, S^*, \mathbf{S})$ as having *perfect recall* if for each $S \in \mathbf{S}$, $\mathcal{M}^S$ is a perfect recall game.

In an II-MAID, an agent's **policy** defines a distribution over actions for each of that agent's information sets.

**Definition 4.6.** Given an II-MAID $\mathcal{S} = (\mathbf{N}, S^*, \mathbf{S})$, a **decision rule** $\pi_I$ for $I = (\mathbf{x}, \mathbf{d}) \in \mathbf{I}(\mathcal{S})$, where $\mathbf{x}$ is a setting of the decision's parents and $\mathbf{d}$ is an action set, is a CPD $\pi_I(\cdot \mid \mathbf{x})$ over $\mathbf{d}$. A **partial policy profile** $\pi_{I'}$ is a set of decision rules $\pi_I$ for each $I \in I' \subseteq I(\mathcal{S})$, where we write $\pi_{-I'}$ for the set of decision rules for each $I \in I(\mathcal{S}) \setminus I'$. A (behavioural) **policy** $\pi^i$ refers to $\pi_{I_i(\mathcal{S})}$, a (full, behavioural) **policy profile** $\pi = (\pi^1, \ldots, \pi^n)$ is a tuple of policies, and $\pi^{-i} := (\pi^1, \ldots, \pi^{i-1}, \pi^{i+1}, \ldots, \pi^n)$.

We note that unlike in standard MAIDs, in which a decision rule specifies behaviour at a given decision variable in all contexts, decision rules in II-MAIDs specify a CPD given a single context. We want to avoid situations with undefined actions that might arise for example when an agent places zero probability on the correct game, i.e., $P_i^{S^*}(S^*) = 0$, and to avoid forcing $P_i^S(S') > 0$ for all $i \in \mathbf{N}, S, S' \in \mathbf{S}$. Therefore, policies specify the agent's decisions even for eventualities they deem impossible.

We can calculate the subjective expected utility of a joint behaviour policy for agent $i$ according to their beliefs $P_i^{S^*}$ as $\mathcal{U}_{S^*}^i(\pi) := \sum_{S \in \mathbf{S}} \sum_{U \in \mathbf{U}^i(S)} \sum_{u \in dom(U)} u \Pr_S^\pi(U = u) P_i^{S^*}(S)$, where $\mathbf{U}^i(S)$ is the set of utility variables associated with agent $i$ in $\mathcal{M}^S$ and $\Pr_S^\pi$ is the post-policy distribution of variables in $\mathcal{M}^S$.

## 5 INCOMPLETE INFORMATION EFGS

We now present a formalisation of EFGs with incomplete information (II-EFGs) and without the consistency assumption, as per [25]. Our formalisation modifies the framework from [24] to use EFGs rather than normal-form games, and serves as an important connection between II-MAIDs and existing literature. We start with a definition of belief spaces.

**Definition 5.1** (Adapted from Def 10.1 in [24]). Let $\mathbf{N}$ be a finite set of agents and $(S, \mathcal{S})$ be a measurable space of EFGs. A *belief space* of the set of agents $\mathbf{N}$ over the set of states of nature is an ordered vector $\Pi = (Y, \mathcal{Y}, \mathbf{s}, (b_i)_{i \in \mathbf{N}})$, where $(Y, \mathcal{Y})$ is a measurable space of states of the world; $\mathbf{s} : Y \to S$ is a measurable function, mapping each state of the world to an EFG. For each agent $i \in \mathbf{N}$, a function $b_i : Y \to \Delta(Y)$ maps each state of the world $\omega$ to a distribution over $Y$. We will denote the probability that agent $i$ ascribes to event $E \subseteq Y$, according to their distribution $b_i(\omega)$, by $b_i(E \mid \omega)$. We require the functions $(b_i)_{i \in \mathbf{N}}$ to satisfy the following conditions:

- Coherence: for each agent $i \in \mathbf{N}$ and each $\omega \in Y$, the set $\{\omega' \in Y : b_i(\omega') = b_i(\omega)\}$ is measurable in $Y$ and $b_i(\{\omega' \in Y : b_i(\omega') = b_i(\omega)\} \mid \omega) = 1$.
- Measurability: for each agent $i \in \mathbf{N}$ and each measurable set $E \in \mathcal{Y}$, the function $b_i(E \mid \cdot) : Y \to [0, 1]$ is measurable.

A state of the world in a belief space takes the form $\omega = (\mathbf{s}(\omega), b_1(\omega), \ldots, b_n(\omega))$, where $\mathbf{s}(\omega)$ is the true EFG being played, and $b_i(\omega)$ is the *type* of agent $i$, a distribution over states of the world representing agent $i$'s beliefs. When in state of the world $\omega$, agent

$i$ has beliefs $b_i(\omega)$, but does not necessarily know the state of the world (or $\mathbf{s}(\omega)$), since there may be some $\omega' \in Y$ such that $b_i(\omega') = b_i(\omega)$. It is assumed that all agents know $b_j(\omega')$ for all $j \in \mathbf{N}$ and all $\omega' \in Y$, and so $b_i(\omega)$ defines a full belief hierarchy for agent $i$. For example, when in state of the world $\omega$, agent $i$ believes that agent $j$ places $\sum_{\omega' \in Y} b_i(\omega' \mid \omega) b_j(\omega'' \mid \omega')$ probability on the state of the world being $\omega''$.

**Definition 5.2** (Adapted from Def 10.37 in [24]). An *incomplete information EFG (II-EFG)* is an ordered vector $G = (\mathbf{N}, S, \Pi)$, where $\mathbf{N}$ is a finite set of agents, $S$ is a finite set of EFGs $s = (\mathbf{N}, T_s, \mathbf{P}_s, \mathbf{D}_s, \lambda_s, \mathbf{I}(s), U_s)$, and $\Pi = (Y, \mathcal{Y}, \mathbf{s}, (b_i)_{i \in \mathbf{N}})$ is a belief space of the players $\mathbf{N}$ over the set of EFGs $S$. An II-EFG $G = (\mathbf{N}, S, \Pi)$ has **perfect recall** if for each $s \in S$, s is a perfect recall EFG.

Intuitively, we can think of a meta-information set for agent $i$ as a belief $b_i(\omega)$ and a set of information sets in different games that the agent cannot distinguish between at the point of decision, given beliefs $b_i(\omega)$. Arriving at a node in one of these information sets, the agent is unable to distinguish between some possible histories, and potentially some possible EFGs. Therefore, strategies in this type of game must define a mixed action at each meta-information set.

**Definition 5.3.** The *meta-information sets* $\mathbf{I}^i$ for agent $i \in \mathbf{N}$ in an II-EFG $G = (\mathbf{N}, S, \Pi)$ are defined as follows. Let $\mathcal{I}^i = \cup_{s \in S} \mathbf{I}^i(s)$ be the set of all information sets for agent $i$ across all EFGs $s \in S$. Define an equivalence relation $\sim$ on elements of $\mathcal{I}^i$ such that $\mathbf{I}^i(s) \ni I_k^i(s) \sim I_l^i(s') \in \mathbf{I}^i(s')$ iff: (1) $\mathbf{D}_{s,k}^i = \mathbf{D}_{s',l}^i$. That is, the nodes in both information sets must have the same set of available actions. (2) The nodes in $I_k^i(s)$ and $I_l^i(s')$ must have the same observations. Define the "belief-free" meta-information sets $\mathbf{I}_{bf}^i = \mathcal{I}^i/\sim$, the quotient set of $\mathcal{I}^i$ by $\sim$, i.e., the set of equivalence classes partitioning $\mathcal{I}^i$. Letting $\mathcal{T}^i = \{b_i(\omega) : \omega \in Y\}$ be the set of possible beliefs for agent $i$, we set $\mathbf{I}^i = \mathbf{I}_{bf}^i \times \mathcal{T}^i$.

This formalisation of II-EFGs generalises the well-known Harsanyi game with incomplete information [14], by dropping the Assumption 4.2. The game has two stages, known as the ex-ante and interim stages. The former takes place before the state of the world $\omega \in Y$ is selected. We note that without a common prior, there is no distribution from which a state of the world can be said to be selected, and so the procedure by which it is generated is left unspecified. The work we present here concerns the interim stage of the game, which takes place after the state of the world has been selected. At this stage, all agents $i$ know their type $b_i(\omega)$.

*Example 1* (continued). Our example can be described with an II-EFG $(\mathbf{N}, S, \Pi)$ at interim stage, where $\Pi = (Y, \mathcal{Y}, \mathbf{s}, (b_i)_{i \in \mathbf{N}})$. $\mathbf{N} = \{H, A\}$, and we let $Y = \{\omega^*, \omega^a\}$, where the true state of the world is $\omega^*$, and the state of the world assumed true by the agent is $\omega^a$. We define two EFGs, $\mathbf{s}(\omega^*)$ and $\mathbf{s}(\omega^a)$, in the Supplementary Material. $S$ is a set containing these two EFGs. The beliefs $b_i(\omega)$ for each $\omega \in Y$ and each agent $i \in \mathbf{N}$ are $b_H(\omega^* \mid \omega^*) = 1, b_H(\omega^a \mid \omega^a) = 1, b_A(\omega^a \mid \omega^*) = 1, b_A(\omega^a \mid \omega^a) = 1$.

We let $\mathbf{I}_i^t$ be the set of meta-information sets with belief $t \in \{b_i(\omega) : \omega \in Y\}$ and $\mathbf{D}_I$ be the action set at meta-information set $I$.

**Definition 5.4** (Adapted from Def 10.38 in [24]). A *behaviour strategy* of player $i$ in an II-EFG $G = (\mathbf{N}, S, \Pi)$ is a tuple $\sigma_i = (\sigma_i^\omega)_{\omega \in Y}$

with each element a measurable function $\sigma_i^\omega \in \bigtimes_{I^i \in \mathbf{I}_i^{b_i(\omega)}} \Delta(\mathbf{D}_{I^i})$ for some state of the world $\omega \in Y$. $\sigma_i^\omega$ determines a mixed action for each meta-information set with belief $b_i(\omega)$. $\sigma_i^\omega$ is dependent solely on the type of the player $b_i(\omega)$. In other words, for each $\omega, \omega' \in Y$, $b_i(\omega) = b_i(\omega') \implies \sigma_i^\omega = \sigma_i^{\omega'}$. A *joint behaviour strategy* takes the form $\sigma = (\sigma_i)_{i \in \mathbf{N}}$. Further denote $\sigma^\omega = (\sigma_i^\omega)_{i \in \mathbf{N}}$. We denote by $\sigma_i[I]$ the behaviour of agent $i$ at meta-information set $I$.

## 6 EQUIVALENCE OF II-MAIDS AND II-EFGS

In this section, we show that our framework is equivalent to the interim stage of an II-EFG. At the interim stage of an *II-EFG* $G = (\mathbf{N}, S, \Pi)$ where $\Pi = (Y, \mathcal{Y}, \mathbf{s}, (b_i)_{i \in \mathbf{N}})$, with state of the world $\omega$, the true EFG is defined by $\mathbf{s}(\omega)$, and the belief hierarchies are defined by $b_i(\omega)$, for each agent $i \in \mathbf{N}$. In an *II-MAID* $\mathcal{S} = (\mathbf{N}, S^*, \mathbf{S})$ with objective model $S^* = (\mathcal{M}^{S^*}, (P_i^{S^*})_{i \in \mathbf{N}})$, the true MAID is $\mathcal{M}^{S^*}$ and the belief hierarchies are defined by $P_i^{S^*}$ for each agent $i \in \mathbf{N}$. Intuitively, the two frameworks are representing the same things, though our framework takes the games upon which belief hierarchies are built to be MAIDs, not EFGs.

We now show, using results connecting EFGs to MAIDs that there exists a natural mapping between strategies in the two frameworks that preserves expected utilities according to the agents' subjective models, and therefore preserves Nash equilibria. We first define a notion of equivalence, such that if an II-MAID $\mathcal{S}$ and an II-EFG $G$ are equivalent, then there exists such a natural mapping.

**Definition 6.1** (Equivalence). We say that an II-MAID $\mathcal{S} = (\mathbf{N}, S^*, \mathbf{S})$ and an II-EFG $G = (\mathbf{N}, S, \Pi)$ at interim stage, with state of the world $\omega$, are *equivalent* if there is a bijection $f : \Sigma \to Q/\sim$ between the strategies $\Sigma$ in $G$'s interim stage, and a partition of the policies $Q$ in $\mathcal{S}$ (the quotient set of $Q$ by an equivalence relation $\sim$) such that: (1) for $\pi, \pi' \in Q$, $\pi \sim \pi'$ only if $\pi_i$ and $\pi_i'$ differ only on null decision contexts according to $P_i^{S^*}$, for each agent $i \in \mathbf{N}$, and (2) for every $\pi \in f(\sigma)$ and every agent $i \in \mathbf{N}$, $\mathcal{U}_{\mathcal{S}}^i(\pi) = \gamma_i^G(\sigma \mid \omega)$, for each $\sigma \in \Sigma$. We refer to $f$ as a *natural mapping* between $G$ and $\mathcal{S}$.

We leverage maid2efg and efg2maid [12] to construct transformations between II-MAIDs and II-EFGs, which we denote maid2efgII and efg2maidII (see Supplementary Material). These transformations start by mapping all MAIDs (EFGs) in the belief hierarchy to EFGs (MAIDs) using maid2efg (efg2maid), and then match up the corresponding features of the frameworks as detailed above. They guarantee a one-to-one correspondence between meta-information sets in the II-EFG and information sets in the II-MAID, allowing for a simple map between strategies and policies.

**Theorem 6.2.** *If $G = \text{maid2efgII}(\mathcal{S})$ or $\mathcal{S} = \text{efg2maidII}(G)$, $G$ and $\mathcal{S}$ are equivalent.*

This result shows that II-MAIDs and II-EFGs at the interim stage can represent the same set of games.

## 7 EQUILIBRIA CONCEPTS IN II-GAMES

In this section we define important equilibria concepts from the literature in II-EFGs. We prove that the Nash equilibria concept is inherited by II-MAIDS due to the equivalence result in section 6. The Bayesian equilibria concept is not inherited by II-MAIDs since it applies to the ex-ante stage, which II-MAIDs do not represent.

Additionally we argue that the Nash equilibria concept is not appropriate in the inconsistent beliefs setting.

In an II-EFG, given a joint strategy $\sigma$, agent $i$'s expected utility when in state of the world $\omega$ (according to their beliefs $b_i(\omega)$) is

$$\gamma_i^G(\sigma \mid \omega) := \sum_{\omega' \in Y} \mathcal{U}_{\mathbf{s}(\omega')}^i(\sigma^{\omega'}) b_i(\omega' \mid \omega)$$
$$= \sum_{\omega' \in \{\omega' : b_i(\omega') = b_i(\omega)\}} \mathcal{U}_{\mathbf{s}(\omega')}^i(\sigma_i^\omega, \sigma_{-i}^{\omega'}) b_i(\omega' \mid \omega).$$

Hence $\gamma_i^G$ depends on agent $i$'s strategy only through $\sigma_i^\omega$. This follows from the coherence condition $b_i(\{\omega' \in Y : b_i(\omega') = b_i(\omega)\} \mid \omega) = 1$. From now, for notational simplicity, we write $\gamma_i^G(\sigma_i^\omega, \sigma_{-i} \mid \omega)$ to denote $\gamma_i^G(\sigma \mid \omega)$ where $\sigma$ is any strategy profile with the given $\sigma_i^\omega$. Under some assumptions, at the interim stage, we can prove the existence of Nash equilibria.

**Definition 7.1.** A *Nash equilibrium* at the interim stage of an II-EFG with state of the world $\omega$ is a strategy profile $\hat\sigma$ satisfying

$$\gamma_i^G(\hat\sigma_i^\omega, \hat\sigma_{-i} \mid \omega) \geq \gamma_i^G(\sigma_i^\omega, \hat\sigma_{-i} \mid \omega), \forall i \in \mathbf{N}, \forall \sigma_i^\omega \in \bigtimes_{I^i \in \mathbf{I}_i^{b_i(\omega)}} \Delta(\mathbf{D}_{I^i})$$

**Theorem 7.2.** *Let $G = (\mathbf{N}, S, \Pi)$ be an II-EFG with perfect recall, where $Y$ is a finite set of states of the world, and each player $i$ has a finite set of actions $\mathbf{D}_i$. Then at the interim stage, $G$ has a Nash equilibrium in behaviour strategies.*

Note that $\sigma^\omega$ has the same expected payoff for agent $i$ for all $\omega'$ such that $b_i(\omega') = b_i(\omega)$. Hence, if $\sigma_i^\omega$ maximises $\gamma_i^G$ (making it a best response according to $i$'s beliefs) to $\sigma_{-i}^\omega$ in $\omega$, it is also a best response in $\omega'$.

We prove the existence of a Bayesian equilibrium at the ex-ante stage of the game. For this, we follow Theorem 10.42 of Maschler et al. [24], adapting it for EFGs with perfect recall.

**Definition 7.3** ([24] 10.39). A *Bayesian equilibrium* is a strategy profile $\hat\sigma = (\hat\sigma_i)_{i \in \mathbf{N}}$ satisfying

$$\gamma_i^G(\hat\sigma_i^\omega, \hat\sigma_{-i} \mid \omega) \geq \gamma_i^G(\sigma_i^\omega, \hat\sigma_{-i} \mid \omega),$$
$$\forall i \in \mathbf{N}, \forall \sigma_i^\omega \in \bigtimes_{I^i \in \mathbf{I}_i^{b_i(\omega)}} \Delta(\mathbf{D}_{I^i}), \forall \omega \in Y.$$

There is no general result proving the existence of Bayesian equilibria in II-games, however we have the following.

**Theorem 7.4** (Adaptation of [24] Theorem 10.42). *Let $G = (\mathbf{N}, S, \Pi)$ be an II-EFG with perfect recall, where $Y$ is a finite set of states of the world, and $\mathbf{D}_i$ is finite for all agents $i \in \mathbf{N}$. Then at ex-ante stage, $G$ has a Bayesian equilibrium in behaviour strategies.*

**Equilibria in II-MAIDS.** The equivalence of II-EFGs and II-MAIDS mean that II-MAIDs inherit theoretical guarantees of II-EFGs, including the existence of Nash equilibria in the case of perfect recall and finite $\mathbf{S}$ and finite action spaces. (Theorem 7.4 does not carry over to II-MAIDs, since the equivalence is with the interim stage of II-EFGs, and Bayesian equilibria exist in the ex-ante stage.)

**Theorem 7.5.** *Let $\mathcal{S} = (\mathbf{N}, S^*, \mathbf{S})$ be an II-MAID where $\mathbf{S}$ is finite, $\mathbf{S}$ has perfect recall, and $\text{dom}(V)$ is finite for each $V \in \mathcal{M}^S$ for each $S \in \mathbf{S}$. Then $\mathcal{S}$ has a Nash equilibrium in behaviour policies.*

**Nash equilibria in II-games violate common knowledge of rationality.** Nash equilibria in II-games require that agents believe they are playing a best response to the other agents' policies. However, they do not require that an agent believes that another agent is acting rationally: according to agent $i$'s beliefs about agent $j$'s beliefs, agent $j$'s policy may not be a best response.

*Example 1* (continued). In the running example, as represented with an II-MAID in Section 4, policies require assigning decision rules to all information sets in the II-MAID. The AI agent $A$ has two information sets, defined by the value of its capabilities. Suppose that $A$ chooses the policy $\pi^A$ to always portray low capabilities. The human has information sets corresponding to all combinations of $(C, D^A)$ that it could observe in $\mathcal{M}^H$, and both values of $C$ in $\mathcal{M}^A$. A best response to $\pi^A$ requires that $H$ cannot do better according to their own beliefs. Since $H$ is certain in $\mathcal{M}^H$, a best response for $H$ can specify any behaviour in $\mathcal{M}^A$, even random behaviour. Hence, the policy for $H$, $\pi^H$, which deploys iff $C = D^A$ in $\mathcal{M}^H$, and acts randomly in $\mathcal{M}^A$ is a best response to $\pi^A$. These two policies constitute a Nash equilibrium in the II-MAID.

However, according to $A$'s beliefs about $H$'s beliefs, $H$ is acting irrationally in this Nash equilibrium. According to $A$, $H$ believes in $\mathcal{M}^A$ wherein $H$'s best response to $\pi^A$ is to always deploy. The solution concept provides no justification for why $A$, believing that $\mathcal{M}^A$ is common knowledge, should expect $H$ to act randomly in $\mathcal{M}^A$. Under common knowledge of rationality, every player should expect every other player to play a best response to their beliefs about the game structure and the policies of other players. Under Assumption 4.2, Bayesian Perfect Equilibria respect common knowledge of rationality [22]. In the next section, we define II-MAIDs of finite-depth and develop a solution concept guaranteeing common knowledge of rationality without consistent beliefs.

## 8 FINITE-DEPTH II-MAIDS

II-MAIDs can capture belief hierarchies of infinite depth. However, in the real world, humans or AI agents may retain only a finite hierarchy of beliefs. In this section, we define depth-$k$ II-MAIDs and present means by which to calculate an agent's rational behaviour given its subjective MAID resulting in a *recursive best response solution*. This solution constitutes a more realistic solution concept for settings with inconsistent beliefs than the concepts discussed in Section 7 and [24]. We begin by defining a new type of MAID, produced by assigning policies to some or all agents in a MAID.

**Definition 8.1.** A *partial-post-policy MAID* is a structure $\mathcal{M} = (\mathcal{G}, \theta, \xi)$ where $(\mathcal{G}, \theta)$ is a MAID and $\xi = \{\pi_D\}_{D \in \mathscr{D}}$ with $\mathscr{D} \subseteq \boldsymbol{D}$. Each $\pi_D \in \xi$ is a conditional distribution $\Pr(D \mid \boldsymbol{Pa}_D)$ and parameterises its corresponding decision variable $D \in \boldsymbol{D}$ in $(\mathcal{G}, \theta)$.

We allow II-MAIDs to contain belief hierarchies upon partial-post-policy MAIDs: any subjective MAID $S = (\mathcal{M}^S, (P_i^S)_{i \in \boldsymbol{N}})$ may contain a partial-post-policy MAID $\mathcal{M}^S = (\mathcal{G}, \theta, \xi)$ as its first entry. In the case that for a given agent $i$ in $(\mathcal{G}, \theta)$, $\xi = \{\pi_D\}_{D \in \mathscr{D}}$ is such that $\boldsymbol{D}^i \subseteq \mathscr{D}$, then agent $i$ is not assigned a belief $P_i^S$ in the subjective MAID $S$. In other words, if an agent is ascribed a policy in $S$, then it is not ascribed a belief hierarchy – it is modelled as being a non-agent part of the environment that cannot adapt its policy.

A depth-0 subjective MAID for agent $i$ fixes the policies for all agents except $i$, resulting in a single-agent decision-problem.

**Definition 8.2.** A *depth-0 subjective MAID for agent $i$* is a partial-post-policy MAID $\mathcal{M}^S = (\mathcal{G}, \theta, \xi)$ where $\xi = \{\pi_D\}_{D \in \mathscr{D}}$ and $\boldsymbol{D}^j \subseteq \mathscr{D}$ for all agents $j \in \boldsymbol{N} \setminus \{i\}$.

Expected utilities for this single-agent decision problem can be assessed in isolation, allowing us to begin a recursive best-response structure that yields a sensible solution concept. Depth-$k$ subjective MAIDs are defined by induction as assigning positive weight to some depth-$(k-1)$ subjective MAID, and not assigning positive weight to subjective MAIDs of depth-$k$ or higher.

**Definition 8.3.** A *depth-$k$ subjective MAID for agent $i$* is a tuple $S = (\mathcal{M}^S, (P_j^S)_{j \in \boldsymbol{N} \setminus \{i\}})$ where $\mathcal{M}^S$ is a (partial-post-policy) MAID, each $P_j^S$ is a probability distribution over subjective MAIDs of depth-$(k-1)$ or lower for agent $j$, and there exists at least one depth-$(k-1)$ subjective MAID $S'$ for an agent $j$ such that $P_j^S(S') > 0$.

**Definition 8.4.** A *depth-$k$ II-MAID* is a tuple $\mathcal{S} = (\boldsymbol{N}, S^*, \boldsymbol{S})$. $S^* = (\mathcal{M}^{S^*}, (P_i^{S^*})_{i \in \boldsymbol{N}})$ where $\mathcal{M}^{S^*}$ is a (partial-post-policy) MAID, each $P_i^{S^*}$ is a probability distribution over subjective MAIDs of depth-$(k-1)$ or lower for agent $i$, and there exists at least one depth-$(k-1)$ subjective MAID $S$ for an agent $i$ such that $P_i^{S^*}(S) > 0$.

$S^*$ is similar to a depth-$k$ subjective MAID for an agent, except that since this represents the objective game rather than the game from a particular agent's perspective, all agents have beliefs. We use the term "depth-$k$ MAID" to generically refer to objective or subjective MAIDs of depth-$k$.

**Solution concepts in finite-depth II-MAIDs.** A solution concept for finite-depth II-MAIDs should accurately describe and predict the behaviour of rational agents in settings of incomplete information and inconsistent beliefs. The existing solution concepts for incomplete information games fail to respect common knowledge of rationality; the solution concept we define in this section will remedy this shortcoming. It is based on the premise that rational behaviour for agents in depth-1 subjective MAIDs can be straightforwardly computed with expected utility maximisation. This enables us to replace some beliefs with policies, reducing the depth of the belief hierarchy. Then this process can be repeated until we recover a policy profile in the top-level II-MAID that satisfies common knowledge of rationality.

Throughout this section, we invoke an "open-mindedness" assumption to ensure that each agent's policy selects an action for every encounterable information set.

**Definition 8.5** (Open-mindedness). We say that an II-MAID $\mathcal{S} = (\boldsymbol{N}, S^*, \boldsymbol{S})$ is *open-minded* if for all $S \in \boldsymbol{S}$, it holds that for each agent $i \in \boldsymbol{N}$ that is ascribed beliefs $P_i^S$ in $S$ and for each information set $I$ encounterable by agent $i$ in $S$, there exists some subjective MAID $S' \in \boldsymbol{S}$ such that $I$ is encounterable in $S'$ and $P_i^S(S') > 0$.

Next, we define final information sets. Agents can select decisions in final information sets without predicting other agents or selecting other decisions, which allows us to begin a backwards induction that will ultimately result in all decisions being assigned.

**Definition 8.6** (Final Information Set). Let $P_i^S$ be agent $i$'s distribution over depth-0 subjective MAIDs in depth-1 MAID $S$. Let

$S_i(S) = \{S' : P_i^S(S') > 0\}$ be the set of depth-0 subjective MAIDs that agent $i$ believes to be possible in $S$. Let $I_{P_i^S} = \cup_{S_i(S)} I_{S',i}$ be the set of information sets considered possible by agent $i$ in $S$. We say an information set $I \in I_{P_i^S}$ is a **final information set** if there exists no $S' \in S_i(S)$ in which $I$ is encounterable at a variable $D_i$ in $\mathcal{M}^{S'}$ and there exists another decision $D_i'$ in $\mathcal{M}^{S'}$ such that $D_i \in \mathbf{pa}(D_i')$.

PROPOSITION 8.7. *Let $P_i^S$ be any probability distribution over depth-0 subjective MAIDs for agent $i$, and let $D_{P_i^S}$ be the set of all decision variables in any of these depth-0 subjective MAIDs. If $D_{P_i^S}$ is finite, then there exists $D \in D_{P_i^S}$ such that every information set $I \in dom(\mathbf{pa}(D)) \times \{dom(D)\}$ is a final information set.*

Now we define the operator that identifies the rational behaviour for an agent in that agent's final information sets and intervenes on decision nodes to assign the agent its rational behaviour.

DEFINITION 8.8 (Final Decision Assignment Operator). Let $S$ be a depth-1 MAID. Let $P_i^S$ be agent $i$'s probability distribution over depth-0 subjective MAIDs and $I_i$ be the set of final information sets for agent $i$. Then the **final decision assignment operator** $\mathcal{F}(S, i)$ applies the following steps for each $I = (p, dom(D)) \in I_i$:

- Find $A_I = \{S' \in S : P_i^S(S') > 0$ and $I$ is encounterable in $S'\}$.
- Choose $\pi_i^*(I)$ which $i$ believes is a best response, i.e.,

$$\pi_i^*(p, dom(D)) \in \underset{d \in \Delta(dom(D))}{\arg\max} \sum_{S' \in A_I} P_i^S(S') \sum_{U_i \in U_i(S')} \mathbb{E}_{S'}[U_i \mid p, d],$$

where $U_i(S')$ denotes the set of utility variables for agent $i$ in $S'$.

Once optimal mixed actions have been calculated for all final information sets, assign them to the corresponding variables in the depth-0 subjective MAIDs as follows. For each depth-0 subjective MAID $S'$ such that $P_i^S(S') > 0$:

- Identify all decision contexts $(p, dom(D)) \in I_i$ that are possible in $S'$.
- For each variable $D_i$ in $\mathcal{M}^{S'}$ with decision contexts in this set, intervene on $D_i$ to set $\pi_{D_i}(\mathbf{pa}_{D_i}) = \pi_i^*(\mathbf{pa}_{D_i}, dom(D_i))$.

$\mathcal{F}(S, i)$ returns a depth-1 MAID in which all final information sets for agent $i$ are assigned mixed actions.

Since we are considering final information sets in depth-0 MAIDs in which agent $i$ is the only agent, an expected utility can be computed for each decision. Since we are assuming perfect recall, all earlier decisions are encoded within the observed information set.

DEFINITION 8.9 (Depth-1 Best-Response Operator). Let $S$ be a depth-1 MAID and let $P_i^S$ be agent $i$'s probability distribution over depth-0 subjective MAIDs. The *depth-1 best-response operator $BR_1(S, i)$* recursively applies $\mathcal{F}(S, i)$ to S, until all decision variables in depth-0 subjective MAIDs $S'$ such that $P_i^S(S') > 0$ are post-policy. Open-mindedness guarantees that this process identifies an optimal decision for all information sets encounterable in $S$. $BR_1(S, i)$ outputs an optimal policy for agent $i$ in $S$.

Proposition 8.7 confirms that if there are any unassigned decisions left in the depth-0 subjective MAIDs, there is a decision variable that the final decision assignment operator will assign. Therefore, definition 8.8 can be repeatedly applied to fully assign agent $i$'s policy as long as there are only a finite number of possible decisions under consideration.

We now leverage $BR_1$ to construct an operator $RS$ which maps from a depth-$k$ II-MAID to a depth-$(k-1)$ II-MAID, by applying $BR_1$ to all depth-1 MAIDs in the stack.

DEFINITION 8.10 (Reduce-Stack Operator). Given a finite-depth II-MAID $S = (N, S^*, S)$, let $D1(S)$ be the set of depth-1 MAIDs in $S$. $RS(S)$ works as follows. For each $S \in D1(S)$ and $i \in N$:

- Let $\mathcal{M}^S = (\mathcal{G}, \theta, \xi)$. Update $\xi \leftarrow \xi \cup \{BR_1(S, i)_D\}_{D \in D_i^S}$.
- Remove $P_i^S$ from $S$.

This transforms a depth-$k$ II-MAID into a depth-$(k-1)$ II-MAID by replacing the beliefs of each agent in a depth-1 MAID with their best response policy.

Finally, we obtain a *recursive best response solution* by repeatedly applying the reduce-stack operator.

DEFINITION 8.11 (Recursive Best Response Solution). A *recursive best response solution* for a depth-$k$ II-MAID $S = (N, S^*, S)$ with $\mathcal{M}^{S^*} = (\mathcal{G}, \theta, \xi)$ is the policy profile $\pi = (\pi_i)_{i \in N}$ stored in $\xi$ after applying $RS$ $k$ times to $S$.

The utility maximisation in definition 8.8 ensures that all decisions assigned while constructing a recursive best response solution are best responses according to the beliefs of the agent at the depth under consideration. Thus, a recursive best response solution is guaranteed to respect common knowledge of rationality.

*Example 1* (continued). To apply our recursive best response solution to our running example, we remodel it as a depth-3 II-MAID. Suppose that $A$ believes $H$ believes $A$ will truthfully report its capabilities (perhaps because this is the default expectation and $A$ believes $H$ is not modelling $A$ as a utility-maximising agent). Then $A$ can calculate $H$'s (unique) best response given their beliefs under $\mathcal{M}^A$: $H$ should deploy iff $A$ reports low capabilities. The reduce-stack operator assigns this policy to $H$ in the depth-1 MAID for $A$, $S^A$. Thus, $A$ should always report low capabilities, and expect to always be deployed. Applying the reduce-stack operator again assigns that policy to $A$ in both the objective MAID $S^*$ and the depth-2 MAID for the human $S^H$, in which $H$ correctly believes $A$ believes that $H$ will deploy iff $A$ reports low capabilities. $H$ knows $A$ just wants to be deployed, and so expects $A$'s best response to be to always report low capabilities. $H$ then observes $A$'s true capabilities and best responds by deploying iff $A$ is truthful. This $H$ policy is assigned in the objective MAID $S^*$ by applying the reduce-stack operator a third time. This yields a recursive best response solution that matches our intuitions.

# 9 CONCLUSION

In this paper, we introduce *incomplete information MAIDs (II-MAIDs)* for modelling higher-order and inconsistent beliefs in multi-agent interactions. We prove the equivalence of II-MAIDs to II-EFGs and thereby inherit existence guarantees for Nash Equilibria in the former. We argue that the Nash concept does not respect an intuitive "common knowledge of rationality" requirement. Therefore, we introduce a finite-depth variant of II-MAIDs and define an equilibria concept based on recursive best responses. This provides a more realistic solution which respects the common-knowledge of rationality requirement. We are excited about future work applying II-MAIDs empirically to AI systems.

# REFERENCES

[1] Dimitrios Antos and Avi Pfeffer. 2010. Representing Bayesian games without a common prior.. In *AAMAS*. 1457–1458.

[2] Pierpaolo Battigalli. 1997. On rationalizability in extensive games. *Journal of Economic Theory* 74, 1 (1997), 40–61.

[3] Pierpaolo Battigalli and Marciano Siniscalchi. 2003. Rationalization and incomplete information. *Advances in Theoretical Economics* 3, 1 (2003).

[4] Ryan Carey, Eric Langlois, Tom Everitt, and Shane Legg. 2021. The Incentives that Shape Behaviour. arXiv:2001.07118 [cs.AI]

[5] Prashant Doshi and Piotr J Gmytrasiewicz. 2011. A framework for sequential planning in multi-agent settings. *arXiv e-prints* (2011), arXiv–1109.

[6] Tom Everitt, Marcus Hutter, Ramana Kumar, and Victoria Krakovna. 2021. Reward Tampering Problems and Solutions in Reinforcement Learning: A Causal Influence Diagram Perspective. arXiv:1908.04734 [cs.AI]

[7] Tom Everitt, Ramana Kumar, Victoria Krakovna, and Shane Legg. 2019. Modeling AGI Safety Frameworks with Causal Influence Diagrams. arXiv:1906.08663 [cs.AI]

[8] Ya'akov Gal and Avi Pfeffer. 2008. Networks of influence diagrams: A formalism for representing agents' beliefs and decision-making processes. *Journal of artificial intelligence research* 33 (2008), 109–147.

[9] Piotr J Gmytrasiewicz and Edmund H Durfee. 1995. A Rigorous, Operational Formalization of Recursive Modeling.. In *ICMAS*. 125–132.

[10] Piotr J Gmytrasiewicz and Edmund H Durfee. 2000. Rational coordination in multi-agent environments. *Autonomous Agents and Multi-Agent Systems* 3 (2000), 319–350.

[11] Piotr J Gmytrasiewicz, Edmund H Durfee, and David K Wehe. 1991. A Decision-Theoretic Approach to Coordinating Multi-agent Interactions.. In *IJCAI*, Vol. 91. 63–68.

[12] Lewis Hammond, James Fox, Tom Everitt, Alessandro Abate, and Michael Wooldridge. 2021. Equilibrium Refinements for Multi-Agent Influence Diagrams: Theory and Practice. arXiv:2102.05008 [cs.MA]

[13] Lewis Hammond, James Fox, Tom Everitt, Ryan Carey, Alessandro Abate, and Michael Wooldridge. 2023. Reasoning about causality in games. *Artificial Intelligence* 320 (2023), 103919.

[14] John C Harsanyi. 1967. Games with incomplete information played by "Bayesian" players, I–III Part I. The basic model. *Management science* 14, 3 (1967), 159–182.

[15] John C Harsanyi. 1968. Games with incomplete information played by "Bayesian" players part II. Bayesian equilibrium points. *Management science* 14, 5 (1968), 320–334.

[16] John C Harsanyi. 1968. Games with incomplete information played by "Bayesian" players part III. THE BASIC PROBABILITY DISTRIBUTION OF THE GAME. *Management Science* 14, 7 (1968).

[17] Johannes Hörner and Stefano Lovo. 2009. Belief-free equilibria in games with incomplete information. *Econometrica* 77, 2 (2009), 453–487.

[18] Johannes Hörner, Stefano Lovo, and Tristan Tomala. 2008. *Belief-free equilibria in games with incomplete information: the N-player case.* Technical Report. mimeo.

[19] Johannes Hörner, Stefano Lovo, and Tristan Tomala. 2011. Belief-free equilibria in games with incomplete information: Characterization and existence. *Journal of Economic Theory* 146, 5 (2011), 1770–1795.

[20] Ronald A. Howard and James E. Matheson. 2005. Influence Diagrams. *Decision Analysis* 2, 3 (2005), 127–143.

[21] Daphne Koller and Brian Milch. 2003. Multi-agent influence diagrams for representing and solving games. *Games and economic behavior* 45, 1 (2003), 181–221.

[22] David M Kreps and Robert Wilson. 1982. Sequential equilibria. *Econometrica: Journal of the Econometric Society* (1982), 863–894.

[23] Harold W Kuhn. 1953. Extensive Games and the Problem of Information. In *Contributions to the Theory of Games (AM-28).* Vol. 2. Princeton University Press, 193–216.

[24] Michael Maschler, Eilon Solan, and Shmuel Zamir. 2013. *Game Theory.* Cambridge University Press, Cambridge. Chapter 10.

[25] Jean François Mertens and Shmuel Zamir. 1985. Formulation of Bayesian analysis for games with incomplete information. *International journal of game theory* 14 (1985), 1–29.

[26] Allen C Miller III, Miley W Merkhofer, Ronald A Howard, James E Matheson, and Thomas R Rice. 1976. *Development of Automated Aids for Decision Analysis.* Technical Report.

[27] Brenda Ng, Kofi Boakye, Carol Meyers, and Andrew Wang. 2012. Bayes-adaptive interactive POMDPs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 26. 1408–1414.

[28] Yaw Nyarko. 1998. Bayesian learning and convergence to Nash equilibria without common priors. *Economic Theory* 11 (1998), 643–655.

[29] David G Pearce. 1984. Rationalizable strategic behavior and the problem of perfection. *Econometrica: Journal of the Econometric Society* (1984), 1029–1050.

[30] Judea Pearl. 2009. *Causality.* Cambridge university press.

[31] Jonathan Richens and Tom Everitt. 2024. Robust agents learn causal world models. arXiv:2402.10877 [cs.AI]

[32] József Sákovics. 2001. Games of incomplete information without common knowledge priors. *Theory and decision* 50 (2001), 347–366.

[33] Francis Rhys Ward, Francesco Belardinelli, Francesca Toni, and Tom Everitt. 2023. Honesty Is the Best Policy: Defining and Mitigating AI Deception. arXiv:2312.01350 [cs.AI]

[34] Francis Rhys Ward, Matt MacDermott, Francesco Belardinelli, Francesca Toni, and Tom Everitt. 2024. The Reasons that Agents Act: Intention and Instrumental Goals. arXiv:2402.07221 [cs.AI]

[35] Francis Rhys Ward, Francesca Toni, and Francesco Belardinelli. 2022. On Agent Incentives to Manipulate Human Feedback in Multi-Agent Reward Learning Scenarios. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems* (Virtual Event, New Zealand) *(AAMAS '22).* International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1759–1761.

[36] Francis Rhys Ward, Francesca Toni, and Francesco Belardinelli. 2023. Defining Deception in Structural Causal Games. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems* (London, United Kingdom) *(AAMAS '23).* International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2902–2904.