# The Role of Cognitive Biases and Theory of Mind in Human Coordination

**Pedro Alexandre Caetano Lopes Ferreira**
pedro.lopes.ferreira@tecnico.ulisboa.pt

## Abstract

Theory of mind models the way humans handle social situations by managing internal models of others in order to predict their actions. A particular recursive scheme that captures such behavior is the level-$k$ model where behavior is hierarchically arranged in increasing sophistication level such that a level-$k$ behavior is the best response against a level-$(k-1)$ policy of the other agent. Expected utility theory (EUT) is commonly used in decision-making models together with the level-$k$ model, as a theory of value. However, EUT is known to not capture biases due to human interpretations of risk associated with decisions and their value. In this work, we develop a framework that enables us to consider level-$k$ recursive theory of mind where biases and risk associated with decisions are captured by cumulative prospect theory (CPT), a risk-sensitive theory of value. We show that risk-sensitive agents (i.e. using CPT) are better at coordinating, with increasing sophistication level $k$, compared to risk-neutral agents (i.e. using EUT). We find that both settings suggest that the reason why humans are good at coordination may stem from the fact that we are cognitively biased to do so. As a consequence, we might be able to leverage the proposed framework to equip artificial intelligence decision-making schemes with human-like reasoning.

## Introduction

Coordination emerges when two or more agents choose actions such that individuals payoffs are maximized as a collective. In society, coordination is of paramount importance; manufacturing companies must coordinate on which units of measurement to use, drivers must choose which lane to drive on and orators in political debates must understand when it is their turn to speak so that useful discourse may be allowed to happen.

Although coordination comes naturally to humans and emerges in societies we live in, the mechanisms through which humans attain coordination are still not fully understood. Similarly, the need for self-organized collective action is becoming increasingly important in engineering applications involving multiple robots or software agents. Artificial entities are often expected to uphold cooperative activities in the absence of pre-defined guidelines (Wooldridge 2009; Stone et al. 2010; Barrett and Stone 2015; Bonabeau et al. 1999). Moreover, as we move towards a society where both humans and machines need to interact with each other and achieve coordination, we do not only need to unveil such mechanisms, but understand how to foster collective action in populations comprising humans and artificial entities (Paiva, Santos, and Santos 2018; Santos et al. 2019; Rahwan et al. 2019). Particularly, as a stepping stone to create computational agents capable of self-organized prosocial behaviors, we may imbue them with some of the features known to promote cooperation among humans, such as reciprocity, empathy, morality or theory of mind, among others (Rand and Nowak 2013).

The emergence of coordination in systems of human agents can be conveniently studied through the lens of game theory and coordination games, normal-form games with two or more pure Nash equilibria (Cooper 1999). The game of stag-hunt, a two-agent coordination game in which two agents (hunters) must choose between a safe with low payoff outcome (hunting a hare) and a risky with high payoff outcome (hunting a stag), has been one of the most studied coordination games due to the strong analogy between the theoretical setting and real-world conflicts (Skyrms 2004).

During human interactions, humans make decisions using several cognitive mechanisms. On one hand, in psychology, *theory of mind* is defined as one's ability to attribute mental states (e.g. beliefs, knowledge, goals) to others and to realize those mental states may be different from one's own (Premack and Woodruff 1978). The level-$k$ model is a model of theory of mind in which agents first assume a stereotyped behavior and progressively make use of previous behaviors to calculate more sophisticated ones in a recursive fashion (Stahl 1993). Truncating the level of recursion to a fixed level is one way to simulate the so-called *bounded rationality*.

On the other hand, in decision theory, agents (i.e. humans) determine the values of their actions and choose whichever action yields the highest payoff. The determination of that value is often done using *expected utility theory* (EUT). Expected utility theory provides a simple and parsimonious model to determine expected payoffs of uncertain outcomes (also referred to as *lotteries* or *prospects* in the literature) but it is regarded as a prescriptive model of decision-making and is not viewed as a good descriptor of how humans assign value to uncertain outcomes (Kahneman and Tversky 1979).

The assumptions an agent makes to be able to "predict" what the other agents will choose – in order to be able to calculate the payoff of his actions – are based on the rationality of preferences. This axiomatic of choice, first proposed in (von Neumann and Morgenstern 1944) in the form of four axioms, is the basis for expected utility theory and has been observed not to be a very good descriptor of how people make decisions. The Allais paradox (Allais 1979) and the Ellsberg paradox (Ellsberg 1961) are examples of how people break these axioms on a regular basis.

In contrast, *cumulative prospect theory* (CPT) attempts to model the way people make decisions by framing prospects relative to a reference point (Tversky and Kahneman 1992). Specifically, above the reference point the outcomes are viewed as gains and below as losses. The utility associated with outcomes are determined by different utility functions, both of which showing diminishing marginal returns but the one associated with losses is steeper, simulating loss aversion by weighting losses more heavily. Furthermore, in CPT, the perceived probability of extreme but rare events is overestimated. This "distorted" perception of prospects is another way of reproducing agents with human-like rationality. Thus, applying CPT to game-theoretical models where agents mimic human decisions may yield more human-like behaviors.

Due to the superior ability of CPT to describe human decisions and the importance of the development of a theory of mind, we are motivated to study coordination between agents equipped with both CPT and bounded rationality (up to a level-$k$ recursion). We create agents with theory of mind and cognitive bias on risk-sensitivity, and study how they coordinate in normal-form and Markov stag-hunt games. Specifically, we aim at answering the following main questions:

1. Can cognitive biases concerning risk promote coordination?
2. Can increasingly sophisticated levels of theory of mind promote coordination?

We show that both of these questions are answered affirmatively. To do so, we assess the emergence of coordination by combining both theories, compare with previous results using EUT, and analyze coordination with increasing $k$. This indicates that, while these mechanisms often create sub-optimal individual behavior, they greatly facilitate collective action among humans.

This suggests that the reason why humans are good at coordination may stem from the fact that we are cognitively biased to do so. As a consequence, we might be able to leverage the proposed framework to equip artificial intelligence decision-making schemes with human-like reasoning and steer self-organized coordination among artificial entities.

## Related Work

CPT is an improvement over the previously developed *prospect theory* (PT) (Kahneman and Tversky 1979). PT non-linearly transforms individual probabilities rather than cumulative probabilities but is known to violate first-order stochastic dominance (Tversky and Kahneman 1992). Both PT and CPT have been used to explain human behavior in many scenarios. For instance, it has been shown that private bankers and fund managers behave according to PT and violate EUT (Abdellaoui, Bleichrodt, and Kammoun 2013). Also, inexperienced consumers in a well-functioning marketplace behave according to PT while those with more experience behave according to more recent economic theories, showing that learning plays an important role in risk perception (List 2004). A study using a model inspired by PT helped explain properties seen in asset prices in an economy where investors derive direct utility not only from consumption but also from fluctuations in the value of their financial wealth (Cxvi et al. 2001). The presence of reference points was observed in a large database of firms, together with risk-seeking behavior for firms below their reference point and risk-averse behavior for firms above their reference point, and risk-seeking behavior was more intense than risk-averse behavior (Fiegenbaum 1990). Prospect theory was used to explain why political actors pursue risky reforms, in spite of political resistance that counteracts change (Vis and Van Kersbergen 2007). The feasibility of a basic income guarantee was studied using PT (Pech 2010). Repeated Stackelberg security games were studied with PT-agents (Kar et al. 2015). CPT is a behavioral alternative to EUT, and empirical evidence suggests that CPT is a better model of human decision-making than EUT. Decision-making models achieved state of the art performance on human judgment datasets by creating neural networks with human-like inference bias by pre-training them with synthetic data generated by CPT (Peterson et al. 2019).

In a more abstract setting, coordination was shown to be promoted for increasing sophistication levels with agents determining value using EUT and equipped with a level-$k$ model (Yoshida, Dolan, and Friston 2008). However, it was assumed that agents made decisions in a sequential fashion and, to the author's best knowledge, no work has been done to study coordination in simultaneous decision-making scenarios with a CPT objective function and level-$k$ model. Cumulative prospect theory has been applied to normal-form games (Metzger and Rieger 2019) and CPT objective functions have been shown to have similar dynamic programming equations to EUT in Markov decision processes (Lin and Marcus 2013; Lin 2013), but no effort has been made to understand coordination with agents measuring value using CPT from a game-theoretic perspective.

Lastly, our approach is also related with the broad literature on intention recognition, opponent modelling, and models that aim at predicting the opponents' sequence of actions through machine learning techniques (Foerster et al. 2018; Mealing and Shapiro 2013; Rabinowitz et al. 2018).

## Model

### Determining the Value of Outcomes

Let $R$ be a discrete random variable with distribution over the set of outcomes $\{r_j\}_{j=1}^{m}$ (sorted in increasing order), and $p_j = \mathbb{P}(R = r_j)$ be the probability that outcome $r_j$

occurs. The value of $R$ under EUT, given utility function $u : \mathbb{R} \to \mathbb{R}$, is the expected value of its utility given by:

$$V^{\text{EUT}}(R) = \sum_{j=1}^{m} u(r_j)p_j. \tag{1}$$

To determine the value under CPT, let us first define the following auxiliary functions:

$$\psi^+(r_j) = w^+(P(R \geq r_j)) - w^+(P(R > r_j)),$$
$$\psi^-(r_j) = w^-(P(R \leq r_j)) - w^-(P(R < r_j)),$$

where $w^+ : [0,1] \to [0,1]$ and $w^- : [0,1] \to [0,1]$ are probability weighting functions for gains and losses, respectively. In contrast, the value of $R$ under CPT is given by

$$\begin{aligned} V^{\text{CPT}}(R) = &\sum_{j \geq k} u^+(r_j)\psi^+(r_j) \\ &+ \sum_{j<k} u^-(r_j)\psi^-(r_j), \end{aligned} \tag{2}$$

where $k \in \mathbb{Z}$ is the reference point, which splits the prospect into gains and losses, and $u^+$ and $u^-$ are the utility functions for gains and losses, respectively. CPT is a generalization of EUT in the sense that $V^{\text{CPT}} = V^{\text{EUT}}$ if $u^+(x) = u^-(x) = u(x)$, for all $x \in \mathbb{R}$ and $w^+(p) = w^-(p) = p$, for all $p \in [0,1]$.

## The Normal-form Game

A normal-form game is composed of a set of agents (or players) $N = \{1, ..., n\}$ and, for each agent $i$, a set of actions $A_i$ and a utility function $u_i : A_1 \times ... \times A_n \to \mathbb{R}$. A play of a normal-form game consists in all agents simultaneously (or, equivalently, without knowing the others' chosen actions) choosing an action from their respective action sets and receiving their respective utilities based on the chosen actions.

The behavior of each agent $i$ is represented as a policy denoted by $\pi_i$ ($\pi_i : A_i \to [0,1]$ and $\sum_{a \in A_i} \pi_i(a) = 1$). For instance, if agent 1 decides to choose an action $a_1$ (out of two possible actions, i.e., $\{a_1, a_2\}$) with probability $0.2$, then his policy is $\pi_1(a_1) = 0.2$ and $\pi_1(a_2) = 0.8$. A joint policy $\boldsymbol{\pi}$ is the collection of the chosen policies of all agents. In general, to solve a normal-form game one needs to find the Nash equilibria (NEs) (i.e. joint policies where no agent will receive a higher payoff by unilaterally changing its action).

## The Markov Game

Repeated games allow for the study of the interaction between immediate gains and long-term incentives. We provide a brief review of the *Markov decision process* (MDP) – a single-agent decision model – and explain how they relate to the Markov game – a model of collective decision-making.

A MDP is a model of discrete-time decision-making in a stochastic environment (Howard 1960). Formally, it consists of a set of states $\mathcal{S}$, a set of actions $\mathcal{A}$, a probability transition function $p : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$, and a reward function $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$. In general, in this framework, an agent is attempting to find a policy $\pi : \mathcal{S} \times \mathcal{A} \to [0,1]$, that maximizes some value functional $V(s, \pi)$, for all states $s \in \mathcal{S}$.

To determine the policies of a MDP, for the standard infinite-horizon EUT value functional, we can use dynamic programming, given a discount factor $\beta \in (0,1)$ that accounts for the importance of short-term versus long-term rewards. Similarly, dynamic programming can be used when the infinite-horizon CPT value functional is considered (Lin and Marcus 2013; Lin 2013).

To determine the policies of a MDP, for the standard infinite-horizon EUT value functional, we can use dynamic programming. This technique makes use of a discount factor $\beta \in (0,1)$ that accounts for the importance of short-term versus long-term rewards. Similarly, dynamic programming can be used when the infinite-horizon CPT value functional is considered (Lin and Marcus 2013; Lin 2013).

A Markov game is a generalization of a MDP to accommodate several agents (Shapley 1953). The main difference is that, in Markov games, each agent must take into account the possible policies of other agents whilst maximizing his own payoff at each state. Formally, let $n > 1$ be the number of agents in a Markov game with joint action space $\boldsymbol{\mathcal{A}} = \mathcal{A}_1 \times ... \times \mathcal{A}_n$ (where $\mathcal{A}_i$ is the action space of agent $i$ and $\mathcal{A}_{-i}$ is the joint action space of all agents except agent $i$) and joint state space $\boldsymbol{\mathcal{S}} = \mathcal{S}_1 \times ... \times \mathcal{S}_n$ (where $\mathcal{S}_i$ is the state space of agent $i$) with stochastic dynamics prescribed by the transition function $P^{\boldsymbol{a}}_{\boldsymbol{s}}(\cdot) = \mathbb{P}(\cdot | \boldsymbol{s}_t = \boldsymbol{s}, \boldsymbol{a}_t = \boldsymbol{a})$, $(s_1, ..., s_n) = \boldsymbol{s} \in \boldsymbol{\mathcal{S}}$, $(a_1, ..., a_n) = \boldsymbol{a} \in \boldsymbol{\mathcal{A}}$.

The CPT-value that agent $i$ places on a joint state $(s_1, ..., s_n) = \boldsymbol{s} \in \boldsymbol{\mathcal{S}}$, given a joint policy $\boldsymbol{\pi} = (\pi_i, \boldsymbol{\pi}_{-i})$, can be obtained by generalizing the MDP result from (Lin and Marcus 2013) to this Markov game via successive iterations of

$$\begin{aligned} V_i^{\pi_i, \boldsymbol{\pi}_{-i}}(\boldsymbol{s}) = &\int_0^\infty w_i^+ \left( \sum_{a_i \in \mathcal{A}_i} P^{a_i, \boldsymbol{\pi}_{-i}}_{i, \boldsymbol{s}, +}(\epsilon) \pi_i(a_i | \boldsymbol{s}) \right) d\epsilon \\ &- \int_0^\infty w_i^- \left( \sum_{a_i \in \mathcal{A}_i} P^{a_i, \boldsymbol{\pi}_{-i}}_{i, \boldsymbol{s}, -}(\epsilon) \pi_i(a_i | \boldsymbol{s}) \right) d\epsilon, \end{aligned} \tag{3}$$

where

$$P^{a_i, \boldsymbol{\pi}_{-i}}_{i, \boldsymbol{s}, +}(\epsilon) = \sum_{\boldsymbol{a}_{-i} \in \boldsymbol{\mathcal{A}}_{-i}(\boldsymbol{s})} P^{a_i, \boldsymbol{a}_{-i}}_{i, \boldsymbol{s}, +}(\epsilon) \boldsymbol{\pi}_{-i}(\boldsymbol{a}_{-i} | \boldsymbol{s}),$$
$$P^{a_i, \boldsymbol{\pi}_{-i}}_{i, \boldsymbol{s}, -}(\epsilon) = \sum_{\boldsymbol{a}_{-i} \in \boldsymbol{\mathcal{A}}_{-i}(\boldsymbol{s})} P^{a_i, \boldsymbol{a}_{-i}}_{i, \boldsymbol{s}, -}(\epsilon) \boldsymbol{\pi}_{-i}(\boldsymbol{a}_{-i} | \boldsymbol{s}),$$

and

$$P^{a_i, \boldsymbol{a}_{-i}}_{i, \boldsymbol{s}, +}(\epsilon) = P^{\boldsymbol{a}}_{\boldsymbol{s}}(u_i^+((r_i(\boldsymbol{s}) + \beta_i V_i^{\pi_i, \boldsymbol{\pi}_{-i}}(\boldsymbol{S}) - b_i)_+) > \epsilon),$$
$$P^{a_i, \boldsymbol{a}_{-i}}_{i, \boldsymbol{s}, -}(\epsilon) = P^{\boldsymbol{a}}_{\boldsymbol{s}}(u_i^-((r_i(\boldsymbol{s}) + \beta_i V_i^{\pi_i, \boldsymbol{\pi}_{-i}}(\boldsymbol{S}) - b_i)_-) > \epsilon).$$

Each agent $i$ tries to maximize his value $V_i$ by choosing the optimal policy $\pi_i$ given the joint policy of every other agent $\boldsymbol{\pi}_{-i}$, i.e.,

$$\pi_i^*(\boldsymbol{s}) = \underset{\pi_i}{\text{argmax}}\, V_i^{\pi_i, \boldsymbol{\pi}_{-i}}(\boldsymbol{s}), \forall \boldsymbol{s} \in \boldsymbol{\mathcal{S}}. \tag{4}$$

However, to obtain the solution to (4), agent $i$ requires knowledge of $\boldsymbol{\pi}_{-i}$. Therefore, instead of assuming rational

| | | 2 | |
|---|---|---|---|
| | | $S$ | $H$ |
| 1 | $S$ | 5,5 | 0,1 |
| | $H$ | 1,0 | 1,1 |

Table 1: Utility functions of the stag hunt. Specifically, agent 1's choices are cast in rows and agent 2's choices are cast in columns. The utility given to agent 1 and agent 2 are the first and second numbers, respectively, of a given cell.

agents, we will consider bounded rationality (i.e. the level-$k$ model) (Stahl 1993). In a two-agent setting, agent 1 assumes a stereotyped policy $\pi_2^{(0)}$, which describes the behavior of an independently operating agent 2. Similarly, agent 2 assumes a stereotyped policy $\pi_1^{(0)}$, which describes the behavior of an independently operating agent 1. With this "new" information, both agents can determine their first-order policies $\pi_1^{(1)}$ and $\pi_2^{(1)}$. However, each of them can also assume that the other did the same and is, therefore, operating under a first-order policy. Subsequently, they should operate under a second-order policy, and so forth. Thus, if such reasoning is used $k$ times, we obtain the following recursive scheme (each line is an iteration and the left and right hand sides represent agent 1 and 2 decisions, respectively):

$$\begin{array}{ccc} \pi_2^{(0)} & \pi_1^{(0)} & \\ \downarrow & \downarrow & \\ \operatorname*{argmax}_{\pi_1} V_1^{\pi_1,\pi_2^{(0)}} & = \pi_1^{(1)}\pi_2^{(1)} = & \operatorname*{argmax}_{\pi_2} V_2^{\pi_1^{(0)},\pi_2}, \\ \vdots & \vdots & \\ \operatorname*{argmax}_{\pi_1} V_1^{\pi_1,\pi_2^{(k-1)}} & = \pi_1^{(k)}\pi_2^{(k)} = & \operatorname*{argmax}_{\pi_2} V_2^{\pi_1^{(k-1)},\pi_2}. \end{array}$$

## Experiments

The stag-hunt game describes the interaction between individuals when they are given a choice between a safe but low payoff outcome and a risky but high payoff outcome (Skyrms 2004). In what follows, we analyze the stag-hunt game, both as a normal-form or a Markov game. We consider the case where these games are played by agents that use CPT and that are equipped with a level-$k$ bounded rationality.

### Normal-form Stag-hunt

The stag-hunt game is a normal-form game with two agents $N = \{1, 2\}$, each with two actions $A_1 = A_2 = \{S, H\}$, where $S$ stands for *Stag* and $H$ stands for *Hare*. The utility functions are symmetric (i.e., $u_1(a_1, a_2) = u_2(a_1, a_2)$) and, throughout this paper, we shall consider the utility functions presented in Table 1.

The Nash equilibria of a stag-hunt game between EUT- and CPT-agents are a useful tool to analyze the effects of CPT-based risk perception. Next, let us assume that utility functions are identical and that the only difference between EUT and CPT is the way agents perceive likelihoods which are captured by the probabilities of an outcome. If, from the perspective of one agent, the other agent will choose $S$ with probability $p$, then the values of his actions, under EUT and CPT, are given by

$$\begin{aligned} V^{\text{EUT}}(S) &= 5p, \quad V^{\text{EUT}}(H) = 1, \text{ and} \\ V^{\text{CPT}}(S) &= 5w(p), \quad V^{\text{CPT}}(H) = 1. \end{aligned} \tag{5}$$

For EUT and CPT, the Nash equilibria correspond to the probabilities $p^{\text{EUT}}$ and $p^{\text{CPT}}$ that make $V^{\text{M}}(S) = V^{\text{M}}(H)$ for $M \in \{\text{EUT}, \text{CPT}\}$, respectively. Also, we consider $w(x) = \exp\{-0.5(-\log(x))^{0.9}\}$, so we get $p^{\text{EUT}} = 0.2$ and $p^{\text{CPT}} \approx 0.028$.

Consequently, it readily follows that CPT dramatically increases coordination in the stag-hunt game, since both agents choose the same action more often. In other words, whereas two hunters hunting a stag yields the largest reward, the risk of hunting a stag alone is overshadowed by the safety of hunting a hare. However, the average sum of rewards for both agents (i.e. the hunters) does not decrease significantly since the expected values are as follows:

$$\begin{aligned} &\mathbb{E}_{a_1,a_2}[r_1^{\text{EUT}}(a_1, a_2) + r_2^{\text{EUT}}(a_1, a_2)] = 2, \\ &\mathbb{E}_{a_1,a_2}[r_1^{\text{CPT}}(a_1, a_2) + r_2^{\text{CPT}}(a_1, a_2)] \approx 1.95. \end{aligned} \tag{6}$$

Thus, an increase in coordination reduces the likelihood of either agent choosing to hunt stags alone and, consequently, getting a reward of zero.

### Markov Stag-hunt

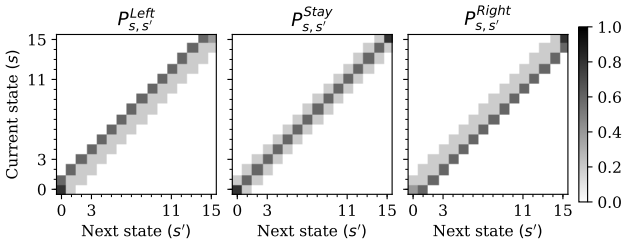
To compare the coordination of EUT-agents and CPT-agents (i.e., those agents using EUT and CPT, respectively), we set up a Markov game based on the stag-hunt game – similar to the stag-hunt in (Yoshida, Dolan, and Friston 2008). In this Markov game, two agents live in one of 16 states ($\mathcal{S}_1 = \mathcal{S}_2 = \{0, ..., 15\}$, such that the joint state space is $\boldsymbol{\mathcal{S}} = \mathcal{S}_1 \times \mathcal{S}_2$), starting in one of them at random. These 16 states represent 16 areas within a hunting region, on the bottom of a long canyon.

Each agent $i$ can move around in this canyon, by choosing an action $a_i$ from their action set $\mathcal{A}_i$ (with $\mathcal{A}_1 = \mathcal{A}_2 = \{Left, Stay, Right\}$). These actions have some probability to fail, in which case they may still go to in the desired direction, stay at the current location, or go in the opposite direction – see Figure 1a. In this canyon, only two of the 16 states have prey. Specifically, in state 3 there are hares and in state 11 there are stags. A single hunter can hunt hares alone, but coordination between the hunters is required in order to hunt stags – refer to Figure 1b for summary.

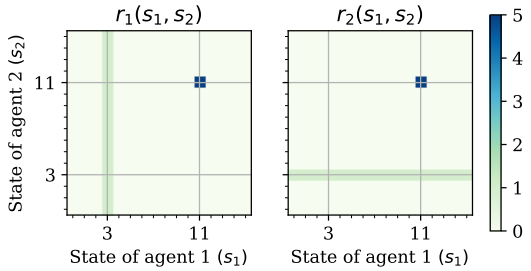
To create a simultaneous decision-making scenario, the dynamics in the joint state space and the individual transition probabilities (in Figure 1a) must be combined in a proper manner, i.e.,

$$P^{a_1,a_2} = \frac{I \otimes P^{a_1} + P^{a_2} \otimes I}{2}, \tag{7}$$

where the Kronecker product $\otimes$ ensures an action from agent 1 does not change the state of agent 2 and vice-versa. The av-




(a) Transition probabilities for each agent.




(b) Rewards as functions of the current state.

Figure 1: Markov stag-hunt transition probabilities and reward functions. (Top) Agents choose one of three actions (Left, Stay, or Right) and, depending on their current state, move to another state according to the respective transition probabilities. (Bottom) Agents receive a reward at each time step depending on their state and the state of the other agent. The hare state (state 3) can be obtained regardless of where the other agent is. This also allows us to model situations in which an agent can only obtain a big reward if the other agent is willing to coordinate with him. In our case, the stag state (i.e., state 11) has one such big, but difficult to obtain reward.

erage of this transformed agent transition probability function ensures the joint state space dynamics is independent of who acts first.

Notice that both EUT-agents and CPT-agents place increasingly more value on the stag state but the latter place substantially more value on the stag state – even in the lowest *sophistication levels* (i.e., low values of level-$k$ bounded rationality). From Figure 2, it readily follows that CPT-agents put higher value in the stag than EUT-agents, which suggests a better outcome for the former.

We can further study the behavior of these agents by capitalizing on the stationary distribution of the Markov chains that result from conditioning the Markov game's transition probability function on some policies $\pi_1^{(k_1)}$ and $\pi_2^{(k_2)}$. The conditioned transition probability function can be obtained as:

$$P^{\pi_1,\pi_2}_{\boldsymbol{s},\boldsymbol{s}'} = \sum_{a_1,a_2} P^{a_1,a_2}_{\boldsymbol{s},\boldsymbol{s}'} \pi_1(a_1|\boldsymbol{s})\pi_2(a_2|\boldsymbol{s}).$$

The stationary distribution $\rho(k)$ (which dictates the likelihood of finding the agents in a certain state) when agents use $\pi_1^{(k)}$ and $\pi_2^{(k)}$ (i.e. the same policy at $k$-level of bounded

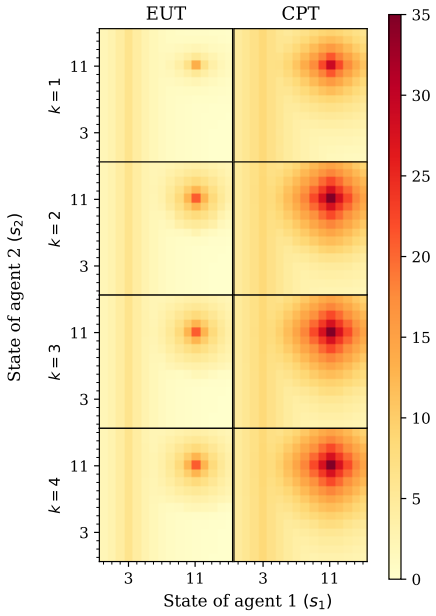



Figure 2: EUT and CPT values as functions of the agent states, for sophistication levels $k = 1, 2, 3$ and 4. We assumed reference points $b_1 = b_2 = 0$, discount factors $\beta_1 = \beta_2 = 0.9$, utility function $u(x) = x$ and weighting function $w(x) = x$ for EUT and $w(x) = e^{-0.5(-\log(x))^{0.9}}$ for CPT.

rationality) can be obtained via

$$\rho(k) = \rho(k)P^{\pi_1^{(k)},\pi_2^{(k)}}.$$

Stationary distributions allow us to easily compare dynamics of EUT- and CPT-agents – see Figure 3. Specifically, from Figure 3, we have that while both EUT- and CPT-agents eventually prefer state 11 (hunting stags) with increasing sophistication levels, it is interesting to notice that CPT-agents dramatically do so.

## Discussion

Our results show that CPT-agents in a normal-form stag-hunt game choose to coordinate to hunt hares with higher probability than EUT-agents. While hunting hares is sub-optimal, the total reward does not decrease substantially from the EUT-agents because the probability of hunting stags alone also decreases. Further, this suggests the risk aversion of human-like agents in a one-shot setting.

However, the conflict between short- and long-term rewards is of particular interest in domains where time is a relevant factor, and it is also in these domains where a theory of mind may prove useful. To that end, we studied how EUT- and CPT-agents coordinate in a Markov game version of stag-hunt inspired by (Yoshida, Dolan, and Friston 2008), where both types of agents were equipped with a level-$k$ theory of mind. Hence, predicting several (increasingly sophisticated) behaviors in the form of policies. Our results suggest that, with increasing $k$ (the sophistication level of

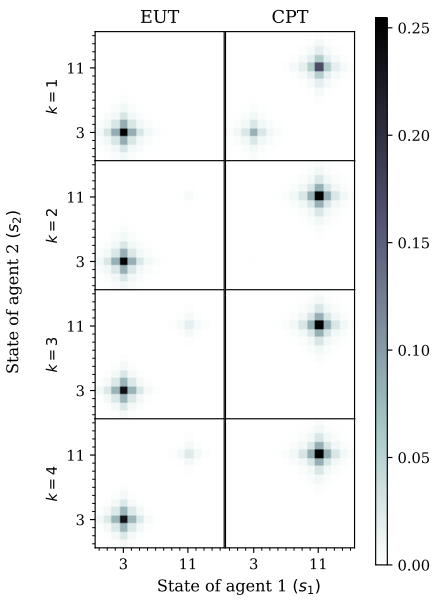



Figure 3: Stationary distributions of the resulting Markov chains obtained by conditioning the Markov game to increasingly sophisticated policies, $k = 1, 2, 3$ and $4$, for EUT- and CPT-agents. We assumed reference points $b_1 = b_2 = 0$, discount factors $\beta_1 = \beta_2 = 0.9$, utility function $u(x) = x$, and weighting function $w(x) = x$ for EUT and $w(x) = e^{-0.5(-\log(x))^{0.9}}$ for CPT.

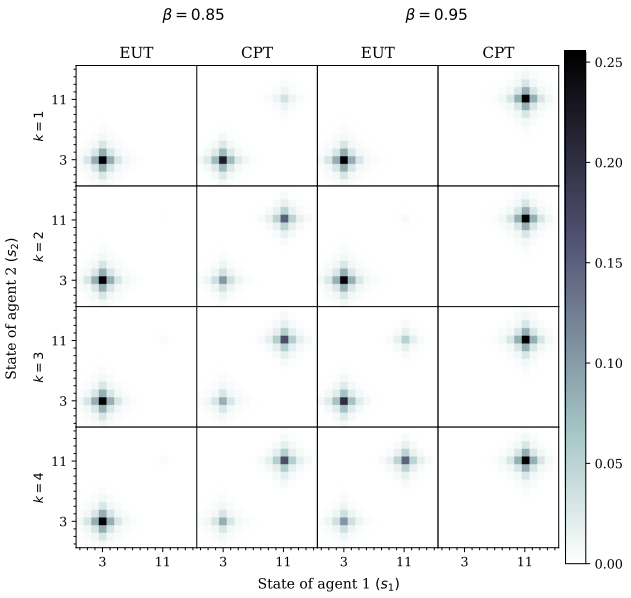



Figure 4: Stationary distributions of the resulting Markov chains obtained by conditioning the Markov game to increasingly sophisticated policies, $k = 1, 2, 3$ and $4$, for EUT-agents and CPT-agents. We assumed reference points $b_1 = b_2 = 0$, utility function $u(x) = x$ and weighting function $w(x) = x$ for EUT and $w(x) = e^{-0.5(-\log(x))^{0.9}}$ for CPT. (Left) Stationary distribution for EUT- and CPT-agents using discount factor $\beta = 0.85$. (Right) Stationary distribution for EUT- and CPT-agents using discount factor $\beta = 0.85$.

policies), *the CPT-agents coordinate faster and choose the optimal stag state, whereas EUT-agents fail to do so.*

This is a remarkable finding given that most multi-agent systems (MAS) use expected utility theory – likely due to the parsimonious mathematical setting it provides – and may be missing out on naturally occurring coordinating behaviors due to their focus on optimality.

To analyze the robustness to parametric choices in our setting, we further studied the sensitivity of the coordination to the parameters of the model. We specifically looked at the discount factor $\beta$ and the reference point $b$. Figure 4 provides evidence that increasing the discount factor (thus increasing the perceived "goodness" of long-term rewards) also increases coordination of both EUT- and CPT-agents, and that the latter still generate more coordination than EUT.

Additionally, we have considered different sophistication levels and reference points. This analysis is captured in Figure 5 where we show the stationary distribution of CPT-agents for different sophistication levels and reference points. It readily follows that higher reference points decrease coordination. In other words, hunting stags is perceived as a not-so-good solution when agents have a negative skewed view of the rewards; notice that CPT-agents with a higher reference point have a more bleak perception of rewards.

## Limitations

Unfortunately, obtaining the solution to the optimization problem defined in Equation 4 is often a daunting task as the numerical approaches suffer from (well known) instability issues for some initial configurations. Consequently, when this occurred, a different but valid initial configuration was selected at random until convergence.

Additionally, the weighting function $w(x) = e^{-0.5(-\log(x))^{0.9}}$ is both computationally expensive and its implementation has to be truncated as $x \to 0$. Therefore, a posynomial approximation $w(x) = 0.00231x^{0.05} + 0.00128x^{0.1} + 0.19578x^{0.35} + 0.59897x^{0.4} + 0.15968x^{0.95} + 0.03318x^{3} + 0.00847x^{23}$ was used, similar to (Cubuktepe and Topcu 2018).

The reader should also be made aware that the optimization algorithm itself is slow and relatively unstable to some parameter configurations and, therefore, theoretical work on techniques regarding CPT value optimization would prove useful and allow us to readily study agent-based systems with more than two agents and at more extreme parameter configurations.

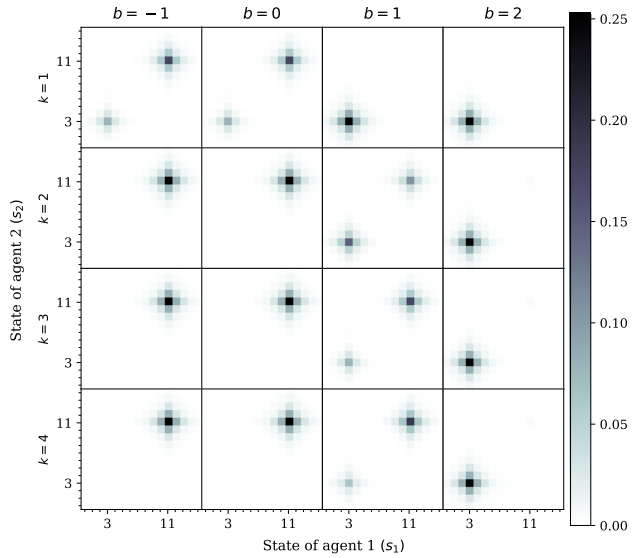



Figure 5: Stationary distributions of the resulting Markov chains obtained by conditioning the Markov game to increasingly sophisticated policies, $k = 1, 2, 3$ and $4$, for CPT-agents with several reference points $b = -1, 0, 1, 2$. We assumed discount factors $\beta_1 = \beta_2 = 0.9$, utility function $u(x) = x$ and weighting function $w(x) = e^{-0.5(-\log(x))^{0.9}}$.

## Conclusion

Cognitive biases and theory of mind are a fundamental part of being human. We have shown that, by including cognitive biases and theory of mind in the dynamics of a coordination game, agents are able to coordinate much more easily. Specifically, equipping agents with cumulative prospect theory helps coordination compared to the (standard) expected utility theory.

Furthermore, increasingly sophisticated policies in the context of bounded rationality (i.e. increasing value of level-$k$ theory of mind) help coordination between both EUT- and CPT-agents. Additionally, we also provided evidence that higher sophistication levels (i.e. higher than $k = 3$) do not seem to change the outcome of the two agent setting in the long run. Thus, this latter provides more evidence that unbounded rationality is not only practically unfeasible, but also unnecessary for coordination.

Also, we have shown how the consideration of long-term rewards over short-term ones affects the coordination of EUT- and CPT-agents. Specifically, for lower values of the discount factor $\beta$, agents will increasingly prefer short-term rewards over long-term rewards. Besides, we have shown that preferring short-term rewards inhibits the coordination of both EUT- and CPT-agents, while the opposite promotes coordination. Furthermore, we remarkably observed that more sophisticated policies in the theory of mind help agents coordinate, even if the long-term reward consideration makes it unlikely at first.

Lastly, we looked at the sensitivity of the coordination to the reference points of the agents. In particular, we have observed that higher reference points decrease coordination, which suggests that the framing of gains and losses plays an important role in the emergence of human coordination.

**Future Work** As we have shown, behavioral agent models provide significantly different system dynamics compared to prescriptive agent models and, therefore, several interesting research directions naturally arise. For instance, multi-agent systems where agents represent people should use a descriptive behavioral model instead of a prescriptive model. Upon realizing this, one can start to develop and study human-based models such as idealized forms of democracy (e.g., liquid democracy (O'Donell 1994)), video-game artificial intelligence with human-like behavior (or that is able to understand human-like behavior) and policy-making, or even revisiting already known conflict problems such as the tragedy of the commons and the diffusion of responsibility.

It would also prove interesting to create an inference model to obtain the optimal parameters of this model, similar to (Yoshida, Dolan, and Friston 2008). For instance, a Bayesian method to infer the reference point, discount factor, utility and weighting function parameters, and policy sophistication level would enable machines to learn to act in a more personalized manner.

One caveat of the bounded rationality using a level-$k$ model is the assumption that stereotype policies are uniform, which may be rather unrealistic. Therefore, a way of creating more realistic stereotyped policies would be an interesting problem to tackle. One such way is self-play, a reinforcement learning method to train agents by pitting them against themselves and, in an evolutionary manner, preserving winners and discarding losers (Silver and others 2017).

In the two-agent level-$k$ model, it is known that humans, in general, do not use more sophistication than level-3 (Stahl II and Wilson 1994). This creates a finite hypothesis space for the policy levels (i.e., with $k = 0, 1, 2$ and 3). However, when multiple interacting agents are a part of the environment, it is not enough to specify policy levels as a single number because each agent may have a policy which is a best response against several other policies of different levels. Therefore, there exists a problem of finding a behaviorally plausible hypothesis space for the inferred orders of each agent, which, if solved, would allow inference to be done on a collective level. Specifically, we would like reasoning such as "what you think about what he things that she thinks..." to be described in a simple, yet well-structured manner. The team theory of mind model proposed in (Shum et al. 2019) is an interesting setting that tackles some of the problems but its solution is computationally costly. At last but not least, experimental verification of the proposed framework could be done via a sociological study, which may also generate interesting data to further validate and expand the proposed model. These and other related research paths may lead to new knowledge of the dynamics of systems comprised of people and, in turn, unlock the knowledge we lack to build artificial entities capable of understanding and simulating human behavior.

## References


Abdellaoui, M.; Bleichrodt, H.; and Kammoun, H. 2013. Do financial professionals behave according to prospect theory? An experimental study. *Theory Decis.* 74(3):411–429.

Allais, M. 1979. The Foundations of a Positive Theory of Choice Involving Risk and a Criticism of the Postulates and Axioms of the American School (1952). *Expected utility hypotheses and the Allais paradox* 27–145.

Barrett, S., and Stone, P. 2015. Cooperating with unknown teammates in complex domains: A robot soccer case study of ad hoc teamwork. In *AAAI-15*. AAAI press.

Bonabeau, E.; Marco, D. d. R. D. F.; Dorigo, M.; Theraulaz, G.; et al. 1999. *Swarm intelligence: from natural to artificial systems*. Number 1. OUP.

Cooper, R. 1999. *Coordination games*. Cambridge UP.

Cubuktepe, M., and Topcu, U. 2018. Verification of Markov Decision Processes with Risk-Sensitive Measures. In *Proc. of the American Control Conf., ACC'18*, 2371–2377. IEEE.

Cxvi, V.; Barberis, N.; Huang, M.; and Santos, T. 2001. Prospect Theory and Asset Prices. *Q. J. Econ.* (1).

Ellsberg, D. 1961. Risk, ambiguity and the Savage Axioms. *Q. J. Econ.* 643–669.

Fiegenbaum, A. 1990. Prospect theory and the risk-return association. *Journal of Economic Behavior & Organization* 14(2):187–203.

Foerster, J.; Chen, R. Y.; Al-Shedivat, M.; Whiteson, S.; Abbeel, P.; and Mordatch, I. 2018. Learning with opponent-learning awareness. In *AAMAS-18*, 122–130. IFAAMAS.

Howard, R. A. 1960. *Dynamic Programming and Markov Processes*.

Kahneman, D., and Tversky, A. 1979. Prospect Theory: An Analysis of Decision Under Risk. *Econometrica* 47(2):263–291.

Kar, D.; Fang, F.; Fave, F. D.; Sintov, N.; and Tambe, M. 2015. "A Game of Thrones": When Human Behavior Models Compete in Repeated Stackelberg Security Games. In *AAMAS-15*, 1381–1390. IFAAMAS.

Lin, K., and Marcus, S. I. 2013. Dynamic Programming with Non-Convex Risk-Sensitive Measures. In *American Control Conference*, 6778–6783. Washington, DC, USA: IEEE.

Lin, K. 2013. *Stochastic Systems with Cumulative Prospect Theory*. Ph.D. Dissertation.

List, J. A. 2004. Neoclassical theory versus prospect theory: Evidence from the marketplace. *Econometrica* 72(2):615–625.

Mealing, R., and Shapiro, J. L. 2013. Opponent modelling by sequence prediction and lookahead in two-player games. In *Int Conf. Art. Intell. Soft Computing*, 385–396. Springer.

Metzger, L. P., and Rieger, M. O. 2019. Non-cooperative games with prospect theory players and dominated strategies. *Games Econ. Behav.* 115:396–409.

O'Donell, G. A. 1994. Delegative Democracy. *Journal of Democracy* 5(1):55–69.

Paiva, A.; Santos, F. P.; and Santos, F. C. 2018. Engineering pro-sociality with autonomous agents. In *AAAI-18*. AAAI press.

Pech, W. 2010. Behavioral Economics and the Basic Income Guarantee. *Basic Income Studies* (5):3–3.

Peterson, J.; Bourgin, D.; Reichman, D.; Griffiths, T.; and Russell, S. 2019. Cognitive model priors for predicting human decisions. In *ICML-19*, 5133–5141.

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

Rabinowitz, N. C.; Perbet, F.; Song, H. F.; Zhang, C.; Eslami, S.; and Botvinick, M. 2018. Machine theory of mind. *arXiv:1802.07740*.

Rahwan, I.; Cebrian, M.; Obradovich, N.; Bongard, J.; Bonnefon, J.-F.; Breazeal, C.; Crandall, J. W.; Christakis, N. A.; Couzin, I. D.; Jackson, M. O.; et al. 2019. Machine behaviour. *Nature* 568(7753):477.

Rand, D. G., and Nowak, M. A. 2013. Human cooperation. *Trends Cogn. Sci.* 17(8):413–425.

Santos, F. P.; Pacheco, J. M.; Paiva, A.; and Santos, F. C. 2019. Evolution of collective fairness in hybrid populations of humans and agents. In *AAAI'19*, volume 33, 6146–6153.

Shapley, L. S. 1953. Stochastic Games. In *Proc Natl Acad Sci USA*, 1095–1100.

Shum, M.; Kleiman-Weiner, M.; Littman, M. L.; and Tenenbaum, J. B. 2019. Theory of minds: Understanding behavior in groups through inverse planning. *arXiv:1901.06085*.

Silver, D., et al. 2017. Mastering the game of Go without human knowledge. *Nature* 550(7676):354–359.

Skyrms, B. 2004. *The stag hunt and the evolution of social structure*. Cambridge Univ. Press.

Stahl II, D. O., and Wilson, P. W. 1994. Experimental evidence on players' models of other players. 25:309–327.

Stahl, D. O. 1993. Evolution of smartn players. *Games and Economic Behavior* 5:604–617.

Stone, P.; Kaminka, G. A.; Kraus, S.; and Rosenschein, J. S. 2010. Ad hoc autonomous agent teams: Collaboration without pre-coordination. In *AAAI-10*. AAAI press.

Tversky, A., and Kahneman, D. 1992. Advances in Prospect Theory: Cumulative Representation of Uncertainty. 5:297–323.

Vis, B., and Van Kersbergen, K. 2007. Why and how do political actors pursue risky reforms? *J Theor Politics* 19(2):153–172.

von Neumann, J., and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. Princeton University Press.

Wooldridge, M. 2009. *An introduction to multiagent systems*. John Wiley & Sons.

Yoshida, W.; Dolan, R. J.; and Friston, K. J. 2008. Game theory of mind. *PLoS Comput. Biol.* 4(12).