Evolution favours positively biased reasoning in sequential interactions with high future gains

Marco Saponara[1,a,*], Elias Fernández Domingos[1,2,b],
Jorge M. Pacheco[3,4,c], Tom Lenaerts[1,2,5,d]

[1]Machine Learning Group, ULB, Campus la Plaine, Brussels 1050, Belgium
[2]Artificial Intelligence Lab, VUB, Pleinlaan 9, Brussels 1050, Belgium
[3]INESC-ID, IST-Tagusparque, 2744-016 Porto Salvo, Portugal
[4]ATP-group, P-2744-016 Porto Salvo, Portugal
[5]Center for Human-Compatible AI, UC Berkeley, 2121 Berkeley Way, Berkeley, CA 94720, USA
[a]ORCID iD: 0009-0008-8092-6220 [b]ORCID iD: 0000-0002-4717-7984
[c]ORCID iD: 0000-0002-2579-8499 [d]ORCID iD: 0000-0003-3645-1455
[*]Corresponding author. Email: marco.saponara@ulb.be

Publication details

This work has been published in the Royal Society Interface journal on date 27/08/2025. DOI: `https://doi.org/10.1098/rsif.2025.0153`

Artificial intelligence (AI) systems can perform more and more complex tasks with minimal human intervention. These systems, commonly referred to as *agentic AI* [1], are developed with a rationalism that allows them to reason and adapt to different scenarios. Human decision making, on the other hand, is influenced by several biases and consistently deviates from rational choices [2]. As training data can transmit these biases to an agent, it is important to understand which biases can lead to more successful social interactions.

To address this question, we investigate the co-evolution of strategic reasoning and cognitive biases using evolutionary game theory [3]. The goal is to understand whether reasoning strategies with particular properties may have been favoured by evolution, thus interfering with the development of rational behaviour [4]. Borrowing ideas from evolutionary biology, a strategy propagates by *imitation* if it is associated with a higher fitness, i.e., yields a higher gain than the others [5].

Our analysis is applied in the context of sequential interactions, represented here by the Centipede Game [6]. This social dilemma involves two individuals who take turns, over a total of L steps, to decide the division of a shared resource. Here we focus on the *Incremental Centipede Game* (ICG), where the resource starts at 0.5 and doubles at every step (see Figure 1) [7]. The payoff structure of the ICG leads a rational self-interested individual to stop the game as early as possible. However, this outcome is rarely observed in behavioural experiments [8].

In our evolutionary model, individuals in a population use a variety of unbiased and biased reasoning strategies to anticipate others' behaviour in the ICG. Reasoning is represented through level–k theory [9]. In this framework, the degree of rationality of an individual is represented by an integer $k \geq 0$, where $k = 0$ corresponds to a naive type who follows a predetermined strategy, while higher values mean more sophisticated reasoning. Different cognitive biases can then influence the reasoning process through a noise parameter ε, which perturbs the inferred action away from the optimal action at each reasoning step.

We find that a reasoning strategy with a systematic inference bias towards higher but uncertain rewards is favoured and becomes dominant under strong selection, whereas rational behaviour undergoes extinction, as Figure 2 shows. Individuals employing such a biased inference process think, at each step, that higher-payoff outcomes are more likely to occur compared to individuals using the analogous unbiased strategy. This *positively biased* reasoning strategy, which may be linked to the notion of *wishful thinking* [10], co-evolves with bounded rationality. In fact, the sophistication of each individual, i.e., the depth of their reasoning processes, remains limited, allowing for a better alignment with experimental data [7, 11].

This work addresses the emergence of the non-rational behaviour that is frequently observed in behavioural experiments. Our theoretical model, applied in the context of reciprocal exchanges of resources, corroborates the idea that certain cognitive biases, despite leading to systematic deviations from a rational judgment, can constitute an adaptive feature to successfully interact with our peers in this type of social dilemmas.

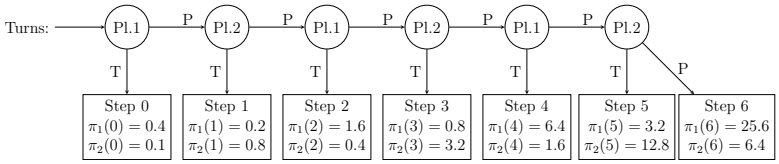

Figure 1: Extensive form of the six-step Incremental Centipede Game with exponential growth. Player 1 (Player 2) plays at even (odd) steps. Playing $T = Take$ at a given step means ending the game and receiving a larger share of the resource than the co-player. Playing $P = Pass$ means letting the other player decide what to do in the next step, unless the last decision node is reached, where Player 2 has to decide between two different splits. The notation $\pi_i(t)$ denotes the payoff Player i would get if the game ends at Step t.

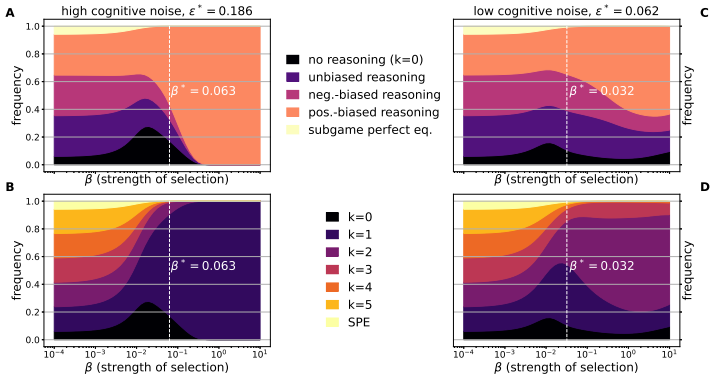

Figure 2: Positively biased reasoning co-evolves with bounded rationality. The distribution of reasoning types and reasoning levels in the six-step ICG is shown for variable strength of selection β, under high ($\varepsilon^* = 0.186$) and low ($\varepsilon^* = 0.062$) probability of reasoning errors (panels **A,B** and **C,D**, respectively). The selection strength represents the intensity of imitation: when $\beta = 0$, imitation occurs with 50% chance; when β is large, imitation becomes almost deterministic. For each value of β, we compute the stationary distribution ϕ of our dynamical system under the assumption of rare mutations [5]. The frequencies of each strategy are aggregated by reasoning type (unbiased, positively biased, and negatively biased) and by reasoning level ($1 \leq k \leq L$) (panels **A,C** and **B,D**, respectively). For completeness, the figures also include the frequency of the pure strategy associated with no reasoning (i.e., reasoning level $k = 0$) and the sub-game perfect equilibrium (SPE) strategy of stopping the game as early as possible. The white dotted lines correspond to the selection strength β^* of optimal fitting with the data in [11] and [7] (panels **A,B** and **C,D**, respectively). The population size is fixed to 100 individuals.

References

[1] Acharya DB, Kuppan K, Divya B. 2025 Agentic AI: Autonomous Intelligence for Complex Goals—A Comprehensive Survey. *IEEE Access* **13**, 18912–18936. (10.1109/ACCESS.2025.3532853)

[2] Tversky A, Kahneman D. 1974 Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty.. *Science* **185**, 1124–1131. (10.1126/science.185.4157.1124)

[3] Smith JM, Price GR. 1973 The Logic of Animal Conflict. *Nature* **246**, 15–18. (10.1038/246015a0)

[4] Lenaerts T, Saponara M, Pacheco JM, Santos FC. 2024 Evolution of a theory of mind. *iScience* **27**, 108862. (10.1016/j.isci.2024.108862)

[5] Fernández Domingos E, Santos FC, Lenaerts T. 2023 EGTtools: Evolutionary game dynamics in Python. *iScience* **26**, 106419. (10.1016/j.isci.2023.106419)

[6] Rosenthal RW. 1981 Games of perfect information, predatory pricing and the chain-store paradox. *Journal of Economic Theory* **25**, 92–100. (10.1016/0022-0531(81)90018-1)

[7] McKelvey RD, Palfrey TR. 1992 An Experimental Study of the Centipede Game. *Econometrica* **60**, 803. (10.2307/2951567)

[8] Krockow EM, Colman AM, Pulford BD. 2016 Cooperation in repeated interactions: A systematic review of Centipede game experiments, 1992–2016. *European Review of Social Psychology* **27**, 231–282. (10.1080/10463283.2016.1249640)

[9] Stahl DO. 1993 Evolution of Smartn Players. *Games and Economic Behavior* **5**, 604–617. (10.1006/game.1993.1033)

[10] Aue T, Nusbaum HC, Cacioppo JT. 2012 Neural correlates of wishful thinking. *Social Cognitive and Affective Neuroscience* **7**, 991–1000. (10.1093/scan/nsr081)

[11] Kawagoe T, Takizawa H. 2012 Level-k analysis of experimental centipede games. *Journal of Economic Behavior & Organization* **82**, 548–566. (10.1016/j.jebo.2012.03.010)