Incorporating human plausibility in single- and multi-agent AI systems

Samuel A. Barnett

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Department of
Computer Science
Advisers: Ryan P. Adams and Tom Griffiths

May 2024

Abstract

As AI systems play a progressively larger role in human affairs, it becomes more important that these systems are built with insights from human behavior. In particular, models that are developed on the principle of human plausibility will more likely yield results that are more accountable and more interpretable, in a way that greater ensures an alignment between the behavior of the system and what its stakeholders want from it. In this dissertation, I will present three projects that build on the principle of human plausibility for three distinct applications:

(i) Plausible representations: I present the Priority-Adjusted Reply for Successor Representations (PARSR) algorithm, a single-agent reinforcement learning algorithm that brings together the ideas of prioritisation-based replay and successor representation learning. Both of these ideas lead to a more biologically plausible algorithm that captures human-like capabilities of transferring and generalizing knowledge from previous tasks to novel, unseen ones.

(ii) Plausible inference: I present a pragmatic account of the weak evidence effect, a counterintuitive phenomenon of social cognition that occurs when humans must account for persuasive goals when incorporating evidence from other speakers. This leads to a recursive, Bayesian model that encapsulates how AI systems and their human stakeholders communicate with and understand one another in a way that accounts for the vested interests that each will have.

(iii) Plausible evaluation: I introduce a tractable and generalizable measure for cooperative behavior in multi-agent systems that is counterfactually contrastive, contextual, and customizable with respect to different environmental parameters. This measure can be of practical use in disambiguating between cases in which collective welfare is achieved through genuine cooperation, or by each agent acting solely in its own self-interest, both of which result in the same outcome.

Acknowledgements

This PhD has been an incredible journey, and it would not have been possible to achieve were it not for the selfless help of many mentors and peers. In the sense of Chapter 4, they have been the real "cooperators".

My deepest gratitude goes to Ryan Adams, who has been my primary advisor since I began at Princeton back in 2018. Ryan taught me how to be an independent researcher, helping me develop an instinct for what are the interesting questions to ask, as well as what rabbit holes and underdeveloped software suites should be avoided at all cost. I continue to be inspired by his approach to research, moving from applications to discipline-traversing theory and back to applications again, and I hope to follow the example he has set as I pursue future work in the field. Beyond his advice and support in my PhD work and my next career moves, I am incredibly grateful that Ryan has also been someone I could turn to for moral and pastoral support during the times in this process when my confidence ebbed to a low.

I am also grateful to Tom Griffiths, my secondary advisor. I was enraptured by computational cognitive science and its philosophical underpinnings when I first came to Princeton: it seemed to me the perfect extension to the mathematics and philosophy I had studied during my undergraduate degree, applied to understanding the nature of intelligence itself. Tom has been inordinately generous in teaching me about the field and allowing me to explore my questions and tastes therein. He has also been a very useful resource when considering my career, particularly regarding Antipodean matters.

In addition to my advisors, my gratitude also extends to the excellent teams past and present that they have assembled.

Ryan's Laboratory for Intelligent Probabilistic Systems, my home in 418B, has been a place of friendship, collaboration, as well as devilish deceit in countless games of Avalon. My thanks goes to, amongst others: Aditya Palaparthi, Alex Guerra, Anat Kleiman, Ari Seff, Cindy Zhang, Deniz Oktay, Diana Cai, Eder Medina, Geoffrey Roeder, Gregory Gundersen, Jad Rahme, Jenny Zhan, Jordan Ash, Joshua Aduol, Kathryn Wantlin, Mehran Mirramezani, Nick Richardson, Olga Solodova, Sulin Liu, Xingyuan Sun, and Yaniv Ovadia. I am especially grateful to Alex Beatson, who has gone from friend, to flatmate, to mentor during these past six years.

In Tom's CoCoSci lab, my thanks goes to Evan Russek, Mark Ho, Matt Hardy, Mayank Agrawal, Rachit Dubey, Ruairidh Battleday, and Ted Sumers. I am especially grateful to Robert Hawkins, who in addition to being a patient collaborator and mentor, was also a good friend during difficult times.

I have been lucky to pursue my PhD with not only two labs, but also a host of outside collaborators. This

includes those with whom I worked at Microsoft Research in New York (Arthur Juliani, Brandon Davis, and Ida Momennejad) as well as Siemens Technology in Princeton (Siddarth Bhela, Suat Gumussoy, and Yubo Wang).

My PhD work has been made possible these past few years thanks to a generous grant from the John Templeton Foundation (#62220), alongside help from both Ryan and Tom. I am grateful to the Agency, Directionality, and Function team led by Alan Love, and to the Modeling Agency Formally cluster specifically, coordinated by Armin Schulz. Discussing my work with practitioners from very different backgrounds has been refreshingly fruitful, and I have learned much and had a lot of fun in the process.

I would not be in Princeton at all were it not for my initial year supported by the Daniel M. Sachs Class of 1960 Scholarship at Princeton, which in 2017 first opened the doors for me to cross the pond. I remember vividly the moment I got the call giving me the offer on the train from Oxford to London, knowing I was about to embark on an adventure the contours of which I could not begin to imagine. I soon met many dear friends at the Graduate School and the 2D food co-op who made Princeton feel much more like home, including Casey Eilbert, Laura Channing, Adeline Heck, and Jonathan Balkind.

I turn at last to my family, present and future.

Thank you to Rosalie, my partner (and soon wife!): you have been there for me throughout the PhD, supporting me when completing it seemed insurmountable, and celebrating my accomplishments along the way that I would otherwise have found all too easy to take for granted. You inspire me every day to be the best I can be at what I do (something you handily achieve yourself), and I hope this product of six years' hard work is a testament to that.

Thank you to my mother Mandy and sister Hannah: you have supported me both from afar and during my frequent return visits to Essex. You were eager interlocutors whenever I talked about the directions I was taking my PhD in, even making sure I steadfastly held to my accent rather than succumbing to any trans-Atlantic tendencies.

Finally, I am grateful to my father, Rodger, to whom this dissertation is dedicated. Dad passed away in April 2020 shortly after being diagnosed with pancreatic cancer, in what was already a tumultuous time in the world at large. Despite this, he has been with me always, inspiring me not only as the original Barnett with a PhD, but also as a person as a whole. I hope I have done him proud as I follow in his footsteps.

To my father, Rodger.

Contents

Abstract . . . iii

Acknowledgements . . . iv

List of Tables . . . x

List of Figures . . . xi

List of Algorithms . . . xii

1 Introduction **1**

1.1 Human plausibility in machine learning . . . 1

1.1.1 What is human plausibility? . . . 1

1.1.2 Why design things according to human plausibility? . . . 2

1.2 Background . . . 4

1.2.1 Single-agent Reinforcement Learning . . . 4

1.2.2 Multi-agent Reinforcement Learning . . . 8

1.3 Research contributions in this dissertation . . . 10

1.3.1 Research contributions not included in this dissertation . . . 11

2 PARSR: Priority-Adjusted Replay for Successor Representations **13**

2.1 Introduction . . . 13

2.2 Background . . . 15

2.2.1 The Successor Representation . . . 15

2.2.2 Prioritized Replay . . . 16

2.3 Algorithms . . . 17

2.4 Experiments . . . 20

2.4.1 Six States . . . 20

2.4.2 Six Rooms . . . 21
2.5 Related work . . . 22
2.6 Discussion and Future Work . . . 23

3 A Pragmatic Account of the Weak Evidence Effect 25
3.1 Introduction . . . 25
3.2 Formalizing a pragmatic account of the weak evidence effect . . . 27
3.2.1 Reasoning about evidence from informative speakers . . . 27
3.2.2 Reasoning about evidence from motivated speakers . . . 28
3.3 Experiment: The Stick Contest . . . 29
3.3.1 Participants . . . 31
3.3.2 Design and procedure . . . 31
3.4 Results . . . 32
3.4.1 Behavioral results . . . 32
3.4.2 Model simulations . . . 33
3.4.3 Quantitative Model Comparison . . . 35
3.5 Discussion . . . 37

4 Measuring Cooperation with Counterfactual Planning 40
4.1 Introduction . . . 40
4.2 The challenges of defining cooperation . . . 42
4.3 Desiderata for a measure of cooperation . . . 43
4.4 Measuring cooperation in stochastic games . . . 45
4.5 Experiments in Social Dilemmas . . . 47
4.5.1 Matrix Game Social Dilemmas . . . 47
4.5.2 Iterated Social Dilemmas . . . 49
4.5.3 Tabular Cleanup . . . 50
4.6 Conclusion . . . 52

5 Conclusion 57

Bibliography 59

A Appendix to Chapter 3 **75**

A.1 Exclusions and attention checks 75

A.2 Order effects 78

A.3 Proofs 78

A.4 Results from original sample 80

A.5 Model fitting details 81

A.5.1 RSA model 81

A.5.2 Belief-adjustment models 82

A.5.3 Higher levels of reasoning and the strong evidence effect 83

A.6 Transcript of the Experiment 84

B Appendix to Chapter 4 **88**

B.1 Full results across multiple welfare functions 88

List of Tables

3.1 Weak evidence effect model comparison results 36

A.1 Stricter attention check passage rates broken out by speaker group. 76

A.2 Participant reasoning in the speaker phase of the Stick Contest 87

A.3 Participant reasoning in the judge phase of the Stick Contest 87

List of Figures

2.1 PARSR flowcharts 19
2.2 Six States experiment structure and results 20
2.3 Six Rooms experiment map and results 21

3.1 Stick Contest paradigm diagram 30
3.2 Pragmatic expectations in the weak evidence effect 33
3.3 Weak evidence effect model simulations 34

4.1 Cooperation measure schematic 46
4.2 Matrix game social dilemmas 48
4.3 Cooperativeness in the Iterated Prisoner's Dilemma 53
4.4 Cooperativeness against TFT in the Iterated Prisoner's Dilemma over time 54
4.5 Cooperativeness in 2-player Tabular Cleanup 55
4.6 Cooperativeness in 3-player Tabular Cleanup 56

A.1 Belief revision on second piece of evidence in the Stick Contest 77
A.2 Order effects in the Stick Contest 78
A.3 The raw data distribution of responses for the listener phase of the Stick Contest 85
A.4 Visualization of posterior predictive distribution for the speaker-dependent RSA model and heterogeneous MAS model 85
A.5 Full Bayesian posteriors for the parameters of the speaker-dependent RSA model 86

B.1 Cooperativeness in the Iterated Prisoner's Dilemma for all welfare functions 89
B.2 Cooperativeness in 2-player Tabular Cleanup for all welfare functions 90
B.3 Cooperativeness in 3-player Tabular Cleanup for all welfare functions 90

List of Algorithms

1 Value Iteration 6
2 Temporal Difference (TD) Learning 7
3 Q-Learning 7
4 Dyna 8
5 PARSR 18

Chapter 1

Introduction

AI and machine learning models are an increasingly prevalent mediator in our lives, with applications ranging from autonomous vehicles, chatbots, and recommendation systems shaping much of what we choose to consume online. Whether the integration of AI remains tacit or explicitly advertised, there is a tendency for users of AI-based tools to anthropomorphize their outputs [129]. Users' interactions with these tools are shaped by the expectation that these tools act and reason in a human-like manner and with a certain base level of competency, which can lead to adverse consequences if not managed responsibly [165]. It is always incumbent on those who promote such tools to be open about their design and subsequent limitations in order to mitigate any negative consequences. However, to the extent possible, these tools should also be designed according to principles that can best close the gap between user expectations and reality.

This dissertation proposes *human plausibility* as one such design principle. In this Introduction we define this principle and justify its use. We then give the necessary technical background, focusing on concepts and algorithms in (tabular) single- and multi-agent reinforcement learning. Finally, we outline the research contributions of this dissertation, as they lie within three areas of the human-plausible design framework.

1.1 Human plausibility in machine learning

1.1.1 What is human plausibility?

In this dissertation, we define a *human-plausible system* as one whose design has been informed by our models of human behavior, neuroscience, reasoning, and concepts. Human-plausible design may therefore draw from

many disciplines, including neuroscience, cognitive science, philosophy, economics, and evolutionary biology. Moreover, it can be taken into consideration at all three of Marr's levels for analyzing information-processing systems [95]:

- At the level of *computation* or *function*, it can make reference to the goals that humans have and the concepts humans have developed to understand these goals, as well as the environments that have shaped these goals in the first place. For example, in human-AI interaction, the behavior of an AI can be specified according to a reward that requires knowledge of the human's own goals [10].
- At the level of *representations* and *algorithms*, it can make reference to the strategies humans use to solve their problems and how the choice of representation of inputs, outputs, and the objects to be learned can aid those strategies. For example, different learned representations for decision-making tasks can lead to different degrees of transfer between tasks that can match human-like behavior [108].
- At the level of *implementation*, it can make reference to the physical substrates of the computations that humans make. For example, neurally plausible AI seeks to develop systems whose computations mimic those of different parts of the brain at the cellular or functional levels [105].

Not all AI systems are or ought to be designed according to human-plausibility. This is true even for those systems relating to human action or designed to potentially supersede a human-performed task. For example, an agent that is trained to run a power network must be robust to many kinds of change in the environment [94, 93, 92]. Many of the sources of this change, such as damage from weather events like storms, or the variability of renewable energy sources like wind, do not have an immediate human origin. Even in the case of human-caused changes like cyberattacks, we might prefer to ensure safety and stability through a strategy that can provide robust guarantees against a broad range of attacks whose form may be unpredictable, rather than one that is a tailored best-response to a distribution of attacks that have been seen in the past.

1.1.2 Why design things according to human plausibility?

Though it is not always appropriate, there are nevertheless many use-cases in which we would want to turn towards human-plausible design. Next, we give three reasons why we would want to design our AI systems according to this principle.

Inductive biases Humans are capable of solving a diverse array of tasks with limited computational resources, including tasks that have not previously been encountered [82]. This suggests that the paradigm of human reasoning extends beyond purely making statistical inferences from past experience alone, and that in fact humans have (inductive) biases shaped by all three levels of analysis that endow them with the ability to quickly solve these tasks.

While much energy (literal and figurative) is currently being expended to increase the capabilities of AI systems through the use of greater compute and the collection of more data [144], we might still worry about the potential limitations of an approach that relies solely on statistical relationships between past data [117]. If we wish to move beyond these limitations, we should instead focus on developing our understanding on what it is that enables humans to successfully perform generalization across tasks, as well as engage in abductive reasoning wherein we generate useful new hypotheses to understand the world around us [90].

Interpretability Beyond increasing the capabilities of our AI systems, human-plausible design also allows us to develop these systems along a path in which each design step is more likely understandable and interpretable by an end-user of the system. In particular, human-plausble design works with our innate tendency to anthropomorphize our systems by designing them in a fashion that actively seeks to mimic our own behavior and modes of reasoning. Unlike approaches which seek to use the tools of computational cognitive science in order to understand complex systems *not* designed to be human-plausible [30], human-plausible design is more likely to be successful by making our design choices explicit in the first place, thus further closing the gap between true human behavior and human-like behavior.

Compatibility In order for AI systems to work well, they must also do so *with humans*. In the context of applications such as self-driving cars, the nuances of human behavior often function as the difficult edge case hindering the ultimate success of the research program [73]. In many other contexts, we would like an AI system to act towards a goal in accordance with certain values that are hard to specify even in natural language, let alone computationally [4]. Building systems that are *human-compatible* [128] therefore demands a greater understanding of human behavior and values, and is therefore best achieved through human-plausible design.

1.2 Background

Both Chapters 2 and 4 involve models of sequential decision-making under uncertainty. The paradigmatic approach to formalizing such models and investigating them is *reinforcement learning*, which we review in this section. We first look at decision-making in the single-agent setting before moving to multi-agent systems. In both settings, we assume a *tabular* model in which we observe and can represent the environmental state exactly, rather than having partial observability. For the latter case, in which feature approximation becomes more pertinent, refer to [146] for the single-agent setting and [42] for a treatment of the multi-agent setting. Section 1.2.1 largely follows the material in [104, Ch. 14] and [146], with Section 1.2.2 drawing from [138].

1.2.1 Single-agent Reinforcement Learning

The environment in which an agent is situated, including its interactions and rewards, is modeled as a *Markov decision process* (MDP), a tuple $(\mathcal{S}, \mathcal{A}, P, R, \gamma, \rho)$ consisting of:

- a set of *states* $\mathcal{S}$;
- a set of possible *actions* $\mathcal{A}$;
- a *transition function* $P\colon \mathcal{S} \times \mathcal{A} \to \Delta\mathcal{S}$ mapping state-action pairs (s, a) to a distribution[1] over states s' (where we also write $P(s, a, s') \in [0, 1]$ to refer to an individual probability);
- a *reward function* $R\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ mapping state-action pairs (s, a) to an *expected, scalar* reward;
- a discount rate $\gamma \in [0, 1]$; and
- a distribution over initial states $\rho \in \Delta\mathcal{S}$.

The *Markov* in MDP refers to the fact that the states are *Markovian*, in that the probability of transitioning to state s' is fully determined by the state s and action a.

Agents act in the environment according to a *policy* $\pi\colon \mathcal{S} \to \Delta\mathcal{A}$ mapping states s to a distribution over actions a. A policy is deterministic if for all states $s \in \mathcal{S}$ there exists an action $a \in \mathcal{A}$ such that $\pi(s) = \delta_a$, where δ is the Dirac delta function representing a point probability mass on action a. By abuse of notation, we can also write this as $\pi(s, a) = 1$ or $\pi(s) = a$.

[1] For a set $\mathcal{X}$, we use the notation $\Delta\mathcal{X}$ to refer to the set of probability distributions over $\mathcal{X}$.

By combining policies and transition functions, we also derive an expected transition function

$$T_\pi(s, s') = \sum_a P(s, a, s')\pi(s, a). \tag{1.1}$$

As agents interact with the environment according to some policy π, they accumulate reward. The key term for evaluating a policy is the *value function* $V_\pi \colon \mathcal{S} \to \mathbb{R}$, defined as the expected discounted sum of rewards by an agent starting in state s and acting according to a policy π:

$$V_\pi(s) = \mathbb{E}_{a_t \sim \pi(s_t), s_{t+1} \sim P(s_t, a_t)} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \mid s_0 = s \right]. \tag{1.2}$$

These values obey a system of linear equations known as the *Bellman equations*:

$$\forall s \in \mathcal{S}, \quad V_\pi(s) = \mathbb{E}_\pi \left[R(s, \pi(s)) \right] + \gamma \sum_{s'} T(s, s') V_\pi(s'). \tag{1.3}$$

We also express these equations in matrix form as

$$\boldsymbol{V}_\pi = \boldsymbol{R}_\pi + \gamma \boldsymbol{T}_\pi \boldsymbol{V}_\pi, \tag{1.4}$$

where $\boldsymbol{V}_\pi \in \mathbb{R}^{|\mathcal{S}|}$ is the vector of value functions, $\boldsymbol{R}_\pi \in \mathbb{R}^{|\mathcal{S}|}$ is the vector of expected rewards such that $\boldsymbol{R}_\pi[s] = \sum_a R(s, a)\pi(s, a)$, and $\boldsymbol{T}_\pi \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$ is the expected transition function T expressed in matrix form.

For an MDP with finitely many states, Bellman's equation admits a unique solution given by

$$\boldsymbol{V}_\pi = (\boldsymbol{I} - \gamma \boldsymbol{T}_\pi)^{-1} \boldsymbol{R}_\pi, \tag{1.5}$$

where $\boldsymbol{I} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$ is the identity matrix. For a fully known MDP and a sufficiently small number of states, it is computationally feasible to use Eq. 1.5 to find the value of any policy π.

The objective of the agent is to find an policy π^* that is *optimal* in the sense that it maximizes value for all states, that is, $V_{\pi^*}(s) = \max_\pi V_\pi(s)$ for all $s \in \mathcal{S}$. For ease of notation, we write V^* instead of V_{π^*}.

This gives rise to the optimal state-action value function $Q^* \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, defined to be the expected return for taking action $a \in \mathcal{A}$ at state $s \in \mathcal{S}$ and then following the optimal policy:

$$Q^*(s, a) = R(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s, a, s') V^*(s'). \tag{1.6}$$

Algorithm 1 Value Iteration

1: **Input**: Arbitrary initial value function $\boldsymbol{V} \in \mathbb{R}^{|\mathcal{S}|}$, tolerance $\varepsilon > 0$.
2: **while** True **do**
3: $\quad \forall s \in \mathcal{S}, \quad \boldsymbol{V}'[s] \leftarrow \max_{a \in \mathcal{A}} \left\{ R(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s,a,s') \boldsymbol{V}(s') \right\}$.
4: $\quad$ **if** $\|\boldsymbol{V} - \boldsymbol{V}'\| < \frac{(1-\gamma)\varepsilon}{\gamma}$ **then**
5: $\quad\quad$ **break**
6: **return** $\boldsymbol{V}'$

From this definition, we can see that

$$\forall s \in \mathcal{S}, \quad V^*(s) = \max_{a \in \mathcal{A}} Q^*(s,a). \tag{1.7}$$

Moreover, we have that

$$\forall s \in \mathcal{S}, \quad \pi^*(s) = \arg\max_{a \in \mathcal{A}} Q^*(s,a). \tag{1.8}$$

Eq. 1.8 has two implications. Firstly, for all MDPs there exists an optimal policy that is deterministic. Secondly, that knowledge of the optimal state-action value function Q^* is sufficient for the agent to determine the optimal policy, without any direct knowledge of the reward or transition function.

If we substitute in Q^* in Eq.1.7 for its definition in Eq. 1.6, we get the *Bellman optimality equation*

$$V^*(s) = \max_{a \in \mathcal{A}} \left\{ R(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s,a,s') V^*(s') \right\}. \tag{1.9}$$

Unlike Eq. 1.3, this set of equations is no longer linear due to the max operator. However, we can still find a solution iteratively using the *value iteration algorithm* (Alg. 1). For any initial value function $\boldsymbol{V}$, this algorithm is guaranteed to converge to $\boldsymbol{V}^*$ to within ε accuracy in $\mathcal{O}\left(\log \frac{1}{\varepsilon}\right)$ iterations [104, Thm. 14.2], with each iteration requiring $\mathcal{O}\left(|\mathcal{S}|^2|\mathcal{A}|\right)$ operations. Observe that we can also derive Q^* and π^* from this algorithm, by considering the term inside of the max operator in line 3.

Value iteration is an example of a *planning* algorithm: the MDP is fully known and represented in advance. However, this is often an unrealistic assumption when modeling actual behavior. In particular, we are also interested in considering cases in which the MDP is not known in advance, and an agent learns about the reward and transition structure through interacting with the environment, while possibly also simultaneously learning a corresponding policy that asymptotically converges to an optimal one.

Consider first the case of stochastically learning V_π for a fixed π interacting with the environment. At a state s, the policy takes an action a, yielding a reward r and leading to a subsequent state s'. Combining

Eqs. 1.2 and 1.3, we see that $r + \gamma V_\pi(s')$ is an unbiased estimator of $V_\pi(s)$.

Of course, we do not know V_π. However, assuming we have a current estimate V, we can update this estimate via

$$V(s) \leftarrow (1-\alpha)V(s) + \alpha[r + \gamma V(s')] \tag{1.10}$$

$$= V(s) + \alpha[r + \gamma V(s') - V(s)], \tag{1.11}$$

where $\alpha > 0$ is a learning rate and the term inside the square brackets in Eq. 1.11 gives the *temporal difference* (TD) error between the current and subsequent estimates of $V(s)$. This update forms the basis of the TD-learning algorithm for learning V_π (Alg.2).

Algorithm 2 Temporal Difference (TD) Learning

Hyperparameters: Learning rate $\alpha > 0$, arbitrary initial value function V.
Draw initial state $s \sim \rho$.
while True **do**
    Draw $a \sim \pi(s)$.
    Take action a; observe reward, r, and state, s'.
    $V(s) \leftarrow V(s) + \alpha[r + \gamma V(s') - V(s)]$.
    $s \leftarrow s'$.

By adapting the TD update slightly to reflect the Bellman optimality equation (Eq. 1.9) rather than the Bellman equation (Eq. 1.3) we arrive at the Q-learning algorithm (Alg 3), which can be used to learn the optimal state-action value function Q^* [164]. Rather than acting according to a fixed policy π, we now derive our actions from our current state-action value function Q, while also exploring other actions. A common way to do this is via an ε-greedy policy, which chooses the best action $a^* = \arg\max_a Q(s,a)$ with probability $(1-\varepsilon)$, and otherwise chooses an action a at random. This approach guarantees that the full space will be explored.

Algorithm 3 Q-Learning

Hyperparameters: Learning rate $\alpha > 0$, exploration parameter $\varepsilon > 0$, arbitrary initial state-action value function Q.
Draw initial state $s \sim \rho$.
while True **do**
    $a \leftarrow \varepsilon$-*greedy*(s, Q).
    Take action a; observe reward, r, and state, s'.
    $\delta_Q \leftarrow r + \gamma \max_{a'} Q(s', a') - Q(s, a)$.
    $Q(s,a) \leftarrow Q(s,a) + \alpha\delta_Q$.
    $s \leftarrow s'$.

Given a finite-state MDP and a specific learning rate schedule, the Q-learning algorithm converges almost surely to Q^* [104, Thm. 14.9]. A similar result guarantees the convergence for TD learning. However, this convergence may still be slow, due either to the costliness of interacting with the environment, or the general sparsity of reward signals in the environment. To speed this up, the Dyna algorithm (Alg. 4) interleaves updates to Q from direct environmental interactions with updates from *replay* via a progressively learned model of the environment [145, 146].[2] This model, written $\mathit{Model}(s, a)$, can record an individual next state s' or a learned distribution over such states, along with the reward r.

Algorithm 4 Dyna

Hyperparameters: Learning rate $\alpha > 0$, exploration parameter $\varepsilon > 0$, arbitrary initial state-action value function Q, number of replay cycles n.
Draw initial state $s \sim \rho$.
while True **do**
    $a \leftarrow \varepsilon\text{-}\mathit{greedy}(s, Q)$.
    Take action a; observe reward, r, and state, s'.
    $\mathit{Model}(s, a) \leftarrow r, s'$
    $\delta_Q \leftarrow r + \gamma \max_{a'} Q(s', a') - Q(s, a)$.
    $Q(s, a) \leftarrow Q(s, a) + \alpha\delta_Q$.
    $s \leftarrow s'$.
    loop n times
        $\overline{s} \leftarrow$ random previously observed state.
        $\overline{a} \leftarrow$ random action previously taken in $\overline{s}$.
        $\overline{r}, \overline{s}' \leftarrow \mathit{Model}(\overline{s}, \overline{a})$.
        $Q(\overline{s}, \overline{a}) \leftarrow Q(\overline{s}, \overline{a}) + \alpha[\overline{r} + \gamma \max_{\overline{a}'} Q(\overline{s}', \overline{a}') - Q(\overline{s}, \overline{a})]$.

1.2.2 Multi-agent Reinforcement Learning

We can represent multi-agent systems as a *stochastic game* [137, 138], a tuple

$$\left(\mathcal{S}, N, \prod_{i=1}^{N} \mathcal{A}_i, P, \prod_{i=1}^{N} R_i, \gamma, \rho\right)$$

consisting of:

- a set $\mathcal{S}$ of states;
- N agents, indexed by $i = 1, ..., N$;
- a set of available actions $\mathcal{A}_i$ for each agent i;
- a transition function $P\colon \mathcal{S} \times \prod_{i=1}^{N} \mathcal{A}_i \to \Delta\mathcal{S}$;

[2] We alternatively refer to this algorithm as the *Dyna-Q* algorithm in Chapter 2 to emphasize the function being learned.

- a reward function $R_i \colon \mathcal{S} \times \prod_{i=1}^{N} \mathcal{A}_i \to \mathbb{R}$ for each agent i;
- a discount rate γ; and
- a distribution over initial states $\rho \in \Delta\mathcal{S}$ as before.

Stochastic games are a useful broad formalism encompassing many other modelling scenarios as special cases:

- An MDP is a stochastic game where $N = 1$.
- A normal form matrix game, the simplest form of a game studied in game theory, can be viewed as a stochastic game with a single state that is played for one time step.[3]
- An iterated game, in which a normal form matrix game is played *ad infinitum*, is a stochastic game with with a single state s. We can incorporate memory into the iterated game by having states which consist of possible actions taken at previous time steps. However, for a memory of size m, this requires an exponentially scaling state space of size $|\mathcal{S}| = \mathcal{O}\left(|\mathcal{A}|^{mN}\right)$.

Since there are multiple reward functions, there is no single generalization of the notion of optimality to the multi-agent setting. However, by deriving MDPs from a stochastic game, we can in turn derive solution concepts which are useful in different contexts.

Firstly, imagine that all agents are controlled by a single controller who chooses the actions $\boldsymbol{a}$ for all agents. If all reward functions are equal, so that $R_i = R$ for all $i = 1, ..., N$, we get the *cooperative* setting for multi-agent systems. Otherwise, we can still derive a single reward function via the sum of all rewards $R = \sum_i R_i$, which we refer to as the prosocial setting. The prosocial optimum is the optimal policy of the MDP consisting of this reward function and an action space where each action is a tuple of all individual agents' actions.

Not all stochastic games involve agents with aligned or altruistic motives, or that are capable of coordinating with one another. Let $\boldsymbol{\pi}_{-i}$ be the policies of all agents except for agent i. Holding this fixed, we derive an MDP for agent i with a reward given by R_i. Let $\mathrm{BR}_i(\boldsymbol{\pi}_{-i})$ be the set of optimal policies to this MDP, also referred to as the *best responses* to $\boldsymbol{\pi}_{-i}$. A set of policies $\pi_1, ..., \pi_N$ are a *Nash equilibrium* for a stochastic game if for all $i = 1, ..., N$,

$$\pi_i \in \mathrm{BR}_i(\boldsymbol{\pi}_{-i}). \tag{1.12}$$

[3] We can fit this more formally in to the stochastic game framework via a game with an initial dummy state s_0 and a terminal state s_T, such that for $s \in \{s_0, s_T\}$, for all action combinations $\boldsymbol{a} \in \prod_i \mathcal{A}_i$, and for $i = 1, ..., N$, we have $P(s, \boldsymbol{a}, s_T) = 1$ and $R_i(s_T, \boldsymbol{a}) = 0$.

That is, in a Nash equilibrium, no agent has an incentive to defer from their current policy as it is the optimal response to all of the other policies. In a stochastic game, at least one Nash equilibrium is guaranteed to exist [138, Thm. 6.2.6].

1.3 Research contributions in this dissertation

The remainder of this dissertation (prior to the Conclusion in **Chapter 5**) is organized according to three application domains for the principle of human plausibility, tracking three projects that were completed in those domains.

Plausible representations As a core component of Marr's second level of analysis, *representations* are an important part of understanding human behavior, including how humans learn and how they adapt to unseen changes in their environment. Developing representations that balance compactness with flexibility is a challenge in both the space of evolutionary design and engineering design, and within the field of single-agent reinforcement learning there has been much study in finding representations that are human-plausible, both to bootstrap more competent behavior as well as to understand better that of humans [106].

In **Chapter 2** we present *priority-adjusted replay for successor representations* [13, 11], an algorithm combining prioritized replay with the successor representations, a form of representing the long-term transition structure of an environment in a way that more easily permits transfer between tasks in a human-plausible fashion. This work was presented at RLDM 2022 and CogSci 2022 and was carried out in collaboration with Ida Momennejad as part of an internship at Microsoft Research.

Plausible inference Humans learn about their environment from a diversity of sources: as passive observers, through active interaction with the world, or socially through interactions with other humans. Knowing that the information can come from different sources, we can account for it in different ways, leading to different inferences that are derived from the same information [65, 151, 150, 136, 63, 112]. If we wish to design AI systems that can usefully inform us and help us to achieve our goals, we need those systems to have an accurate model of how its outputs will be processed by us—in other words, we need to accurately model our *data-generating* assumptions and our subsequent inference procedures.

In **Chapter 3** we present *a pragmatic account of the weak evidence effect* [12]. This chapter develops a model of listener and speaker behavior where speakers are active agents, presenting evidence according to their own persuasive goals rather than at random. This generalizes the Rational Speech Acts (RSA)

framework [49], which provides a recursive Bayesian model for pragmatic inference. This work was published in Open Mind and was carried out in collaboration with Tom Griffiths and Robert Hawkins.

Plausible evaluation In a scenario in which humans with mixed motives deputize AI agents to act in a shared environment, these agents will have to coordinate and cooperate with one another in order to achieve their goals to the extent possible. Cooperative AI is an emerging discipline that develops algorithms and test frameworks for designing systems that can learn to cooperate in complex environments [32, 31]. Yet, in its reliance on evaluation metrics for cooperation that are arbitrary, acontextual, or fail to consider relevant counterfactuals regarding the environment's external reward structure, these recent efforts have failed to factor in insights about what constitutes cooperative behavior from cognitive science and evolutionary biology [154, 166].

In **Chapter 4** we present *a counterfactual measure of cooperative behavior*. This is a tractable and generalizable measure within the stochastic game framework that better accords with a *human-plausible* account of what constitutes cooperative (as well as anti-cooperative) behavior. After proposing the measure, we evaluate it on a number of games of increasing complexity that have been studied in the multi-disciplinary cooperation literature. This work is in preparation for publication, and has been carried out in collaboration with Ryan P. Adams. A poster based on this work was presented at the Cooperative AI Summer School in 2023.

1.3.1 Research contributions not included in this dissertation

In addition to the work covered in this dissertation, two other projects were completed during the course of the PhD:

- *Neuro-Nav: A Library for Neurally-Plausible Reinforcement Learning* [71]: in this work, completed concurrently with Barnett and Momennejad [13], we propose *Neuro-Nav*, an open-source library for neurally plausible reinforcement learning, available at `https://github.com/awjuliani/neuro-nav`. This work was presented at RLDM 2022 and was carried out in collaboration with Arthur Juliani (the primary author), Brandon Davis, Margaret Sereno, and Ida Momennejad.

- *Toeplitz posterior approximation for multivariate time series modeling with Gaussian processes*: in this work, we propose the Independent Toeplitz Variational Strategy (ITVS) for scalable multi-output Gaussian process regression on multivariate time series with latent correlation structure. In addition

to proposing the strategy, we also develop the *Carathéodory* parametrization for initialization and optimization on the space of real, symmetric Toeplitz positive definite matrices. This work is in preparation for publication, and has been carried out in collaboration with Joshua Aduol, Alex Guerra, Suat Gumussoy, and Ryan P. Adams.

Chapter 2

PARSR: Priority-Adjusted Replay for Successor Representations

> Those who cannot remember the past are condemned to repeat it.
>
> *George Santayana*

2.1 Introduction

Intelligent agents are capable of transfer and generalization. Imagine driving to a coffee shop to meet a friend. If we encounter a blocked road, we are able to quickly adapt in choosing a new route that will get us there. Or if we decide that a different coffee shop might be preferable, we can just as easily change course.

To accommodate this flexibility to changes in the environment, contemporary reinforcement learning algorithms often rely on representation learning and replay as human-plausible solutions [108, 106, 169, 14]. Learning flexible yet compact representations of the environment allows us to adapt to changes in its rewards, and replay enables us to adapt to changes in transition structures without the need for significant amounts of further real experience.

One such algorithm combines the successor representation (SR) [36, 143], in which states are represented in terms of long-run, time-discounted visitation frequencies according to a given policy, and memory replay (inspired by Dyna [145, Alg. 4]). This algorithm, Dyna-SR for short, captures human-like task transfer

behavior across a number of tabular tasks [127, Algorithm 3]. A key advantage of the Dyna-SR algorithm is that it is almost as flexible as a model-based RL algorithm (in which the full transition structure is learned and then used for planning), while at decision time it is almost as inexpensive as model-free RL (in which a policy is found solely by learning the long-term reward structure). By caching multi-step trajectories of states offline, Dyna-SR remains more flexible than model-free RL and SR alone, while avoiding model-based RL's high cost of rolling out entire state-action-state-reward trajectories at decision time.

As a caveat, the replay prioritization in current implementations of Dyna-SR focus on more recent memories for efficiency. While this heuristic may capture certain aspects of human memory recall [72], previous work in cognitive neuroscience suggest human-like replay may be better modelled by an error-based replay prioritization [107]. Thus, here we extend the scope of algorithms combining representation and replay, using both reward-based and representation-based errors for replay prioritization. While current approaches remain largely limited to tagging memories with reward prediction errors (PE) for priority [109, 119, 132, 60, 70], our proposed algorithm is inspired by Dyna-SR but can prioritize replay using either reward PE and successor PE.

We propose PARSR (pronounced *PARS*-er), Priority-Adjusted Replay for Successor Representations, which improves on Dyna-SR offering more human-plausible replay prioritization with no effective increase in hyperparameters. As an SR-based algorithm, PARSR learns a representation of the transition structure (i.e., the environment dynamics) and the reward structure separately, allowing either to be quickly relearned for greater generalization. Critically, by decoupling reward and transition representations, PARSR can use the prediction errors from either to prioritize memory replay.

We propose two variants of PARSR based on the choice of prediction error: M-PARSR (prioritizes memories using successor PE) and Q-PARSR (prioritizes memories using value PE). This prioritization for memory selection distinguishes PARSR from Dyna-SR [127], which performs replay-enhanced SR learning with random memory selection with a recency bias.

We test how well PARSR captures human behavior in small tabular experiments (with 6 states) [108] as well as a scaled version of the experiments (with 121 states). We compare PARSR to a number of state of the art algorithms using replay on simple benchmark transfer learning (or *revaluation*) tasks in cognitive neuroscience (Figs. 2.2a, 2.2b, 2.2c) and a scaled up version of these tasks (Figs.2.3a, 2.3b, 2.3c). We find that PARSR matches human-like behavior as well as other algorithms' efficiency in learning speed of the prioritization-based algorithms. To clarify differences among the solutions different replay heuristics provide, we visualize which experiences are prioritized as more important to recall by different prioritization

algorithms (Fig. 2.3b). The code for implementing the algorithm and benchmark experiments is available at `https://github.com/s-a-barnett/PrioritizedSR`.

2.2 Background

In this section, we build on the single-agent reinforcement learning paradigm outlined in Sec. 1.2.1.

2.2.1 The Successor Representation

The successor representation (SR) [36] is defined as the discounted sum over expected future state visitations:

$$M_\pi(a, s, s') = \mathbb{E}_\pi \left[\sum_{t=T}^{\infty} \gamma^{t-T} \mathbf{1}(s_t = s') \mid s_T = s, a_T = a \right]. \tag{2.1}$$

Observe that in addition to the stochasticity inherent to the environment, this expectation is also taken with respect to a specific policy π, since the actions the agent takes also impact the states that are visited in the future. The SR is therefore *policy-dependent*: however, we often drop the π subscript if it is clear from the context, or when the SR is being learned on-policy, i.e., with respect to the actions taken by an SR-learning algorithm that do not correspond to a fixed policy.

Assuming that we take our initial action in expectation according to π, the SR is strongly connected to the one-step transition structure of the environment, T_π—it is in fact the Neumann series for this structure multiplied by the discount factor γ [127]:

$$\boldsymbol{M}_\pi := \sum_{a \in \mathcal{A}} \big(M_\pi(a, :, :)^\top \pi(a, :) \big) = \sum_{t=0}^{\infty} \gamma^t \boldsymbol{T}_\pi^t = (I - \gamma \boldsymbol{T}_\pi)^{-1}. \tag{2.2}$$

Combining Eq. 2.2 with Eq. 1.5, we are able to decompose the value function as

$$\boldsymbol{V}_\pi = \boldsymbol{M}_\pi \boldsymbol{R}_\pi, \tag{2.3}$$

where $\boldsymbol{R}_\pi \in \mathbb{R}^{|\mathcal{S}|}$ gives the expected one-step reward according to policy π. Similarly, after writing $\boldsymbol{R}_\pi$ as a vector of weights $\boldsymbol{w}(s')$, we can also express the state-action value function as a linear combination:[1]

$$Q_\pi(s, a) = \sum_{s'} M_\pi(a, s, s') \boldsymbol{w}(s'). \tag{2.4}$$

[1] We can extend this analysis to rewards defined on state-action pairs, $\boldsymbol{w}(s', a')$, and define our SR on 4-tuples as $M(a, s, a', s')$.

We can therefore split the task of learning Q_π for a given policy π into learning M_π and $\boldsymbol{w}$. If an agent has acted in state s according to π and received reward r, we can update the estimate of $\boldsymbol{w}(s)$ according to the delta rule

$$\boldsymbol{w}(s) \leftarrow \boldsymbol{w}(s) + \alpha_{\boldsymbol{w}} \big[r - \boldsymbol{w}(s) \big], \tag{2.5}$$

where $\alpha_{\boldsymbol{w}}$ is the learning rate for the reward weights.

One approach to learning $\boldsymbol{M}_\pi$ is to initially learn $\boldsymbol{T}_\pi$ (which can easily be achieved by keeping track of frequencies of state-to-state pair transitions), and then computing the inversion on the right-hand side of Eq. 2.2. Combining this with Eq. 2.5 yields the Model-Based SR (SR-MB) algorithm [127, Algorithm 2]. However, inverting $\boldsymbol{T}_\pi$ requires $\mathcal{O}\left(|\mathcal{S}|^3\right)$ operations, which can be costly if there are too many states and is therefore likely to be infeasible as a mechanistic account of how humans would compute a successor representation.

Alternatively, we can exploit the fact that the SR obeys a Bellman-like equation

$$M(a, s, :) = \mathbf{1}_s + \gamma \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} T_\pi(s, s') \pi(s', a') M(a', s', :), \tag{2.6}$$

where $\mathbf{1}_s \in \mathbb{R}^{|\mathcal{S}|}$ is a one-hot vector with a 1 on the index representing state s. Analogously to our combination of Eqs. 1.2 and 1.3, this leads to a temporal difference update similar to to the one for the value function (see Eq. 1.11):

$$M(a, s, :) \leftarrow M(a, s, :) + \alpha_M \underbrace{\big[\mathbf{1}_s + \gamma M(a', s', :) - M(a, s, :) \big]}_{\boldsymbol{\delta}_M}, \tag{2.7}$$

where $\alpha_M > 0$ is the SR learning rate. Combining Eq. 2.7 with Eq. 2.7 yields the Temporal Difference SR (SR-TD) algorithm [127, Algorithm 1].

2.2.2 Prioritized Replay

In addition to learning appropriate representations, we can also integrate the advantages of both model-based planning (flexibility) and model-free learning methods (speed and memory efficiency) by interleaving learning from real experience with learning from simulated experience, or *replay*, sampled from a model learned from previous interactions with the environment. In the deterministic setting considered in this paper, this model is a dictionary whose entries are state-action pairs, and whose values are the next state and received reward,

$$\mathit{Model}(s, a) = (r, s'). \tag{2.8}$$

Within the space of replay algorithms, there are a number of choices of which previous experience is chosen at each time step (for a broad treatment, refer to Sutton and Barto [146, Chapter 8]). The Dyna-Q algorithm [145, Alg. 4] samples past experience from its model uniformly at random. Combining Dyna-style replay mechanics with the temporal difference approach to SR learning yields the Dyna-SR algorithm [127, Algorithm 3].[2]

However, by prioritizing experience that previously led to larger changes in the learned representations, we can improve the efficiency of our replay algorithm. These changes are given by the absolute temporal difference prediction errors for the state-action value function $|\delta_Q|$, as computed in Alg. 3, line 6, and Alg. 4, line 7. Prioritized Sweeping (PS) [109, 119] maintains a queue ordered according to the most recent error $|\delta_Q|$ for each state-action pair, and pops from the queue after every time step. This is then propagated to states that precede the replayed state, allowing the largest changes to flow backwards through the model. In addition to the increased efficiency, this approach is consistent with human studies suggesting that larger prediction errors are followed by more offline replay, and offline replay of predecessors of states tagged with prediction error is correlated with future revaluation behavior [107].

2.3 Algorithms

Though replay alone improves an agent's ability to relearn local aspects of the task at hand, the representation of this task (namely, the Q function) is inflexible inasmuch as it can obscure the different kinds of changes that might take place. This can lead to slower learning, and does not correspond to the representations of such tasks used by biological agents [152, 48, 106].

By simultaneously updating both an SR and reward weights, we can enable flexibility to changes in rewards (reward transfer). However, changes in the transition structure (transition transfer) requires updating the SR for states with reevaluated transitions, via real or replayed experience. Thus, combining the representational flexibility of the SR with efficient forms of planning through replay achieves both reward and transition transfer.

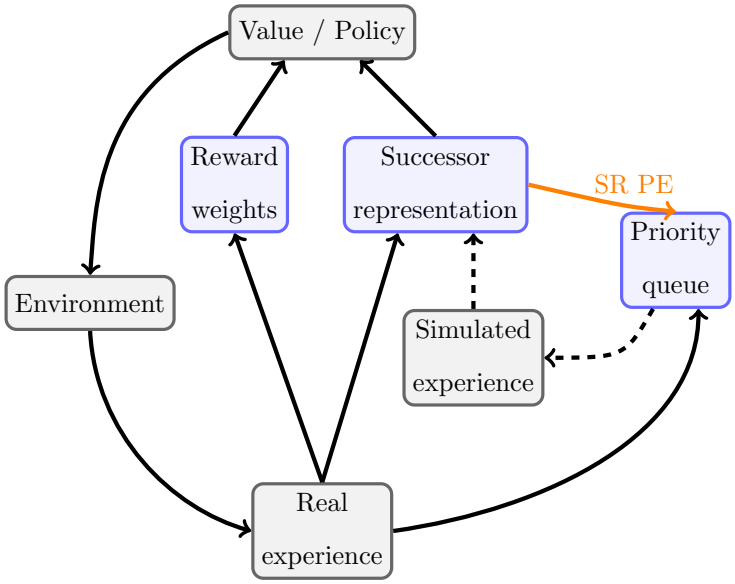
The Dyna-SR algorithm [127] discussed previously is one such hybrid. However, we hypothesized that sampling past experiences according to a prioritization schema, similar to prioritized sweeping, could further improve performance while still capturing human-like behavior.

[2]Technically, the model in the Dyna-SR algorithm stores a list of previously seen successor states and rewards for each state-action pair, which is then drawn based on a recency-weighted bias governed by an exponential distribution with rate $\lambda = 1/5$. In this paper, the difference between storing a list of values and storing an individual value is minimal.

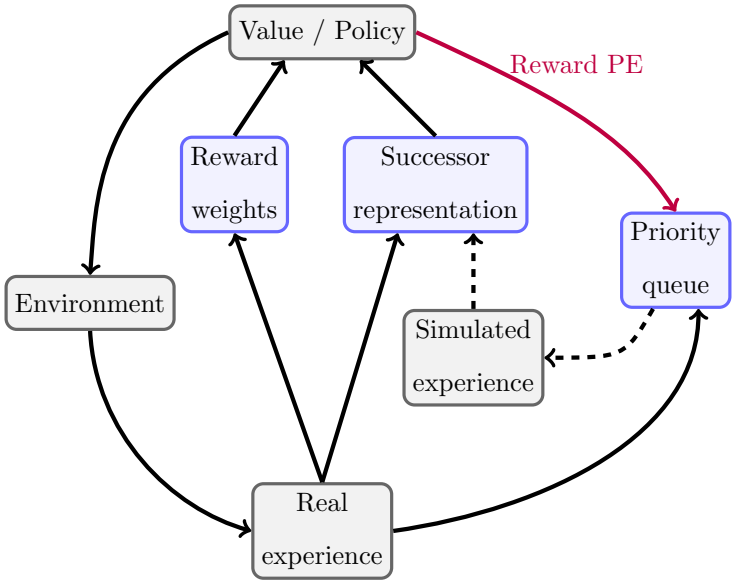
Algorithm 5 PARSR

Hyperparameters: Number of replay cycles n, exploration parameter ε, SR learning rate α_M, reward weights learning rate $\alpha_{\boldsymbol{w}}$, prioritization type *PriType*.
Initialize $M(a, s, s')$, $\boldsymbol{w}(s)$, *Model*(s, a) for all s, a and *PQueue* to empty.
while True **do**
    $s \leftarrow$ current (nonterminal) state.
    $Q \leftarrow \boldsymbol{w}(:)^\top M(a, s, :)$.
    $a \leftarrow \varepsilon$-*greedy*(s, Q).
    Take action a; observe reward, r, and state, s'.
    Model$(s, a) \leftarrow r, s'$.
    $Q \leftarrow \boldsymbol{w}(:)^\top M(a, s, :)$.
    $a' \leftarrow \arg\max_{a''}(Q(s', a''))$.
    $\boldsymbol{\delta}_M \leftarrow \big[\mathbf{1}_s + \gamma M(a', s', :) - M(a, s, :)\big]$. ▷ SR ($M$) prediction error.
    $M(a, s, :) \leftarrow M(a, s, :) + \alpha_M \boldsymbol{\delta}_M$.
    $\boldsymbol{w}(s) \leftarrow \boldsymbol{w}(s) + \alpha_{\boldsymbol{w}}\big[r - \boldsymbol{w}(s)\big]$.
    if *PriType* is M-PARSR **then**
        $p \leftarrow \|\boldsymbol{\delta}_M\|$
    else if *PriType* is Q-PARSR **then**
        $p \leftarrow \delta_Q \equiv \boldsymbol{\delta}_M^\top \boldsymbol{w} - \boldsymbol{w}(s) + r$ ▷ Q prediction error.
    Insert s, a into *PQueue* with priority p.
    loop n times
        if *PQueue* is not empty **then**
            $\overline{s}, \overline{a} \leftarrow$ *first*(*PQueue*).
        else
            $\overline{s} \leftarrow$ random previously observed state.
            $\overline{a} \leftarrow$ random action previously taken in $\overline{s}$.
        $\overline{r}, \overline{s}' \leftarrow$ *Model*$(\overline{s}, \overline{a})$.
        $Q \leftarrow \boldsymbol{w}(:)^\top M(\overline{a}, \overline{s}, :)$.
        $\overline{a}' \leftarrow \arg\max_{\overline{a}''}(Q(\overline{s}', \overline{a}''))$.
        $\boldsymbol{\delta}_M \leftarrow \big[\mathbf{1}_s + \gamma M(\overline{a}', \overline{s}', :) - M(\overline{a}, \overline{s}, :)\big]$.
        $M(\overline{a}, \overline{s}, :) \leftarrow M(\overline{a}, \overline{s}, :) + \alpha_M \boldsymbol{\delta}_M$.

To this end, we propose *PARSR*: Priority-Adjusted Replay for Successor Representations (Alg. 5 and Fig. 2.1, changes from Dyna-SR in blue). Unlike prioritized sweeping, which only relies on reward prediction errors (PE), PARSR can prioritize replay using PE for either the SR or reward weights. When using the latter priority measure for PARSR we call this variant *Q-PARSR*, and when using successor prediction errors ($\|\boldsymbol{\delta}_M\|$), we refer to the variant as *M-PARSR*. Each algorithm represents different choices about what kinds of experience to prioritize, potentially leading to different behavior and training times.



(a) M-PARSR



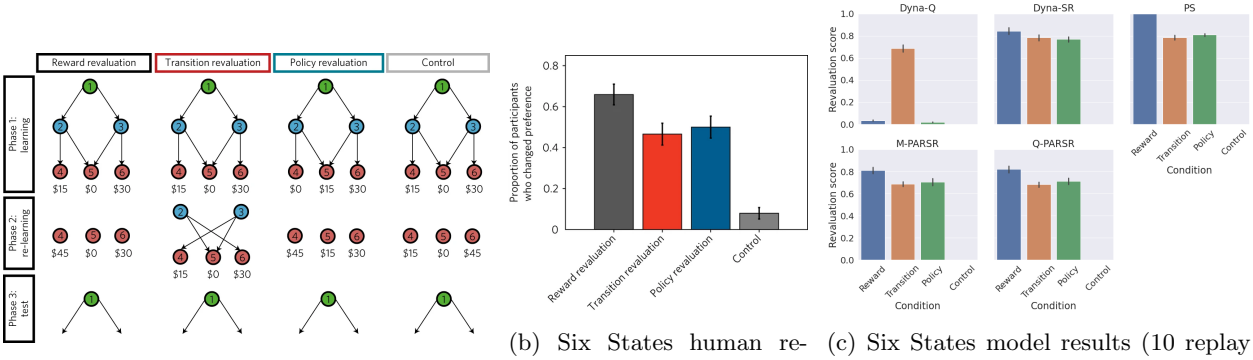
(b) Q-PARSR

Figure 2.1: Flowcharts for both variants of the PARSR algorithm. Components of the agent are in blue. **Real experience** (s, a, r, s') is fed into the priority queue, and used to update the **reward weights** $\boldsymbol{w}(s')$ and **successor representation** $M(a, s, s')$ via temporal difference learning. The **priority queue** returns **simulated experiences** $(\overline{s}, \overline{a}, \overline{r}, \overline{s}')$, which are used to provide further updates to the successor representation. M-PARSR (a) prioritizes according to the successor representation prediction error (SR PE), whereas Q-PARSR (b) prioritizes according to the reward prediction error (Reward PE). The reward weights and successor representation determine the state-action **value** function $Q(s, a) = \sum_{s'} M(a, s, s')\boldsymbol{w}(s')$, which in turn determines the **policy** for acting in the **environment**.

2.4 Experiments

We evaluate the performance of both variants of the PARSR algorithm in comparison to other SR-based and replay-based RL algorithms on revaluation tasks on different scales. Each experiment is evaluated over 10 seeds with $\varepsilon \in \{0.1, 0.3, 0.5, 1.0\}$. The Six States (resp. Six Rooms) experiment is performed for 100 (resp. 10) runs per seed, with 10 (resp. 1000) replay cycles per timestep for each algorithm. Note that PARSR has the same number of hyperparameters as other algorithms, so any improvement in human-like performance is not due to increased model complexity.

2.4.1 Six States



(a) Six States design.

(b) Six States human results.

(c) Six States model results (10 replay cycles).

Figure 2.2: Structure and results for the Six States experiment, reproduced from Experiment 2 in [108]. In (a), numbered circles denote different states, and arrows denote the unidirectional actions available at each state. For a given phase of a given condition, trials begin only in the earliest-stage states that are displayed in the figure for that condition and phase. (b) shows the proportion of participants in the human experiment ($n = 88$) who changed preference following the re-learning phase for each condition. Participants show greater ability to transfer in the case of reward revaluation than they do for the transition or policy conditions, though they are also capable of performing those tasks in some proportion. (c) reproduces these results for different algorithms.

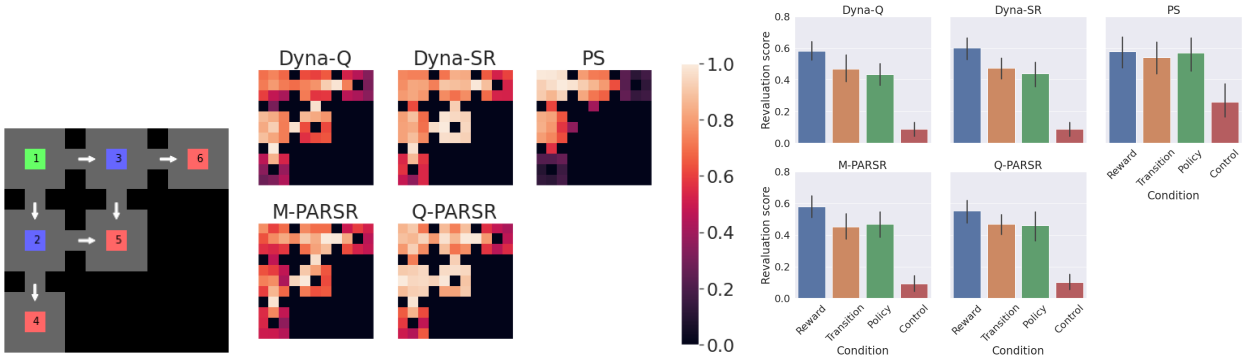
We first reproduce and extend the results of Experiment 2 of [108], which we refer to as the "Six States" experiment. In this decision-making task (in which, unlike Experiment 1, the participant takes actions at every step), participants complete four games, each corresponding to a different experimental condition (Fig. 2.2a). The six states of the environment are structured as a unidirectional, three-stage decision tree, where two actions are available at the first two stages and a scalar reward is received at the terminal states in the final stage.

Each task is divided into three phases: one must pass each phase three times consecutively in order to

progress to the next phase. In phase 1, participants are trained on a specific reward and transition structure. In phase 2, a change in either the reward or transition structure is changed, and participants learn about the changed structure *without revisiting the starting state*. Hence, participants do not get to experience these new contingencies following an action taken from the first stage. In phase 3, participants perform a single test trial beginning from the starting state, with the revaluation score corresponding to the probability (over multiple experimental runs) that the participant changes their action in state 1 between the end of phase 1 and the single trial in phase 3. Results from a human study in [108] show a greater revaluation score for the reward condition than transition or policy conditions, and no significant difference in revaluation between the latter two. Revaluation in the control condition is significantly lower than all of these.

For both Experiments 1 and 2, we find that both variants of PARSR are able to capture the human revaluation behavior in the task equally as well as Dyna-SR.

2.4.2 Six Rooms



(a) Six Rooms (121 states) design.

(b) Six Rooms priority visualization, phase 2 of transition revaluation.

(c) Six Rooms model results (1000 replay cycles).

Figure 2.3: Results for the Six Rooms experiment. (a) shows the map of the Six Rooms environment, with white arrows denoting "trapdoors" between rooms. (c) shows the revaluation scores for each model: all algorithms with the exception of PS attain results that are analogous to those achieved by human participants in Experiment 2 of [108] (Six States). (b) shows the relative frequency of the states prioritized during replay for the transition revaluation condition during phase 2. *M-PARSR has a much narrower focus on the bottlenecks between the rooms at which the revaluation is taking place, whereas Q-PARSR has a more uniform distribution over the prioritized states despite nonetheless employing a priority queue.* Both achieve similar performance in these tasks in spite of these differences.

We wanted to investigate whether the findings from the Six States experiment scaled to tasks with larger state spaces. To test this, we designed Six Rooms, a gridworld analog with 121 states. In this environment, the states in the Six States experiment correspond to the centers of the six rooms, laid out in a similar

structure to the smaller environment. Unidirectionality is enforced through one-way corridors between each room, in order to retain the solution structure of the smaller environment. Each of the conditions and phases of the Six States experiment can be defined analogously for the Six Rooms domain.

Despite their similar latent structure, the Six Rooms environment represents a greater challenge for the agents since moving between the rooms requires a sequence of *several* actions, thus requiring more updates to the agents' representations. To successfully pass each phase, therefore, requires not only that the agent navigates to the correct state given its starting state, but moreover that it do so using the shortest path. During phase 2, agents are initialized in the center states of the second-stage rooms and are required to navigate to center states of the correct final rooms using the quickest path. This matches the difficulty of the exclusion criteria across the four conditions.

Fig. 2.3c shows the results of the experiment for each algorithm. We observe that the revaluation scores for these tasks match that of the human performance on the Six States experiment for all but Priortized Sweeping and most faithfully by PARSR. This suggests that PARSR's performance scales to more complex tasks with a similar latent structure. This further offers the testable prediction that human behavior in the Six Rooms experiment should scale accordingly.

In Fig. 2.3b, we visualize the relative frequencies at which each state was prioritized by each algorithm during transition revaluation (phase 2). We observe a difference between the two PARSR variants in their prioritization strategies: while M-PARSR prioritizes replaying bottleneck states at which the revaluation is occurring, Q-PARSR's prioritization focus is more diffuse. Future work is required to test which conditions and tasks are best served by each prioritization scheme. For instance, adding meta-learning to PARSR could control which type of prioritization is appropriate depending on the task at hand.

2.5 Related work

Experience replay has been incorporated as a core part of many reinforcement learning algorithms, both in the tabular setting in algorithms such as Dyna [145, 83], and in deep reinforcement learning, where Deep Q-Networks (DQN) update the Q-network using a minibatch randomly sampled from a replay buffer [103].

Likewise, the use of prioritization based on prediction error as a heuristic for faster learning has also been explored in both settings. In the tabular setting, Prioritized Sweeping [109, 119] is the most prevalent of these, with a "small backups" variant improving performance by making computation time independent of the number of successor states [158]. In the deep reinforcement learning setting, Prioritized Experience

Replay [132] and its extensions [60, 74] also use a PE-based prioritization heuristic. Unlike in the tabular case, however, experiences are sampled at random with a probability proportional to the PE (with an importance-based sampling correction), as opposed to the experiences being popped from a queue.

The successor representation was first introduced in [36]. In the tabular setting, SR-based algorithms have been proposed and studied for their ability to capture neurally plausible, human-like behavior in transfer tasks [108, 127], planning and hippocampal replay [97], as well as cognitive fatigue and boredom [3].

Extensions to the successor representation can also be found in the deep reinforcement learning setting. The Deep Successor Representation architecture [79, 88] learns the SR as a deep neural network with pixels as its input, and outputs including a deep convolutional decoder to reproduce the input, a linear regressor to predict instantaneous rewards, and a feedforward neural network producing successor activations corresponding to each outcome. Alternatively, the Successor Features framework [16] generalizes its discounted sum from a one-hot vector representing a state visitation to a feature vector, where instantaneous rewards are linear in these features. This framework has been successfully combined with the Generalized Policy Improvement algorithm for combining solutions of previous tasks into a policy for the unseen task [16, 14, 22, 17, 27]. Successor features can also be used for discovering extended courses of actions known as *options*, an initial finite set of which can be combinatorially extended without extra learning [15, 87, 28].

The eigendecomposition of the SR matrix has also been an object of interest. It can be used to show the equivalence of the SR for a uniform policy to the graph Laplacian of the matrix of adjacent states in the environment [88, 87]. Moreover, the eigenvectors can be seen as a reflection of grid cell firing patterns in the brain used for navigation tasks [143]. For comprehensive overviews of the successor representation and its properties, refer to [48] and [106].

2.6 Discussion and Future Work

We have introduced PARSR: an algorithm that combines successor representation based learning with novel replay prioritization heuristics. By drawing from neuroscience [127, 46, 25] and human behavior studies [108], this algorithm is inspired by the principle of human-plausibility [105]. PARSR's two variants prioritize experience replay using either representation-based or reward-based prediction errors. Both PARSR variants show human-like behavior on benchmark tasks with 6 states (Figs. 2.2a, 2.2b, 2.2c) as well as as scaled tasks (the Six Rooms environment) with 121 states (Figs.2.3a, 2.3b, 2.3c). The latter offers novel predictions for human behavior in scaled experiments.

In future work could extend PARSR beyond tabular environments, as a deep RL algorithm with function approximation, similar to that of Prioritized Experience Replay [132]. Even within the tabular setting, the Prioritized Experience Replay algorithm suggests useful modifications that may also be of benefit to the PARSR algorithm, such as incorporating importance sampling to correct for the bias introduced by drawing based on prioritization.

In addition to extending PARSR to deep learning and more complex environments, future work would investigate further the nature of the two prioritization signals in PARSR variants, e.g., to investigate sequence memory activations at given moments in the task [97]. Moreover, we have visualized how error signals determine the relative frequency of prioritized experiences (Fig. 2.3b). Future work is required to investigate which problems are better served by which prioritization schemes. One solution is a novel algorithm combining PARSR with meta-learning of a control parameter learned across tasks and environments that, given the problem at hand, determines which PE signal is appropriate for efficient replay prioritization.

Chapter 3

A Pragmatic Account of the Weak Evidence Effect

> Well, he would [say that], wouldn't he?
>
> *Mandy Rice-Davies*

3.1 Introduction

Communication is a powerful engine of learning, enabling us to efficiently transmit complex information that would be costly to acquire on our own [153, 58]. While much of what we know is learned from others, it can also be challenging to know how to incorporate socially transmitted information into our beliefs about the world. Each source is a person with a "hidden agenda" encompassing their own beliefs and desires and biases, and not all information can be treated the same [63, 112]. For example, when deciding whether to buy a car, we may weight information differently depending on whether we heard it from a trusted family memory or the dealership, as we know the dealership is trying to make a sale. While such reasoning is empirically well-established—even young children are able to discount information from untrustworthy or unknowledgeable individuals [170, 141, 124, 102, 53, 56]—these phenomena have continued to pose a problem for formal models of belief updating, which typically take information at face value.

Recent probabilistic models of social reasoning have provided a mathematical framework for understanding how listeners ought to draw inferences from socially transmitted information. Rather than treating information as a direct observation of the true state of the world, social reasoning models suggest treating

the true state of the world as a *latent variable* that can be recovered by inverting a generative model of how an intentional agent would share information under different circumstances [57, 160, 167, 69, 9, 49, 50]. These models raise new explanations for classic effects in the judgment and decision-making literature, where behavior is often measured in social or linguistic contexts [123, 142, 110, 8, 100, 86].

Consider the *weak evidence effect* [99, 41, 84] or *boomerang effect* [122], a striking case of non-monotonic belief updating where weak evidence in favor of a particular conclusion may backfire and actually reduce an individual's belief in that conclusion. For example, suppose a juror is determining the guilt of a defendant in court. After hearing a prosecutor give a weak argument in support of a guilty verdict—say, calling a single witness with circumstantial evidence—we might expect the juror's beliefs to only be shifted weakly in support of guilt. Instead, the weak evidence effect describes a situation where the prosecutor's argument actually leads to a shift in the opposite direction—the juror may now believe that the defendant is more likely to be *innocent*.

Importantly, social reasoning mechanisms are not necessarily in conflict with previously proposed mechanisms for the weak evidence effect, such as algorithmic biases in generating alternative hypotheses [41, 33], causal reasoning about other non-social attributes of the situation [18] or sequential belief-updating [99, 156]. Both social and asocial models are able to account for the basic effect. To find *unique* predictions that distinguish models with a social component, then, we argue that we must shift focus from the *existence* of the effect to asking *under what conditions* it emerges. Social mechanisms lead to unique predictions about these conditions that purely asocial models cannot generate. In particular, if evidence comes from an intentional agent who is expected to present the strongest possible argument in favor of their case, then weak evidence would imply the absence of stronger evidence [52]; otherwise weak evidence may be taken more at face value. Thus, a pragmatic account predicts a systematic relationship between a listener's social expectations and the strength of the weak evidence effect:[1] *weak evidence should only backfire when the information source is expected to provide the strongest evidence available to them.*

In this paper, we proceed by first extending recent rational models of communication to equip speakers with persuasive goals (rather than purely informative ones) and present a series of simulations deriving key predictions from our model. We then introduce a simple behavioral paradigm, the *Stick Contest*, which allows us to elicit a participant's social expectations about the speaker alongside their inferences as listeners. Based

[1] Harris, Corner, and Hahn [55] presents a related model of the *faint praise* effect, where the omission of any stronger information that a speaker would be expected to know implies that it is more likely to be negative than positive (e.g. "James has very good handwriting.") Importantly, this effect is sensitive to the perceived expertise of the source; no such implication follows for unknowledgable informants [see also 64, 21, 53, for related inferences from omission].

on speaker expectation data, we find that participants cluster into sub-populations of *pragmatic* listeners or *literal* listeners, who expect speakers to provide strongly persuasive evidence or informative but neutral evidence, respectively. As predicted by the pragmatic account, only the first group of participants, who expected speakers to provide persuasive evidence, reliably displayed a weak evidence effect in their belief updates. Finally, we use these data to quantitatively compare our model against prior asocial accounts and find that a pragmatic model accounting for these hetereogenous groups is most consistent with the empirical data. Taken together, we suggest that pragmatic reasoning mechanisms are central to explaining belief updating when evidence is presented in social contexts.

3.2 Formalizing a pragmatic account of the weak evidence effect

To derive precise behavioral predictions, we begin by formalizing the pragmatics of persuasion in a computational model. Specifically, we draw upon recent progress in the Rational Speech Act (RSA) framework [45, 49, 135]. This framework instantiates a theory of recursive social inference, whereby listeners do not naively update their beliefs to reflect the information they hear, but explicitly account for the fact that speakers are intentional agents choosing which information to provide [52].

3.2.1 Reasoning about evidence from informative speakers

We begin by defining a pragmatic listener L who is attempting to update their beliefs about the underlying state of the world w (e.g. the guilt or innocence of the defendant), after hearing an utterance u (e.g. an argument provided by the prosecution). According to Bayes' rule, the listener's posterior beliefs about the world $P_L(w \mid u)$ may be derived as follows:

$$P_L(w \mid u) \propto P_S(u \mid w)P(w) \tag{3.1}$$

where $P(w)$ is the listener's prior beliefs about the world and the likelihood $P_S(u \mid w)$ is derived by imagining what a hypothetical speaker agent would choose to say in different circumstances. This term yields different predictions given different assumptions about the speaker, captured by different speaker utility functions U. In existing RSA models, the speaker is usually assumed to be *epistemically informative*, choosing utterances that bring the listener's beliefs as close as possible to the true state of the world, as measured by information-

theoretic surprisal:

$$P_S(u \mid w) \propto \exp\{\alpha U_{\text{epi}}(u; w)\}$$
$$U_{\text{epi}}(u; w) = \ln P_{L_0}(w \mid u) \tag{3.2}$$

where the free parameter $\alpha \in [0, \infty]$ controls the temperature of the soft-max function and U_{epi} denotes the utility function of an (epistemically) informative speaker. As $\alpha \to \infty$, the speaker increasingly chooses the single utterance with the highest utility, and as $\alpha \to 0$ the speaker becomes indifferent among utterances. If this hypothetical speaker, in turn, aimed to be informative to the same listener defined in Eq. 3.1, it would yield an infinite recursion: the RSA framework instead assumes that the recursion is grounded in a base case known as the "literal" listener, L_0, who takes evidence at face value:

$$P_{L_0}(w \mid u) \propto \delta_{[\![u]\!](w)} P(w). \tag{3.3}$$

Here, $[\![u]\!]$ gives the literal semantics of the utterance u, with $\delta_{[\![u]\!](w)}$ returning 1 if w is consistent with the state of affairs denoted by u, and 0 (or very small ϵ) otherwise.

3.2.2 Reasoning about evidence from motivated speakers

The epistemic utility defined in Eq. 3.2 aims only to produce assertions that most effectively lead to *true* beliefs. Often, however, speakers do not seek to neutrally inform, but to persuade in favor of a particular outcome or "hidden agenda." What is needed to represent such persuasive goals in the RSA framework? We begin by assuming that motivated speakers have a particular goal state w^* that they aim to induce in the listener, where w^* does not necessarily coincide with the true state of affairs w. This naturally yields a persuasive utility U_{pers} that aims to persuade the listener to adopt the intended beliefs w^*:

$$U_{\text{pers}}(u; w^*) = \ln P_{L_0}(w^* \mid u) \tag{3.4}$$

where we say an utterance u is strictly more persuasive than u' if and only if $U_{\text{pers}}(u \mid w^*) > U_{\text{pers}}(u' \mid w^*)$ (i.e. when the utterance results in the listener assigning higher probability to the desired state w^*). Following prior extensions of the speaker utility to other non-epistemic goals [e.g. 172, 173, 19], we then define a combined utility assuming the speaker aims to jointly fulfill persuasive aims (Eq. 3.4) while remaining consistent with

the true world state w (Eq. 3.2):

$$P_S(u \mid w, w^*) \propto \exp\{\alpha \cdot U(u; w, w^*)\} \tag{3.5}$$

$$U(u; w, w^*) = U_{\text{epi}}(u; w) + \beta \cdot U_{\text{pers}}(u; w^*) \tag{3.6}$$

where β is a parameter controlling the strength of the persuasive goal (we recover the standard epistemic RSA model when $\beta = 0$). This motivated speaker forms the foundation for a pragmatic model of the weak evidence effect.[2] A pragmatic listener L_1 who suspects that the utterance was generated by a motivated speaker with non-zero bias β is able to be "skeptical" of the speaker's agenda and discount their evidence accordingly:[3]

$$P_L(w \mid u, w^*, \beta) \propto P_S(u \mid w^*, w, \beta) \cdot P(w) \tag{3.7}$$

To see why this model allows evidence to backfire, note that the probability of different utterances are in competition with one another under the speaker model. In the case that w and w^* coincide, the speaker is expected to choose a utterance that is strongly supportive of that state; weaker utterances have a lower probability of being chosen. Conversely, if w^* deviates from the true state of affairs, stronger utterances in favor of w^* will be dispreferred (because they will be false and violate the epistemic term), hence weaker utterances are more likely. In this way, the absence of strong evidence from a speaker who would be highly motivated to show it statistically implies that no such evidence exists.

3.3 Experiment: The Stick Contest

Empirical studies of the weak evidence effect require a cover story to elicit belief judgments and manipulate the strength of evidence. Typically, this cover story is based on a real-world scenario such as a jury trial [99] or public policy debate [41], where participants are asked to report their belief in a hypothetical state such as the defendant's guilt or the effectiveness of the policy intervention. While these cover stories are naturalistic, they also introduce several complications for evaluating models of belief updating: participants may bring in different baseline expectations based on world knowledge and the absolute scalar argument strength of verbal statements is often unclear. To address these concerns, we introduce a simple behavioral paradigm

[2] Coincident with our work, [161] has proposed a similar formulation to explain how speakers may stretch the truth of epistemic modals like "possibly" or "probably."

[3] Although we formulate the listener's posterior as being conditioned on a *known* value of β, we can also consider the case in which the listener has a prior distribution over biases and can compute (marginal) posteriors accordingly – refer to Appendix A.5 for details.

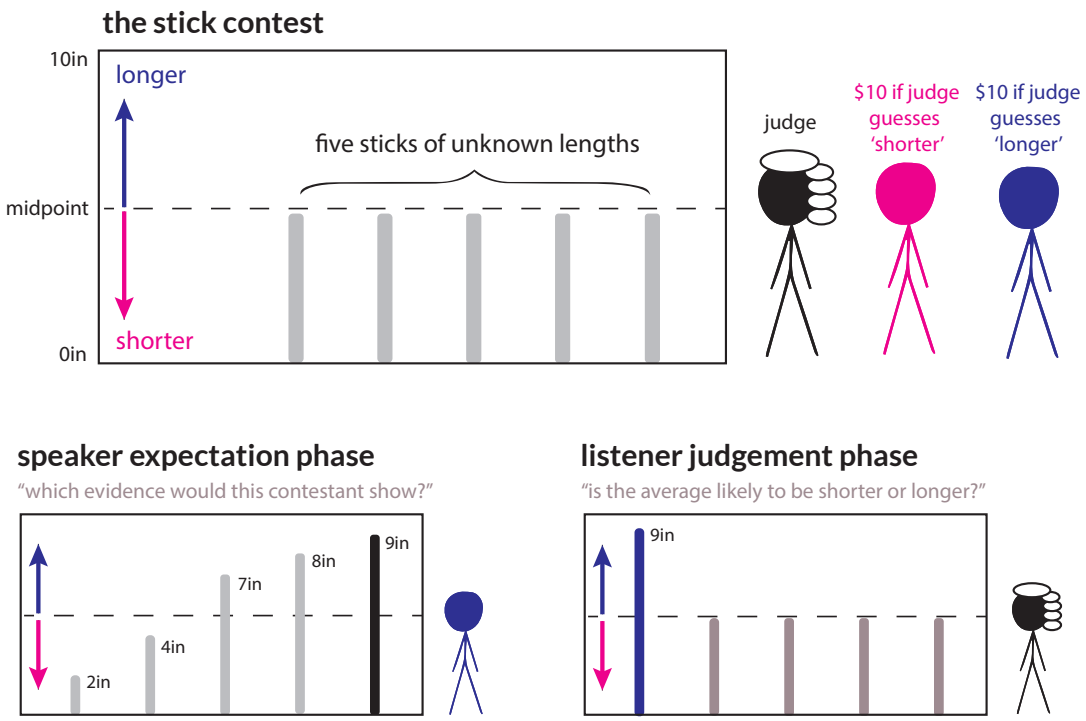


Figure 3.1: In the Stick Contest paradigm, participants are asked to determine whether a set of five hidden sticks is longer or shorter, on average, than a midpoint (dotted line) based on limited evidence from a pair of contestants. In the *speaker expectation* phase (left), participants were asked which one of the five sticks a given contestant would be most likely to show. In the *listener judgment* phase (right), participants were presented with a sequence of sticks from each contestant and asked to judge the likelihood that the overall sample is "longer."

called *the Stick Contest* (see Fig. 3.1). This game is inspired by a courtroom scenario: two contestants take turns presenting competing evidence to a judge, who must ultimately issue a verdict. Here, however, the verdict concerns the average length of $N = 5$ sticks which range from a minimum length of 1" to a maximum length of 9". These sticks are hidden from the judge but visible to both contestants, who are each given an opportunity to reveal exactly one stick as evidence for their case. As in a courtroom, each contestant has a clear agenda that is known to the judge: one contestant is rewarded if the judge determines that the average length of the sticks is longer than the midpoint of 5" (shown as a dotted line in Fig. 3.1), and the other is rewarded if the judge determines that the average length of the sticks is shorter than the midpoint.

This paradigm has several advantages for comparing models of the weak evidence effect. First, unlike verbal statements of evidence, the scale of evidence strength is made explicit and provided as common knowledge to the judge and contestants. The strength of a given piece of evidence is directly proportional to

the length of the revealed stick, and these lengths are bounded between the minimum and maximum values. Second, while previous paradigms have operationalized the weak evidence effect in terms of a sequence of belief updates across multiple pieces of evidence (e.g. where the first piece of evidence sets a baseline for the second piece of evidence), common knowledge about the scale allows the weak evidence effect to emerge from a single piece of evidence. This property helps to disentangle the core mechanisms driving the weak evidence effect from those driving *order effects* [e.g. 156].

3.3.1 Participants

We recruited 804 participants from the Prolific crowd-sourcing platform, 723 of whom successfully completed the task and passed attention checks (see Appendix A.1). The task took approximately 5 to 7 minutes, and each participant was paid $1.40 for an average hourly rate of $14. We restricted recruitment to the USA, UK, and Canada and balanced recruitment evenly between male and female participants. Participants were not allowed to complete the task on mobile or to complete the experiment more than once.

3.3.2 Design and procedure

The experiment proceeded in two phases: first, a *speaker expectation* phase, and second, a *listener judgment* phase (see Fig. 3.1). In the speaker expectation phase, we placed participants in the role of the contestants, gave them an example set of sticks $\{2, 4, 7, 8, 9\}$ and asked them which ones they believed each contestant would choose to show, in order of priority. In the listener judgment phase, we placed participants in the role of the judge and presented them with a sequence of observations. After each observation, they used a slider to indicate their belief about the verdict on a scale ranging from 0 ("average is definitely shorter than five inches") to 100 ("average is definitely longer than five inches"). It was stated explicitly that the judge knows that there are exactly five sticks, and that each contestant's incentives are public knowledge. After each phase, we asked participants to explain their response in a free-response box (see Tables A.2 and A.3 for sample responses).

This within-participant design allowed us to examine individual co-variation between the strength of a participant's weak evidence effect in the listener judgment phase and their beliefs about the evidence generation process in the speaker expectation phase. Critically, while the set of candidate sticks in the speaker expectation phase was held constant across all participants for consistency, the strength of evidence we presented in the listener judgment phase was manipulated in a between-subjects design. The length of

the first piece of evidence was chosen from the set $\{6, 7, 8, 9\}$ when the long-biased contestant went first, and from the set $\{4, 3, 2, 1\}$ when the short-biased contestant went first, for a total of 4 possible "strength" conditions (measured as the distance of the observation from the midpoint; we assigned more participants to the more theoretically important "weak evidence" condition, i.e. $\{4, 6\}$, to obtain a higher-powered estimate). The order of contestants was counterbalanced across participants and held constant across the speaker and listener phase.[4] Although it was not the focus of the current study, we also presented a second piece of evidence from the other contestant to capture potential order effects (see Appendix A.2 for preliminary analyses).

3.4 Results

3.4.1 Behavioral results

Before quantitatively evaluating our model, we first examine its key qualitative predictions. Do participants exhibit a weak evidence effect in their listener judgments at all, and if so, to what extent is variation in the strength of the effect related to their expectations about the speaker? We focus on each participant's first judgment, provided after the first piece of evidence in the listener phase. This judgment provides the clearest view of the weak evidence effect, as subsequent judgments may be complicated by order effects. We constructed a linear regression model predicting participants' continuous slider responses. We included fixed effects of evidence strength as well as expectations from the speaker phase (coded as a categorical variable, expecting strongest evidence vs. expecting weaker evidence), and their interaction, along with a fixed effect of whether the first contestant was "short"-biased or "long"-biased. Because the design was fully between-participant (i.e. each participant only provided a single slider response as judge), no random effects were supported.

As predicted, we found a significant interaction between speaker expectations and evidence strength, $t(718) = 5.2,\ p < 0.001$; see Fig. 3.2. For participants who expected the speaker to provide the strongest evidence (485 participants or 67% of the sample), weak evidence in favor of the persuasive goal backfired and actually pushed beliefs in the opposite direction, $m = 34.7$, 95% CI: $[32.3, 37.3], p < 0.001$. Meanwhile, for participants who expected speakers to "hedge" and not necessarily show the strongest evidence first (238 participants, or 33% of the sample), no weak evidence effect was found ($m = 50.1$, group difference $= -15.4$, post-hoc $t(367) = -6.3,\ p < 0.001$.) We found only a marginally significant asymmetry in slider bias, $p = 0.056$,

[4] An earlier iteration of our experiment only used a `long`-biased speaker; we report results from this version in Appendix A.4.

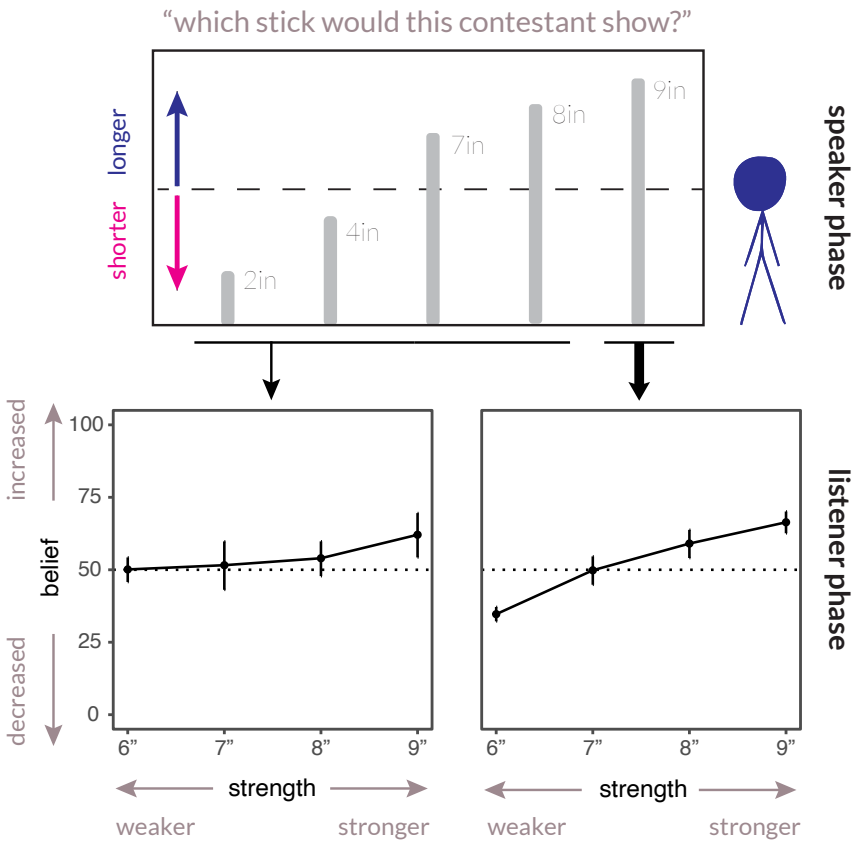


Figure 3.2: Individual differences in the weak evidence effect are predicted by pragmatic expectations. Dotted line represents neutral or unchanged beliefs. Error bars are bootstrapped 95% CIs (see Fig. A.3 for raw distributions).

with short-biased participants giving slightly larger endorsements ($m = 1.6$ slider points) across the board.

3.4.2 Model simulations

The qualitative effect observed the previous section is consistent with our pragmatic account: weak evidence only backfired for participants who expected speakers to provide the strongest available. In this section we conduct a series of simulations to explicitly examine the conditions under which this effect emerges from our model of recursive social reasoning between a speaker (who selects the evidence) and a listener (who updates their beliefs in light of the evidence). Our task is naturally formalized by defining the possible utterances $u \in \mathcal{U}$ as the possible lengths of individual sticks the speaker must choose between, the world state w as the true set of sticks, and the persuasive goals $w^* \in \{\texttt{longer}, \texttt{shorter}\}$ as a binary proposition corresponding to each speaker's incentive. Because the speaker only has access to true utterances, all

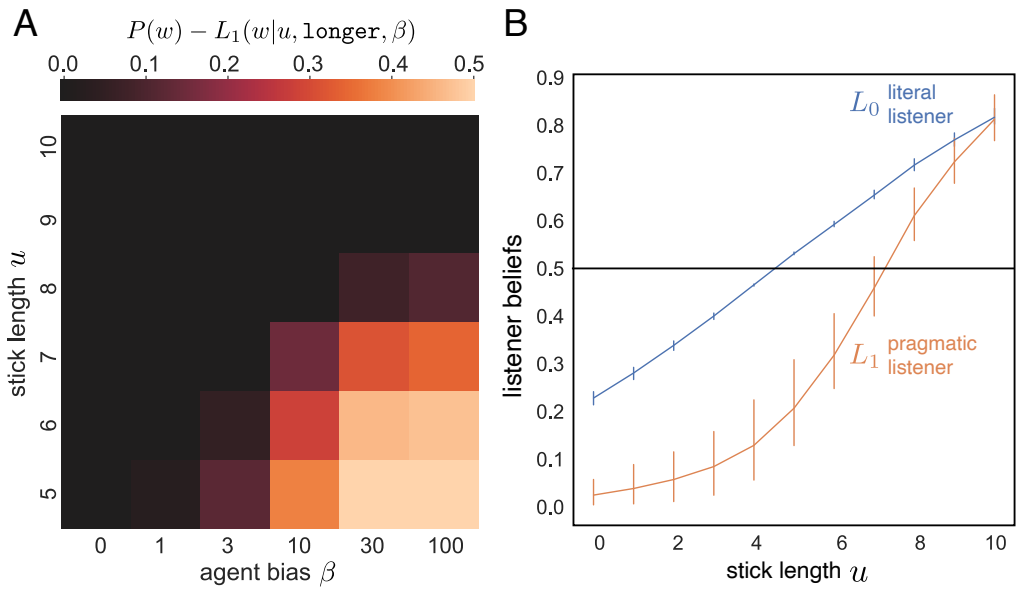


Figure 3.3: *Model simulations.* (A) Our pragmatic listener model predicts a weak evidence effect for a broader range of evidence strengths at higher perceived speaker bias β. The color scale represents the extent to which the listener's posterior beliefs decrease in light of positive evidence, where the black region represents conditions under which no weak evidence effect is predicted. (B) Posterior beliefs of literal and pragmatic listener models as a function of evidence from long-biased speaker. Horizontal line represents prior beliefs. Error bars are given by 10-fold cross-validation across parameter fits on different subsets of our behavior data, with average $\overline{\beta} = 2.03$ and response offset $\overline{o} = -0.13$ (translating the curve down).

utterances have equal epistemic utility (i.e. the speaker must show one of the five *actual* sticks,[5] which has the epistemic effect of reducing uncertainty about the identity of exactly one stick). Hence, the combined utility (Eq. 3.6) simplifies to the following:

$$S(u \mid w, w^*, \beta) \propto \exp\{\alpha \cdot \beta \cdot \ln L_0(w^* \mid u)\} \tag{3.8}$$

and the persuasive utility of an utterance is monotonic in the stick length (see Appendix A.3 for complete proofs). Note that when $\beta = 0$, the pragmatic listener L_1 expects the speaker preferences to be uniform over true evidence, $S_1(u \mid w, w^*, \beta = 0) = \text{Unif}(u)$, thus reducing to the literal listener L_0. When $\beta \to \infty$, the pragmatic listener expects the speaker to maximize utility and choose the single strongest piece of evidence.[6]

In our simulations, we present the listener models with different pieces of evidence $u \in \{5, 6, 7, 8, 9, 10\}$

[5]For related tasks studying outright lying, see Oey, Schachner, and Vul [113], Ransom, Voorspoels, Perfors, and Navarro [125], and Franke, Dulcinati, and Pouscoulous [44] and Oey and Vul [114]. For a more comprehensive and multidisciplinary overview of varieties of deception and misleading, see Saul [131] and Meibauer [101].

[6]Because the product $\alpha \cdot \beta$ is non-zero only if the persuasion weight β is non-zero, these two parameters are redundant in our task. We thus treat their product as a single free parameter, effectively fixing $\alpha = 1$. It is possible that a near-zero α (e.g. low effort from participants) may make it difficult to empirically detect a non-zero β term in our model comparison below, but this would work against our hypothesis.

and manipulate β, which represents the degree to which the pragmatic listener L_1 expects the speaker S to be motivated to show data that prefers target goal state $w^* = \texttt{longer}$ (the case for `shorter` is analogous). We operationalize the size of the weak evidence effect as the decrease in belief for a proposition given positive evidence supporting that proposition. For example, if observing a stick length of 6" *decreased* the listener's beliefs that the sample was longer than 5" from a prior belief of $P(\texttt{longer}) = 0.5$ to a posterior belief of $P(\texttt{longer} \mid u = 6) = 0.4$, then we say the size of the effect is $0.5 - 0.4 = 0.1$.

First, we observe that when $\beta = 0$ (Fig. 3.3A, left-most column), no weak evidence effect is observed: the listener interprets the evidence literally. However, as the perceived bias of the speaker increases, we observe a weak evidence effect emerge for shorter sticks. When the perceived bias grows large (e.g. $\beta = 100$, right-most column), the weak evidence effect is found over a broad range of evidence: if the listener expects the speaker to show the single strongest piece of evidence available, then even a stick of length 8" rules out the existence of any stronger evidence, shifting the possible range of sticks in the sample. To further understand this effect, we computed the beliefs of literal (L_0) and pragmatic (L_1) listener models as a function of the evidence they've been shown (Fig. 3.3B). While the literal listener predicts a near-linear shift in beliefs as a function of positive or negative evidence, the pragmatic listener yields a sharper S-shaped curve reflecting more skeptical belief updating.

3.4.3 Quantitative Model Comparison

Our behavioral results suggest an important role for speaker expectations in explanations of the weak evidence effect, and our simulations reveal how a pragmatic listener model derives this effect from different expectations about speaker bias. In this section, we compare our model against alternative accounts by fitting them to our empirical data (see Appendix A.5 for details).

Fitting the RSA model to behavioral data We considered several variants of the RSA model, which handled the relationship between the speaker and listener phase in different ways. The simplest variant, which we call the *homogeneous* model, assumes the entire population of participants is explained by a pragmatic model ($z = L_1$) with an unknown bias. It is homogeneous because the same model is assumed to be shared across the whole population. The second variant, which we call the *heterogeneous* model, is a mixture model where we predicted each participant's response as a convex combination of the L_0 and L_1 models with mixture weight p_z (i.e. marginalizing out latent assignments z_i). In the third variant, which we call the *speaker-dependent* model, we explicitly fit different mixture weights depending on the participant's response

Model	Variant	Likelihood	WAIC	PSIS-LOO
A&A	Homogeneous	-28.1	57.7 ± 9.9	28.8 ± 9.9
MAS	Homogeneous	8.2	-13.3 ± 9.6	-6.6 ± 9.6
	Heterogeneous	8.2	-11.3 ± 9.5	-5.6 ± 9.5
RSA	Homogeneous	8.1	-13.3 ± 9.5	-6.7 ± 9.5
	Heterogeneous	8.1	-10.5 ± 9.3	-5.2 ± 9.3
	Speaker-dependent	**12.0**	**-16.4** ± 9.1	**-9.2** ± 9.1

Table 3.1: Results of the model comparison, including the likelihood achieved by the best-fitting model as well as the WAIC, and PSIS-LOO (± standard error), which penalize for model complexity.

in the speaker expectations phase. Rather than learning a single mixture weight for the entire population, this variant learns independent mixture weights for different sub-groups z_j, defined by the different sticks j that participants chose in the speaker phase. This model asks whether conditioning on speaker data allows the model to make sufficiently better predictions about the listener data.

Fitting anchor-and-adjust models to empirical data The most prominent family of *asocial* models accounting for the weak evidence effect are *anchor-and-adjust* (AA) models. In these models, individuals compare the strength of new evidence u against a reference point R and adjust their beliefs $P(w|u)$ up or down accordingly:

$$P(w|u) = P(w) + \eta \cdot (s(u) - R), \tag{3.9}$$

where $s(u)$ is the strength of the evidence, and η is an adjustment weight. In the simplest variant [62], the reference point and scaling are fixed to a neutral baseline $\eta = P(w) = 1 - P(w) = .5$ and $R = 0$. In a more complex variant, beliefs are not updated from a *neutral* baseline but instead relative to more stringent level known as the argument's "minimum acceptable strength" [MAS; 99], which is treated as a free parameter: $R \sim \text{Unif}[-1, 1]$. In this case, positive evidence that falls short of R may nonetheless be treated as negative evidence and decrease the listener's beliefs. Although the anchor is typically taken to be a specific earlier observation, it may be interpreted in the single-observation case as the participant's implicit or imagined expectations from the task instructions and cover story. Prior work using anchor-and-adjust models would not predict a relationship between behavior in the speaker phase and in the listener phase. We thus evaluated a homogeneous *AA* model, a homogeneous *MAS* model, and a heterogeneous mixture model predicting responses as a convention combination of the two.

Comparison Results We examined several metrics to assess the relative performance of these models.[7] First, as an absolute goodness of fit measure, we found the parameters that maximized the model likelihood (see Table 3.1). As a Bayesian alternative, which penalizes models for added complexity, we also considered a measure using the full posterior,[8] the Watanabe-Akaike (or Widely Applicable) Information Criterion [163, 47]. The WAIC penalizes model flexibility in a way that asymptotically equates to Bayesian leave-one-out (LOO) cross-validation [1, 47], which we also include in the form of the PSIS-LOO measure [PSIS stands for Pareto Smoothed Importance Sampling, a method for stabilizing estimates 159]. These comparison criteria (Table 3.1) suggest that the added complexity of the speaker-dependent RSA model is justified: it outperforms all *asocial* variants. For this speaker-dependent model, we found a maximum *a posteriori* (MAP) estimate of $\hat{\beta} = 2.26$, providing strong support for a non-zero persuasive bias term. We found that the pragmatic L_1 model best explained the judgments of participants who expected the strongest evidence to be shown during the speaker phase (mixture weight $\hat{p}_z = 0.99$) while the literal L_0 model best explained the judgments of participants who expected weaker sticks to be shown (mixture weight $\hat{p}_z = 0.1$). Full parameter posteriors are shown in Fig. A.5.

3.5 Discussion

Evidence is not a direct reflection of the world: it comes from somewhere, often from other people. Yet appropriately accounting for social sources of information has posed a challenge for models of belief-updating, even as increasing attention has been given to the role of pragmatic reasoning in classic phenomena. In this paper, we formalized a pragmatic account of the *weak evidence effect* via a model of recursive social reasoning, where weaker evidence may backfire when the speaker is expected to have a persuasive agenda. This model critically predicts that individual differences in the weak evidence effect should be related to individual differences in how the speaker is expected to select evidence. We evaluated this qualitative prediction using a novel behavioral paradigm – the Stick Contest – and demonstrated through simulations and quantitative model comparisons that our model uniquely captures this source of variance in judgments.

Several avenues remain important for future work. First, while we focused on the initial judgment as the purest manifestation of the weak evidence effect, subsequent judgments are consistent with the *order effects* that have been the central focus of previous accounts [see Appendix A.2; 5, 34, 156]. Thus, we

[7] All models were implemented in WebPPL [51]; code for reproducing these analyses is available at `https://github.com/s-a-barnett/bayesian-persuasion`.

[8] We drew 1,000 samples from the posterior via MCMC across four chains, with a burn-in of 7,500 steps and a lag of 100 steps between samples.

view our model of social reasoning as capturing an orthogonal aspect of the phenomenon, and further work should explicitly integrate computational-level principles of social reasoning with process-level mechanisms of sequential belief updating. Second, our model provides a foundation for accounting for related *message involvement* effects (e.g., emotion, attractiveness of source), *presentation* effects (e.g. numerical vs. verbal descriptions), and *social affiliation* effects (i.e., whether the source is in-group) that have been examined in real-world settings of persuasion [e.g. 96, 37, 20, 115, 29, 40], These settings also involve uncertainty about the *scale* of possible argument strength, unlike the clearly defined interval of lengths in our paradigm. Third, while the weak evidence effect emerges after a single level of social recursion, it is natural to ask what happens at higher levels: what about a more sophisticated speaker who is *aware* that weak evidence may lead to such inferences? Our paradigm explicitly informed participants of the speaker bias, but uncertainty about the speaker's hidden agenda may give rise to a *strong* evidence effect [120], where speakers are motivated to *avoid* the strongest arguments to appear more neutral (see Appendix A.5). Based on the self-explanations we elicited (Table A.2), it is possible that some participants who expected less strong evidence were reasoning in this way. These individual differences are consistent with prior work reporting heterogeneity in levels of reasoning in other communicative tasks [e.g. 43].

We used a within-participant individual differences design for simplicity and naturalism, but there are also limitations associated with this design choice. For example, it is possible that the group of participants who expected weaker evidence to be shown first could be systematically different from the other group in some way, such as differing levels of inattention or motivation, that explains their behavior on *both* speaker and listener trials. We aimed to control for these factors in multiple ways, including strict attention checks (Appendix A.1) and self-explanations (Tables A.2 and A.3), which suggest a thoughtful rationale for expecting weaker evidence. However, an alternative solution would be to explicitly manipulate social expectations about the speaker in the cover story (e.g. training participants on speakers that tend to show weaker or stronger evidence first). Such a design would license stronger causal inferences, but would also raise new concerns about exactly what is being manipulated. A second limitation of our design is that the speaker phase was always presented before the listener phase. It is already known that the order of these roles may affect participants' reasoning [e.g. 139, 136], but asocial accounts of the weak evidence effect would not predict any relationship between speaker and listener trials under *either* order. Hence, we chose the order we thought would minimize confusion about the task; it is not our goal to suggest that social reasoning is spontaneous or mandatory, and we expect that social-pragmatic factors may be more salient in some contexts than others [e.g. when evidence is presented verbally vs. numerically, as in 96].

Probabilistic models have continually emphasized the importance of the data generating process, distinguishing between assumptions like *weak* sampling, *strong* sampling, and *pedagogical* sampling [65, 136, 151, 150]. Our work considers a fourth sampling assumption, *rhetorical sampling*, where the data are not necessarily generated in the service of pedagogy but rather in the service of persuasive rhetoric. Critically, although we formalized this account in a recursive Bayesian reasoning framework, insights about rhetorical sampling are also compatible with other frameworks: for example, work in the anchor-and-adjust framework may use similar principles to derive a relationship between information sources and reference points.

Such socially sensitive objectives may be particularly key in the context of developing artificial agents that are more closely aligned with human values. As an example, the *AI safety via debate* framework [67] involves agents compete in a debate game to produce the most true, useful information for a human to judge in a decision-making task. While the initial version of this model posed agents playing a zero-sum game using Monte Carlo Tree Search [24], it is plausible that an human-plausible agent model with a theory of mind about the target of persuasion can produce more useful information. This would be in line with other research indicating the importance of human-like AI models for tasks in which humans and AI systems must cooperate [26, 61].

Chapter 4

Measuring Cooperation with Counterfactual Planning

> Antigonus, leader of Socho, received [his Torah] from Shimon the Just. He used to say: "Don't be like servants who serve their master for the sake of receiving a reward; instead be like servants who serve their master with the understanding that they will not receive a reward. And let the awe of heaven be upon you."
>
> *Pirkei Avot 1:3*

4.1 Introduction

In the trees of the Tai Forest in Côte d'Ivoire, chimpanzees hunt for red colobus monkeys in groups. Each chimpanzee shares the goal of hunting the monkey, and each chimpanzee benefits from the participation of the other chimpanzees in order to increase the likelihood that the prey is caught. Therefore, each chimpanzee is acting in a way that is conducive to the good of the group—this would appear to be a paradigmatic case of cooperative behavior.

However, there is another characterization of this sequence of events [154]. One chimpanzee initiates the

hunt in the knowledge that other chimpanzees are in the area, and then each other chimpanzee will in turn take the position that best maximizes its own likelihood of catching the prey. This has the cumulative effect of each chimpanzee blocking the monkey's next best path of escape. Importantly, each of the chimpanzees takes these actions individually and makes their plans solely according to its own self regard; there is no central planning.

A similar dynamic can arise within artificial systems. Although central planning is possible in theory and may be more likely to lead to desirable outcomes, due to its computational demand the approach is often eschewed in favor of agents who learn and act *independently* in an environment without regard to the other agents' utilities [42]. In many cases, this can still lead to an outcome that is beneficial to all of the agents [147].

In order to evaluate the cooperativeness of group behavior in both artificial multi-agent systems and biological species, we need to be able to measure the cooperativeness of these systems [32, 31]. However, as the preceding examples show, behavior that increases the total utility of the group is not necessarily cooperative—in other words, the cooperativeness of behavior is *underdetermined* by the actual sequence of events [116].

Previous work studying cooperation in artificial systems has focused on the design of environments within which cooperation can be understood, using these to investigate what mechanisms can drive cooperation [7, 80, 66, 68, 118, 39, 38]. However, cooperative behavior is typically either declared so by fiat, or is defined only in relation to the actual group outcome, without reference to the actions of uncooperative agents. An alternative to this is to measure the alignment between individual and collective interests in the system as a whole, such as through the *price of anarchy* [77] or the *self-interest level* [168].

In this paper, we propose a family of scalar measures of cooperation capable of precluding cases such as that of unintended mutual benefit by being *counterfactually contrastive*: we subtract from the group's total utility the amount that would have been attained had the agent in question acted purely in their self-interest. Our approach is agnostic to the mechanisms that distinguish between cooperative and competitive modes of group behavior [75, 149, 148], and it does not require any manipulations of the external rewards in the environment [91]. Moreover, we allow our measure to be *contextual*, in that it is relative to other agents' behavior, as well as *customizable* with respect to the time and space horizons, which can help to disambiguate other gray areas of cooperative behavior that have been previously studied.

We define our measure on stochastic games, a formalization of multi-agent systems that allow for the application of our measure on a broad class of artificial agents, as well as biological agents that can be

modelled in this way [138]. Using this definition, we evaluate the behavior of multiple classes of agents with different types of behavior in tabular social dilemmas, a common test bed in a variety of disciplines for understanding cooperation [7]. We show that many of these behaviors are no longer regarded as cooperative when our measure is applied to it, and other seemingly uncooperative behaviors become otherwise according to our measure. Crucially, by making explicit the components of the measure, our measure can provide an interpretable explanation of why behavior is cooperative or uncooperative. Moreover, by making the choice of social welfare function one of these components, our measure also explains the respect in which this behavior is cooperative, either by achieving a greater total utility, or a more equitable or fair outcome.

4.2 The challenges of defining cooperation

Cooperation has long been a subject of study in disciplines ranging from philosophy and economics to evolutionary biology and cognitive science [7, 154, 157, 78]. In these disciplines, we investigate the mechanisms that allow for cooperative behavior, the degree to which it leads to greater flourishing for a system as a whole, and what the motivations are for cooperating in the first place. In turn, this allows us to build artificial multi-agent systems that display cooperative properties, arguably essential as we begin to deploy progressively more complex AIs in the real world.

In order to study cooperation with computational models, an initial approach is simply to declare behaviors as cooperative or defecting by fiat. For example, in the Prisoner's Dilemma, the classic one-shot social dilemma game, the available actions to each player are to "Cooperate" or "Defect". The conclusions drawn from analyses of this game are subsequently generalized about cooperation as a broader concept [7]. However, this approach fails as we begin to examine systems acting in more complex environments that are capable of a richer range of behaviors: these behaviors will arguably now be cooperative to different *degrees*, with the cooperativeness of each behavior not necessarily being obvious [80].

Hence the need to define a *measure* on the cooperativeness of behavior. At a first pass, we might do this by simply evaluating the sum of all utilities attained by the group, also referred to as the *utilitarian welfare* [66, 81]; other welfare metrics such as fairness or sustainability could also be considered [6]. One drawback of this approach is that the cooperativeness of behavior is defined on groups as a whole, whereas it would be desirable for a measure to tell us if one agent were acting *more* cooperatively than others within the group.

More importantly, however, solely evaluating the actual outcome erroneously includes cases such as the aforementioned chimpanzee group hunting in which a mutually beneficial outcome results from individual

agents acting solely in their own self-regard. A related phenomenon occurs in evolutionary biology in which two species feed upon the waste product of the other: this is known as byproduct reciprocity. Unless this behavior is selected for *because* of the beneficial effect on the recipient (or at least partially because of this effect), this is not classed as cooperation [166].

Another issue when defining cooperation relates to the time horizon over which it is evaluated. The utility accrued from a group behavior, either to the individual or the entire group, may vary in its magnitude and valence over time. For example, consider the case of costly punishment as a means of enforcing a societally-beneficial norm: by invoking the punishment in response to a violation, the whole group is harmed in the short term, but with the expectation that these harms will be compensated for by a mutually beneficial outcome in the longer term. The degree to which this punishment is cooperative is therefore ambiguous without reference to this time horizon. A related notion in evolutionary biology is that of reciprocal altruism, in which individuals take turns helping each other in a costly way with the expectation that they will be helped in the future [155]. The term "altruism" is commonly taken to be a misnomer in relation to this phenomenon [54, 166], and this mistake can be clarified with appeal to the time horizon in question: while "reciprocally altruistic" behavior is costly to the agent performing it in the short term, in the long term we expect that the reciprocated benefits will justify this cost, so that the behavior can eventually be considered as self-interested.

4.3 Desiderata for a measure of cooperation

To address these common pitfalls, we divide the desiderata for a measure of cooperation into three broad categories: that it should be *counterfactually contrastive*, *customizable*, and *contextual*.

Counterfactually contrastive The absolute returns in total utility can be a misleading guide to the cooperativeness of a system: in certain situations, these returns might result without any cooperation taking place. We call a measure *counterfactually contrastive* if it sets as a baseline the behavior of agent(s) acting uncooperatively.

However, defining this baseline is not straightforward. For instance, consider a process-level approach whereby we contrast the agent with a counterpart that has had the mechanisms which allow for it to cooperate removed. For a human-like agent, the relevant cognitive mechanism in question might be the capacity to form a theory-of-mind about other agents [23, 175], so that cooperativeness would be measured by the difference between the actual outcome for the group and the outcome for the group if that agent's

theory-of-mind capabilities were removed.

This process-level approach has two limitations. Firstly, it would mean that each class of agent would have a different version of the measure applied to it depending on what its cooperative mechanism is—a measure would ideally be agnostic to these process-level details. Secondly, even within a particular class of agents it is non-trivial to specify what these mechanisms are. In the above example, while a theory-of-mind might be necessary for human-like cooperation to occur, it is certainly not sufficient: a theory-of-mind equally allows for enhanced competitive behavior [75]. In general, the boundary between "cooperative" and "non-cooperative" components of decision-making is too blurred to be of use for a measure of cooperation.

As a mechanism-agnostic alternative, we can base our measure on the contrast between an agent acting in accordance with its own goals, rather than the goals of the group. This is much simpler to evaluate for any given agent, as we need only consider what that agent's best response is to the behavior of the rest of the system, assuming the agent is only concerned with its own goals. This captures the individualized description of what truly occurs during chimpanzee group hunting [154].

One drawback of this approach is that it seems to rule out the idea that cooperation occurs ultimately as a result of agents acting in their own self-interest. In the context of evolutionary biology, much of the cooperation of biological systems is explained in terms of the direct fitness benefits that accrue to the cooperating agent. A contrastive measure of cooperation that used self-interest as a baseline would therefore appear to preclude any form of cooperation that has direct fitness as its basis as being true cooperation. To address this issue, we must go to the next category of desiderata.

Customizable A measure of cooperation should be *customizable* insofar as it allows variations to certain components that are key to determining how cooperative a given behavior is.

One such component is the *time scale* over which the behavior is evaluated. This is important for understanding the challenge posed by direct fitness explanations of cooperation: the self-interested benefits of direct fitness accrue only in the long term, whereas in the short term the behavior in question may seem counterintuitive from the self-interested perspective.

For example, a cleaning symbiosis exists between certain species of fish, in which a "cleaner" fish will enter into a "host" fish to consume the ectoparasites that live within it [155]. The host fish allows the cleaner fish to do this, even allowing the cleaner to exit when the cleaning is done. Though consuming the cleaner fish would provide an immediate benefit to the host, the host forgoes this in return for the opportunity for future cleaning, either by the same cleaner or another of its species. As a function of this time horizon, the

cooperativeness of this behavior would start as being positive before converging to zero. Punishments for norm violations would be an inverse to this case, starting negative and eventually going to zero or even a positive value under a potential measure.

Another important component of any evaluation of cooperative behavior is the way in which social outcomes are valued: while a typical choice would be to take the sum of all relevant agents' utilities, this does not always capture everything that we care about for a given outcome. For instance, we might instead evaluate success in terms of the utility of the worst-off agent, or in terms of the equitability of outcomes for each individual agent. The choice of metric for each social outcome will vary depending on the multi-agent system in question, and should not be held as a fixed component of the cooperative measure.

Contextual Finally, a measure should be *contextual*, so as to reflect the idea that the cooperativeness of an individual agent's actions depends on those of the other agents in the system they are interacting in. Hence, when measuring the cooperativeness of the agent, it will always be relative to the other agents in question. We can nonetheless derive the cooperativeness of the system as a whole by taking the average cooperativeness of each agent in the context of each other agent, though importantly this cooperativeness should fundamentally be a measure on *individual* behavior.

By making the measure contextual, we also make explicit the subgroup of agents on which the social outcomes are considered. While this subgroup may include all of the agents in the environment, this is not a requirement: in the example of predator-prey interactions, we do not consider the utility of the prey to be a factor in the cooperativeness of the group hunting behavior.

4.4 Measuring cooperation in stochastic games

We define our measure of cooperation within the framework of a *stochastic game* [137, 138]

$$\left(\mathcal{S}, N, \prod_{i=1}^{N} \mathcal{A}_i, P, \prod_{i=1}^{N} R_i, \gamma, \rho\right).$$

For an overview of this framework, refer to Sec 1.2.2 of the Introduction.

Assuming the agents follow policies $\boldsymbol{\pi} = (\pi_1, ..., \pi_N)$, we can write the value of a state s for agent i as:

$$V_{\boldsymbol{\pi}}^{(i)}(s) = \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{t=T}^{\infty} \gamma^{t-T} R_i(s_t, a_t) \;\middle|\; s_T = s\right]. \tag{4.1}$$

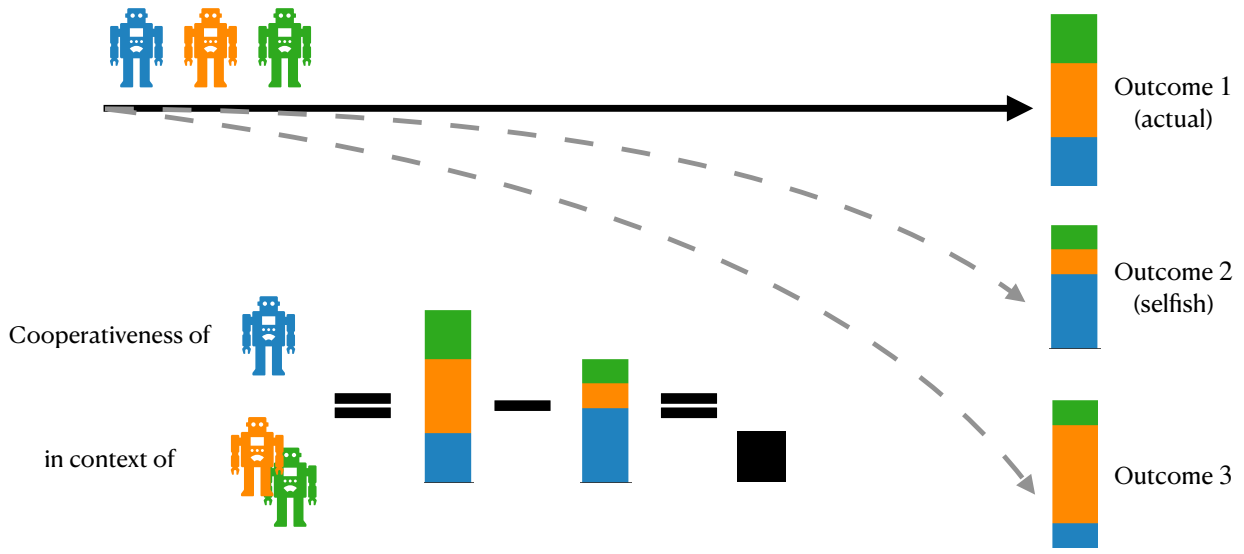


Figure 4.1: A schematic of the cooperation measure. A multiagent system consisting of three robots (Blue, Orange, and Green) has three outcomes, one actual and two potential, with utilities to each agent represented by stacked colored bars. The cooperativeness measure for Blue's policy consists of subtracting from the actual welfare (given here by total utility) the welfare for the selfish outcome, that is, the outcome for which Blue's utility is the largest.

To measure the social outcome of a stochastic game, we use a welfare metric w that is a function of each agent's value function, which we can then weight by the initial state distribution ρ. A typical choice is to use the *utilitarian* welfare as our metric, $w_U : \boldsymbol{\pi} \mapsto \sum_{s \in \mathcal{S}} \sum_{i=1}^{N} \rho(s) V_{\boldsymbol{\pi}}^{(i)}(s)$.

If we fix the policies of all agents except for i (denoting these as $\boldsymbol{\pi}_{-i}$), the stochastic game reduces to a single-agent Markov Decision Process (MDP). Let $\mathrm{BR}_i(\boldsymbol{\pi}_{-i})$ denote the (non-empty) set of optimal policies (or *best responses*) for agent i in the context of the other agents choosing policies $\boldsymbol{\pi}_{-i}$, i.e., the set of solutions to the single-agent MDP.

Finally, we define our measure of cooperation for a policy π_i in the context of $\boldsymbol{\pi}_{-i}$ as the welfare of these policies, minus the best possible welfare of agent i's best response policy:

$$c(\pi_i; \boldsymbol{\pi}_{-i}) = w(\pi_i, \boldsymbol{\pi}_{-i}) - \max_{\pi_i^* \in \mathrm{BR}_i(\boldsymbol{\pi}_{-i})} w(\pi_i^*, \boldsymbol{\pi}_{-i}). \tag{4.2}$$

A schematic diagram explaining this definition can be seen in Fig. 4.1.

By defining a scalar measure for cooperation, we are now able to evaluate the *degree* to which a policy is cooperative or uncooperative, and therefore we can also make comparative judgements between different

policies. Intuitively, the measure evaluates the extent to which policy π_i improves the social welfare over the (best) outcome that would have resulted from the agent acting purely in its self-interest.

This definition is clearly contextual, as the cooperativeness of π_i depends on the context of the other agents' policies $\boldsymbol{\pi}_{-i}$. Moreover, the definition clearly depends on the choice of welfare and discount factor. Each of these serves as an example of the measure's customizability, with the discount factor capturing the notion of a relevant time horizon..

The measure is also counterfactually contrastive in the sense set out above. In this case, we take the relevant counterfactual to hold fixed the policies of the other agents, and consider a self-interested agent to be one who maximizes its value in response to these policies. Since evaluating the counterfactual only requires us to find the best self-interested policy, it is agnostic to the internal mechanisms of the agent in question. In the environments we are interested in, we can compute optimal policies to arbitrary precision with value iteration [146], although for more complex systems we can also approximate the cooperation measure by using reinforcement learning to find approximate solutions to this problem.

4.5 Experiments in Social Dilemmas

To understand how our measure captures the intuitive notion of cooperative behavior, we focus on *social dilemmas* [35]. This class of games, studied in a wide variety of disciplines, involve interactions of agents with mixed motives in which agents acting in their own individual self-interest can effect an outcome that is worse than if all agents had cooperated. These dilemmas are therefore designed so as to clearly differentiate between cooperative and uncooperative behavior in a way that ought to be apparent in our measure.

We measure the cooperativeness of different kinds of behaviors that have been studied in these games. Moreover, by varying the context of other agents, as well as the time horizon, we show the impact that these parameters have on the cooperativeness of the agent behavior in question.

4.5.1 Matrix Game Social Dilemmas

To motivate the applicability of our measure, we begin by evaluating the cooperativeness of different strategies in four matrix games. The first three of these are the canonical *one-shot social dilemmas* that are designed to elucidate the opposing pressures of individual rationality and ideal collective action [85, 126, 89].

In these games, two agents have the choice of actions C (for *Cooperate*) or D (for *Defect*). The agents prefer mutual C to mutual D, mutual C to unilateral C, and mutual C yields a higher total utility than

Figure 4.2: The four varieties of matrix game social dilemmas: the Prisoner's Dilemma, Stag Hunt, Chicken, and Byproduct Mutualism. The top tow shows the payoff matrices for each game, with the colors representing the value of the total utility (calculated by adding the players' payoffs in each cell). The bottom row shows the heatmaps for the cooperativeness score for the row player's action in the context of the column player's action.

mutual D. However, in each game, we have that either unilateral D is preferable to mutual C (so that you can do better by exploiting a cooperator than cooperating with one), or that mutual D is preferable to unilateral C (so that being exploited is worse than not cooperating with a would-be exploiter). Chicken meets only the first of these disjuncts, Stag Hunt only the second, and Prisoner's Dilemma meets both.

The bottom row of Fig. 4.2 shows the cooperativeness of each row player's action in the context of the column player's action, using total utility as welfare. Holding fixed this context, we see that C is always strictly more cooperative than D, supporting the interpretation of C and D as *cooperation* and *defection*, respectively. Notably, in Byproduct Mutualism there is no pair of actions with a positive cooperativeness score. This is due to the fact that in this game the dilemma is completely relaxed: it is better to cooperate irrespective of the partner's decision, and so the choices that lead to the highest collective utility are also precisely the ones that self-interested actors would take.

4.5.2 Iterated Social Dilemmas

When we move to the iterated Prisoner's Dilemma, in which agents interact in a Prisoner's Dilemma *ad infinitum*, there is no strictly dominant individual strategy in this game.[1] Nonetheless, a number of strategies have been proposed with desirable properties [7, 111, 140]. We limit our strategies to those that depend on at most one previous interaction, referred to as *memory-1 strategies*: this includes strategies such as (Suspicious)-Tit-for-Tat (`(S)TFT`), Win-Stay-Lose-Shift (`PAVLOV`), and Grim (`GRIM`) (and their stochastic variants), but excludes others such as Tit-for-Two-Tats or Majority that require keeping track of a longer history. Hence, the MDP that arises from fixing the opponent's strategy to one of these will have five states (one for each possible action combination, and an additional initial state), making it tractable to solve so that we can compute the cooperativeness scores analytically.

Fig. 4.3 shows the cooperativeness measure applied to six deterministic memory-1 strategies, with each strategy being evaluated in the context of the other agent adopting every other strategy from the group. Strategies that take action C in more states generally score higher than strategies that take action D. However, we also see that the context policy plays an important role in determining the cooperativeness of the evaluated policy. In particular, in the context of policies that punish defection (either for one turn as in the case of `(S)TFT` or forever as in the case of `GRIM`), `ALL_C` does not rank as cooperative, as it becomes the best-response strategy. This supports the intuition that cooperating in the face of potential punishment is not as cooperative as unconditional cooperation, allowing us to distinguish between coercion and cooperation [133].

In addition to evaluating the cooperativeness of the most common deterministic memory-1 strategies, we ran a further analysis on all $2^5 = 32$ such strategies. Averaging over the possible deterministic other-agent contexts, we find that the policies that always end up cooperating (such as `ALL_C`) and the policies that always end up defecting (such as `ALL_D`) strategies attain the highest and lowest cooperativeness scores, on average ($+10.5$ and -9.5, respectively).

We also see the impact that the discount factor has on the measure of cooperativeness. If the column player adopts the `TFT` policy, then a row player will be able to exploit the fact that this strategy cooperates in the initial turn, at the expense of a defection in the subsequent turn. Therefore, if future rewards are sufficiently discounted relative to immediate rewards, it is optimal for the row player to initially defect. However, if future rewards are not significantly discounted, then it is in the row player's best interest to always cooperate. Fig. 4.4 shows the cooperativeness of each memory-1 strategy in the context of `TFT` plotted

[1] Refer to Appendix B.1 for analyses of the iterated Chicken and Stag Hunt.

against the discount factor: cooperative policies such as `ALL_C` score positively on cooperativeness for lower values of the discount factor, with the score eventually tending towards zero. On the other hand, defecting policies such as `ALL_D` have cooperativeness scores that begin at zero before tending towards $-\infty$.

4.5.3 Tabular Cleanup

Though iterated matrix games can lead to a richer range of behaviors through the use of memory-based strategies, the actions themselves that these strategies are defined over nonetheless treat *cooperate* and *defect* as primitives. A more faithful depiction of social dilemmas demands more complex strategies that apply to *policies* over richer action and state space. To this end, we investigate a simplified version of the social dilemma *Cleanup* [121, 66, 2, 59, 162]. This an example of a public goods dilemma, in which an individual must pay a personal cost in order to provide a resource that is shared by all [76].

The original version of *Cleanup* is an example of a *sequential social dilemma* [80]. These are typically formulated as being *partially observable*, so that each agent has only incomplete information on the state of the game (in this case, by limiting each agent's field of vision to a small subgrid of pixel values). Training agents to maximize rewards in such games typically requires example-inefficient deep reinforcement learning algorithms such as PPO [134, 174], in addition to policy models based on neural networks that must first learn to map pixel observations onto appropriate features. While our measure of cooperativeness is sufficiently general to capture such cases (any RL algorithm could find an "approximate best-response", giving a cooperativeness upper-bound), we believe that we can still develop insights about public goods dilemmas such as *Cleanup* without resorting to as much complexity.

To this end, we propose a simplified version we refer to as *Tabular Cleanup*: this game consists of N players who can choose between the actions *Clean*, *Eat*, and *Punish Player* i for $i = 1, ..., N$. The state space consists of the actions taken by each player at the previous time-step, and the number of apples currently available, which can range from 0 to $3N - 1$. An apple grows with a probability linearly proportional to the number of agents choosing *Clean*, with the probability ranging from 0 to 1. If an agent chooses to eat an apple, it receives a reward of $+1.0$, unless there are fewer apples available than agents eating, in which case the reward is divided amongst the eaters. If an agent chooses to punish another agent, it imposes a -2.0 reward deduction from the target, at an expense of -0.5 reward. For $N = 2$, we exclude the possibility of self-punishment for simplicity.

We consider two- and three-player instantiations of *Tabular Cleanup*, leading to state spaces of sizes 54 and 1125, respectively. This includes states which are not reachable from any other state: we therefore define

the distribution over initial states according to the states reached by starting with a uniformly random choice over actions and numbers of apples. We evaluate the following policies:

- *Always* X: This policy takes a constant action across all states. When there are three players, *Always Punish* punishes another player at random.
- *Take Turns*: This policy alternates between cleaning and eating.
- *TFT*: In the two-player version, this policy reciprocates the action taken by its co-player in the previous timestep. In the three-player version, *TFT* n will clean if n players are also cleaning, and will eat otherwise. Hence, TFT 2 is a more "suspicious" reciprocator than TFT 1 [2].
- *Nash*: This policy cleans when there are no apples available, and eats otherwise. As the name suggests, this is a Nash equilibrium to both versions of the game.
- *Prosocial*: This policy cleans when there are fewer apples available than the number of players, and eats otherwise. This was derived by solving the MDP derived from the two-player game with a centralized actor controlling both agents, and a reward consisting of the sum of the player's rewards. However, while this policy is maximally prosocial for the two-player game, it is not for the three-player game.[2]

Figures 4.5 and 4.6 show the results of evaluating each of these policies in a variety of contexts. In the two-player case, as expected, the *Prosocial* policy is the most cooperative on average across all contexts, and *Always Punish* the least. However, in many contexts and for both number of players, *Always Clean* is less cooperative than *Always Eat*, and *Take Turns* is more cooperative than both. This can be explained by the fact that eating contributes to the collective reward through adding to your *own* reward, and so choosing to clean in states where there are a sufficient number of apples for all agents needlessly forgoes a reward that contributes to the joint welfare. While such a result is intuitive, it is obscured by discussions of *Cleanup* that simply equate cooperativeness with the frequency at which each agent cleans [66].

These results can also be interpreted as providing a quantitative argument that *specialization* can be crucial to cooperation depending on the context. In the context of an agent that always eats, it is in fact more cooperative to focus on cleaning. However, in the converse context, eating becomes more imperative for increasing the joint welfare.

[2]We in fact find that no maximally prosocial policy exists for the three-player game that is the same for all three players.

4.6 Conclusion

Evaluating cooperativeness in multi-agent systems comprising of both people and AIs requires an understanding of cooperation that is *human-plausible*. In particular, our approach should accord with human-like judgments of behavior that rely on the posing of "what-if" scenarios involving the agent under evaluation acting in its own self-interest.

In this paper, we have motivated and specified a framework for measuring cooperative behavior that is contextual, customizable, and counterfactually contrastive. The cooperativeness measure is defined on a broad class of games and is agnostic to the mechanisms that drive cooperation, making it applicable to a variety of agent models. We then evaluated this measure in the space of deterministic policies in iterated social dilemmas, showing that the measure works in accordance with our intuitions and is capable of precluding examples of non-cooperative group behavior that contingently provide a group benefit. Finally, we expanded our evaluations to *Tabular Cleanup*: a stochastic game with a larger state space representing a temporally-extended social dilemma.

Future work will expand these evaluations to the sequential social dilemmas specified by the Melting Pot experimental suite [2], which introduce further complexity through temporally extended policies whose cooperative properties are not necessarily clear from taking a single action. These environments would also allow us to investigate the properties of the measure for environments involving more than two agents, and whether cooperation arises only within certain subgroups.

Furthermore, our measure could be used to detect collusion between a subgroup of agents by contrasting the agents' cooperativeness within the subgroup to the cooperativeness within the overall group [98].

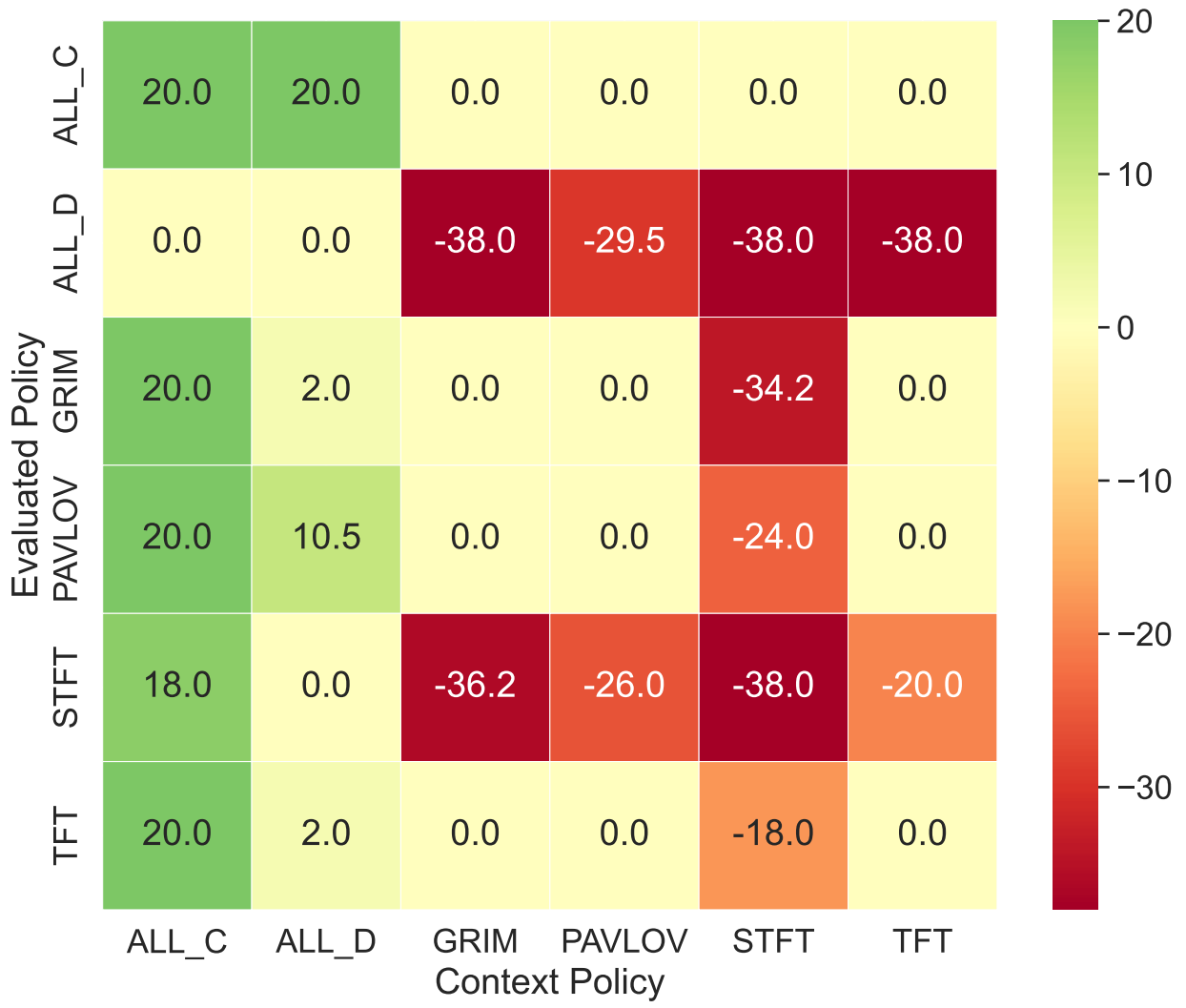


Figure 4.3: Cooperativeness of six common deterministic policies in the Iterated Prisoner's Dilemma, in the context of each other policy. The cooperativeness is valued on the initial state, with a discount factor $\gamma = 0.9$.

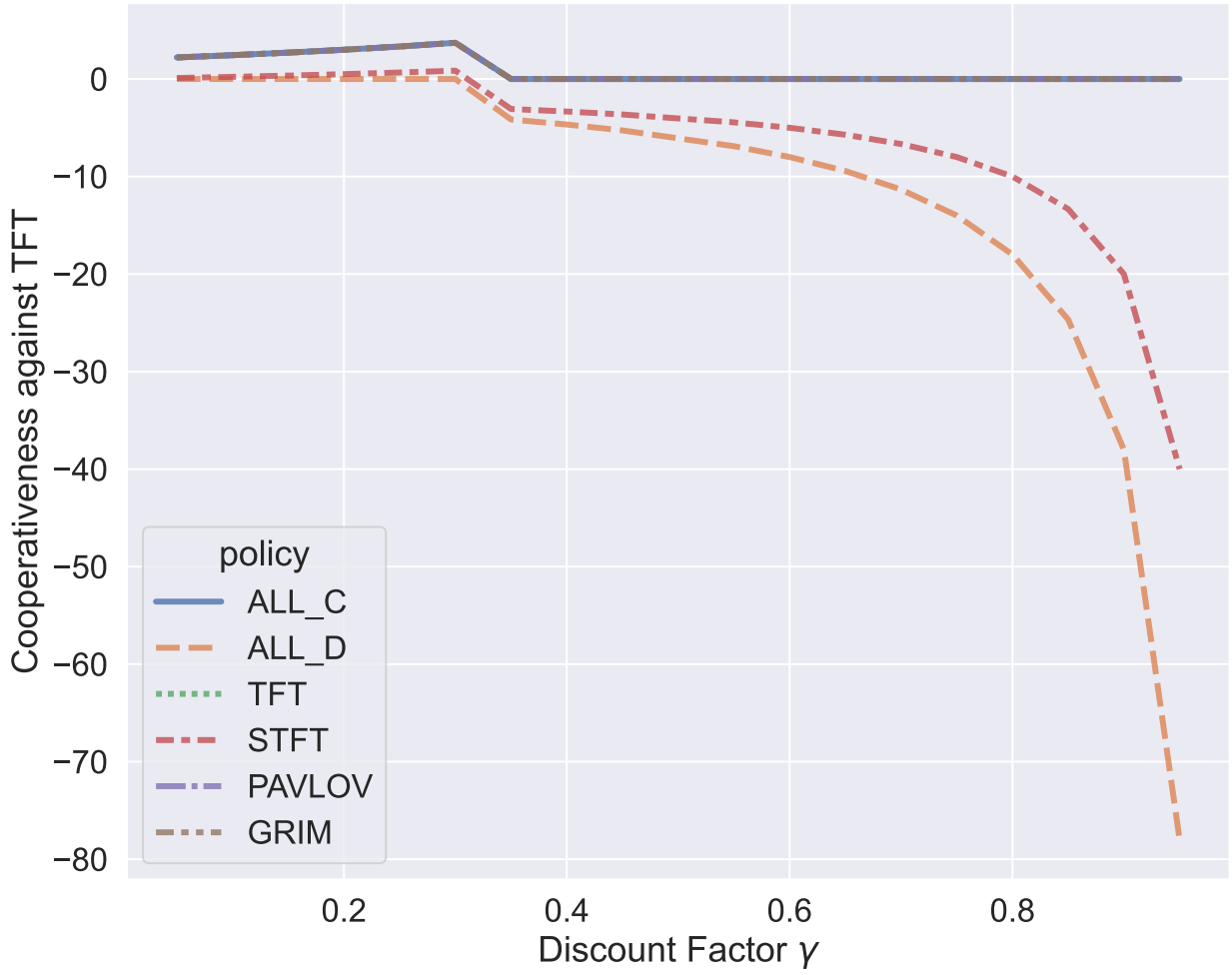


Figure 4.4: Cooperativeness of six common deterministic policies in the context of Tit-for-Tat in the Iterated Prisoner's Dilemma, plotted against the discount factor γ. As the `ALL_C`, `TFT`, `PAVLOV`, and `GRIM` strategies all cooperate on the initial time-step, their outcomes playing against `TFT` are identical and so their cooperativeness ratings overlap.

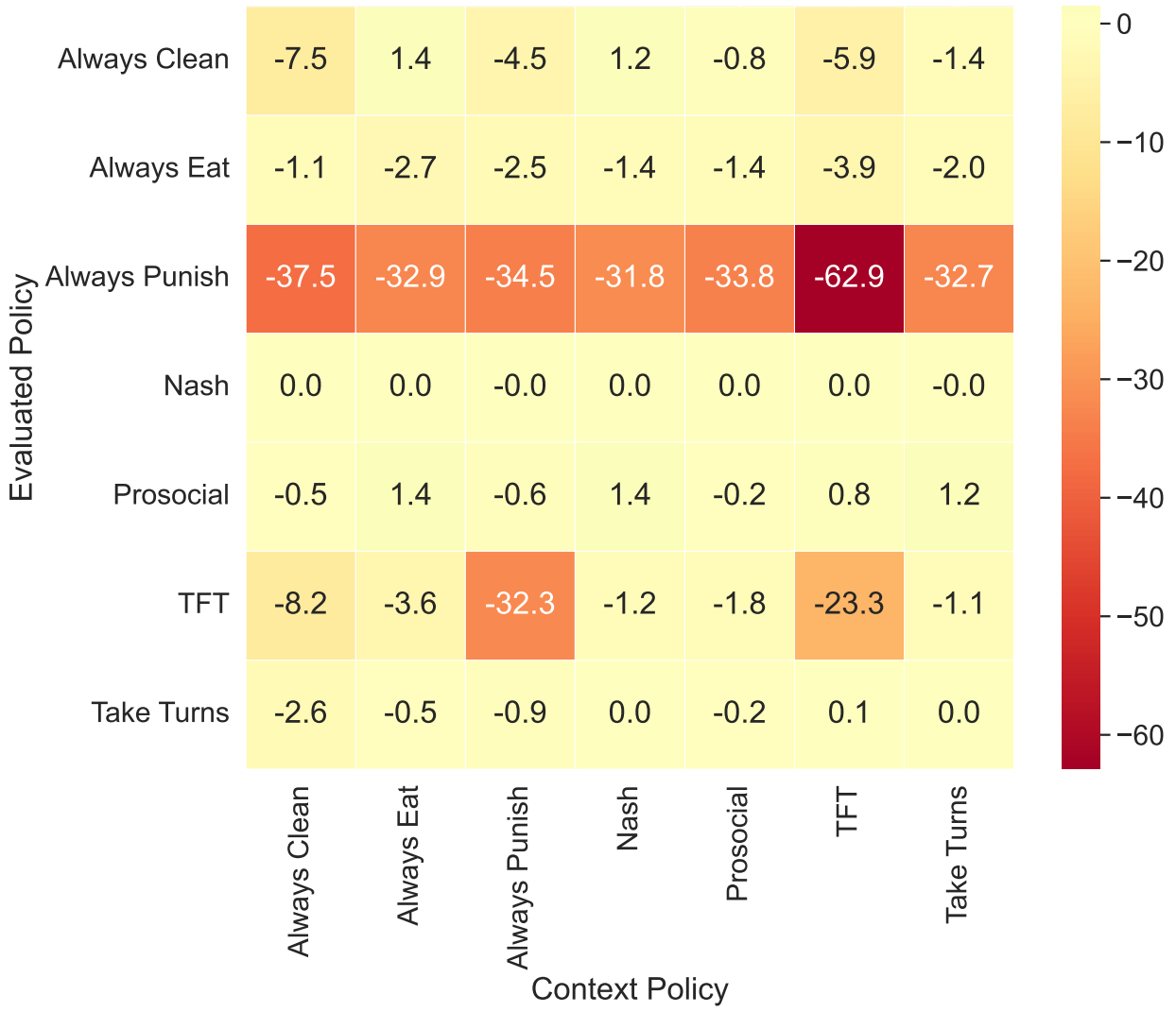


Figure 4.5: Cooperativeness of seven deterministic policies in the 2-player Tabular Cleanup, in the context of each other policy. The cooperativeness is valued on the initial state, with a discount factor $\gamma = 0.9$.

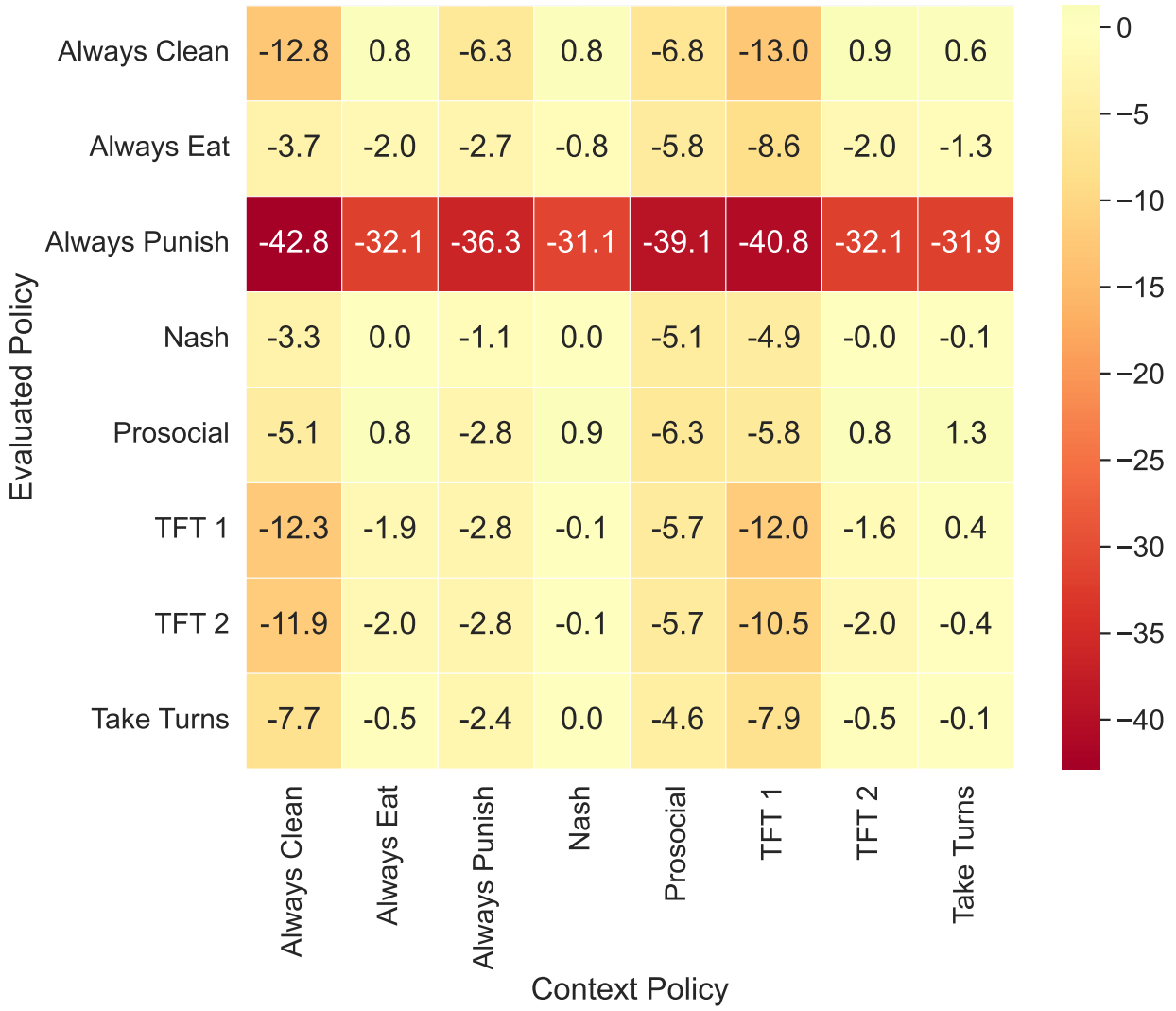


Figure 4.6: Cooperativeness of eight deterministic policies in the 3-player Tabular Cleanup, in the context of two players playing the same context policy. The cooperativeness is valued on the initial state, with a discount factor $\gamma = 0.9$.

Chapter 5

Conclusion

In this dissertation, we defined and proposed *human plausibility* as a unifying framework for building both single- and multi-agent models that can successfully interact with humans. Within this framework, we presented three projects motivated by this principle.

First, we showed how human-plausible representations and learning mechanisms, in the form of the successor representation and prioritized replay, can be used to create algorithms that exhibit human-like transfer and generalization behavior. This can be useful in creating predictive models of how people make split-second decisions when presented with unseen tasks, which can then be leveraged when designing systems that are deployed in time-critical applications.

We then showed how a recursive Bayesian model of speaker-listener interactions in which the speaker has persuasive goals can form a human-plausible account of inference from communicated evidence. This is essential for designing agents that must make choices about what information to present to a human and need to reason intelligently about how that information might be processed.

Finally, we proposed a tractable measure for the evaluation of cooperative behavior based on human-plausible intuitions about what constitutes such behavior in different contexts, where different counterfactuals might hold. This is an important methodological intervention in a field where the misevaluation or misdiagnosis of cooperative behavior may lead to disastrous consequences when an algorithm is deployed in the world at large.

By way of these projects, we hope to have demonstrated the fruitfulness of human plausibility as a concept for organizing and progressing our decision-making about how we can best design AIs that are to be used in everyday life. There is further work to be done in each of these projects, as outlined in the conclusions

of each chapter. Moreover, the general framework of human plausibility will naturally be enriched as we continue to develop our models of human intelligence. Beyond these avenues, the most interesting future projects in this direction might seek to understand further how humans interact with human-plausible AI, and what dynamics this may lead to in turn as the AI learns policies in response. This becomes increasingly relevant as the applications of AI systems expand into more areas of our lives. It also begs the question of how these interactions differ, depending on whether they occur in the context that the presence of AI is made explicit, or in the context of a "Turing test"-like scenario in which this is unknown.

Perhaps, as expansions in computational resources and theoretical models push us further along within this framework, we will build a clearer picture of the true nature of the gap between the behavior of humans and that of intelligent machines.

Bibliography

[1] Luigi Acerbi, Kalpana Dokka, Dora E Angelaki, and Wei Ji Ma. "Bayesian comparison of explicit and implicit causal inference strategies in multisensory heading perception". In: *PLOS Computational Biology* 14.7 (2018), e1006110.

[2] John P. Agapiou, Alexander Sasha Vezhnevets, Edgar A. Duéñez-Guzmán, Jayd Matyas, Yiran Mao, Peter Sunehag, Raphael Köster, Udari Madhushani, Kavya Kopparapu, Ramona Comanescu, DJ Strouse, Michael B. Johanson, Sukhdeep Singh, Julia Haas, Igor Mordatch, Dean Mobbs, and Joel Z. Leibo. *Melting Pot 2.0*. 2023. arXiv: `2211.13746 [cs.MA]`.

[3] Mayank Agrawal, Marcelo G Mattar, Jonathan D Cohen, and Nathaniel D Daw. "The temporal dynamics of opportunity costs: A normative account of cognitive fatigue and boredom." In: *Psychological review* 129.3 (2022), p. 564.

[4] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. *Concrete Problems in AI Safety*. 2016. arXiv: `1606.06565 [cs.AI]`.

[5] Norman H Anderson. *Foundations of information integration theory*. New York, NY: Academic Press, 1981.

[6] Kenneth J Arrow, Amartya Sen, and Kotaro Suzumura. *Handbook of Social Choice and Welfare*. Vol. 2. Elsevier, 2010.

[7] Robert Axelrod and William D Hamilton. "The evolution of cooperation". In: *Science* 211.4489 (1981), pp. 1390–1396.

[8] Maria Bagassi and Laura Macchi. "Pragmatic approach to decision making under uncertainty: The case of the disjunction effect". In: *Thinking & Reasoning* 12.3 (2006), pp. 329–350.

[9] Chris L Baker, Julian Jara-Ettinger, Rebecca Saxe, and Joshua B Tenenbaum. "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing". In: *Nature Human Behaviour* 1.4 (2017), pp. 1–10.

[10] Chris L Baker, Rebecca Saxe, and Joshua B Tenenbaum. "Action understanding as inverse planning". In: *Cognition* 113.3 (2009), pp. 329–349.

[11] Samuel Barnett and Ida Momennejad. "Priority-Adjusted Replay for Successor Representations". In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 44. 44. 2022.

[12] Samuel A Barnett, Thomas L Griffiths, and Robert D Hawkins. "A pragmatic account of the weak evidence effect". In: *Open Mind* 6 (2022), pp. 169–182.

[13] Samuel A Barnett and Ida Momennejad. "PARSR: Priority-Adjusted Replay for Successor Representations". In: *The 5th Multidisciplinary Conference on Reinforcement Learning and Decision Making*. 2022, pp. 80–85.

[14] Andre Barreto, Diana Borsa, John Quan, Tom Schaul, David Silver, Matteo Hessel, Daniel Mankowitz, Augustin Zidek, and Remi Munos. "Transfer in deep reinforcement learning using successor features and generalised policy improvement". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 501–510.

[15] André Barreto, Diana Borsa, Shaobo Hou, Gheorghe Comanici, Eser Aygün, Philippe Hamel, Daniel Toyama, Jonathan J. Hunt, Shibl Mourad, David Silver, and Doina Precup. "The Option Keyboard: Combining Skills in Reinforcement Learning". In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. Ed. by Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett. 2019, pp. 13031–13041. URL: https://proceedings.neurips.cc/paper/2019/hash/251c5ffd6b62cc21c446c963c76cf214-Abstract.html.

[16] André Barreto, Will Dabney, Rémi Munos, Jonathan J. Hunt, Tom Schaul, David Silver, and Hado van Hasselt. "Successor Features for Transfer in Reinforcement Learning". In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. Ed. by Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett. 2017, pp. 4055–4065.

URL: https://proceedings.neurips.cc/paper/2017/hash/350db081a661525235354dd3e19b8c05-Abstract.html.

[17] André Barreto, Shaobo Hou, Diana Borsa, David Silver, and Doina Precup. "Fast reinforcement learning with generalized policy updates". In: *Proceedings of the National Academy of Sciences* 117.48 (2020), pp. 30079–30087.

[18] Rahul Bhui and Samuel J Gershman. "Paradoxical effects of persuasive messages." In: *Decision* 7.4 (2020), pp. 239–258.

[19] Manuel Bohn, Michael Henry Tessler, Megan Merrick, and Michael C Frank. "How young children integrate information sources to infer the meaning of words". In: *Nature Human Behaviour* 5.8 (2021), pp. 1046–1054.

[20] Gerd Bohner, Markus Ruder, and Hans-Peter Erb. "When expertise backfires: Contrast and assimilation effects in persuasion". In: *British Journal of Social Psychology* 41.4 (2002), pp. 495–519.

[21] Elizabeth Bonawitz, Patrick Shafto, Hyowon Gweon, Noah D Goodman, Elizabeth Spelke, and Laura Schulz. "The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery". In: *Cognition* 120.3 (2011), pp. 322–330.

[22] Diana Borsa, André Barreto, John Quan, Daniel J. Mankowitz, Hado van Hasselt, Rémi Munos, David Silver, and Tom Schaul. "Universal Successor Features Approximators". In: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL: https://openreview.net/forum?id=S1VWjiRcKX.

[23] Michael E Bratman. "Shared cooperative activity". In: *The Philosophical Review* 101.2 (1992), pp. 327–341.

[24] Charles R Brown. "Social foraging in cliff swallows: local enhancement, risk sensitivity, competition and the avoidance of predators". In: *Animal Behaviour* 36.3 (1988), pp. 780–792.

[25] Iva K Brunec and Ida Momennejad. "Predictive representations in hippocampal and prefrontal hierarchies". In: *Journal of Neuroscience* 42.2 (2022), pp. 299–312.

[26] Micah Carroll, Rohin Shah, Mark K. Ho, Tom Griffiths, Sanjit A. Seshia, Pieter Abbeel, and Anca D. Dragan. "On the Utility of Learning about Humans for Human-AI Coordination". In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. Ed. by Hanna

M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett. 2019, pp. 5175–5186. URL: https://proceedings.neurips.cc/paper/2019/hash/f5b1b89d98b7286673128a5fb112cb9a-Abstract.html.

[27] Wilka Carvalho, Angelos Filos, Richard L. Lewis, Honglak Lee, and Satinder Singh. "Composing Task Knowledge With Modular Successor Feature Approximators". In: *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL: https://openreview.net/pdf?id=DrtSx1z40Ib.

[28] Wilka Carvalho, Andre Saraiva, Angelos Filos, Andrew Kyle Lampinen, Loic Matthey, Richard L. Lewis, Honglak Lee, Satinder Singh, Danilo J. Rezende, and Daniel Zoran. *Combining Behaviors with the Successor Features Keyboard*. 2023. arXiv: 2310.15940 [cs.AI].

[29] Robert B Cialdini. *Influence: The Psychology of Persuasion*. New York: Morrow, 1993.

[30] Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. *Towards Automated Circuit Discovery for Mechanistic Interpretability*. 2023. arXiv: 2304.14997 [cs.LG].

[31] Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson, and Thore Graepel. *Cooperative AI: machines must learn to find common ground*. 2021.

[32] Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R. McKee, Joel Z. Leibo, Kate Larson, and Thore Graepel. *Open Problems in Cooperative AI*. 2020. arXiv: 2012.08630 [cs.AI].

[33] Ishita Dasgupta, Eric Schulz, and Samuel J Gershman. "Where do hypotheses come from?" In: *Cognitive Psychology* 96 (2017), pp. 1–25.

[34] James H Davis. "Order in the courtroom". In: *Psychology and Law* (1984), pp. 251–265.

[35] Robyn M Dawes. "Social dilemmas." In: *Annual Review of Psychology* (1980).

[36] Peter Dayan. "Improving generalization for temporal difference learning: The successor representation". In: *Neural Computation* 5.4 (1993), pp. 613–624.

[37] Kenneth G DeBono and Richard J Harnish. "Source expertise, source attractiveness, and the processing of persuasive information: A functional approach." In: *Journal of Personality and Social Psychology* 55.4 (1988), pp. 541–546.

[38] Yali Du, Joel Z. Leibo, Usman Islam, Richard Willis, and Peter Sunehag. *A Review of Cooperation in Multi-agent Learning*. 2023. arXiv: 2312.05162 [cs.MA].

[39] Edgar A Duéñez-Guzmán, Suzanne Sadedin, Jane X Wang, Kevin R McKee, and Joel Z Leibo. "A social path to human-like artificial intelligence". In: *Nature Machine Intelligence* 5.11 (2023), pp. 1181–1188.

[40] Emily Falk and Christin Scholz. "Persuasion, Influence, and Value: Perspectives from Communication and Social Neuroscience". In: *Annual Review of Psychology* 69.1 (2018), pp. 329–356. ISSN: 0066-4308, 1545-2085.

[41] Philip M. Fernbach, Adam Darlow, and Steven A. Sloman. "When good evidence goes bad: The weak evidence effect in judgment and decision-making". In: *Cognition* 119.3 (2011), pp. 459–467. ISSN: 00100277.

[42] Jakob N Foerster. "Deep Multi-Agent Reinforcement Learning". PhD thesis. University of Oxford, 2018.

[43] Michael Franke and Judith Degen. "Reasoning in reference games: Individual-vs. population-level probabilistic modeling". In: *PLOS ONE* 11.5 (2016), e0154854.

[44] Michael Franke, Giulio Dulcinati, and Nausicaa Pouscoulous. "Strategies of Deception: Under-Informativity, Uninformativity, and Lies—Misleading With Different Kinds of Implicature". In: *Topics in Cognitive Science* 12.2 (2020), pp. 583–607.

[45] Michael Franke and Gerhard Jäger. "Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics". In: *Zeitschrift für Sprachwissenschaft* 35.1 (2016), pp. 3–44.

[46] Mona M Garvert, Raymond J Dolan, and Timothy EJ Behrens. "A map of abstract relational knowledge in the human hippocampal–entorhinal cortex". In: *elife* 6 (2017), e17086.

[47] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC Press, 2013.

[48] Samuel J Gershman. "The successor representation: its computational logic and neural substrates". In: *Journal of Neuroscience* 38.33 (2018), pp. 7193–7200.

[49] Noah D Goodman and Michael C Frank. "Pragmatic language interpretation as probabilistic inference". In: *Trends in Cognitive Sciences* 20.11 (2016), pp. 818–829.

[50] Noah D Goodman and Andreas Stuhlmüller. "Knowledge and implicature: Modeling language understanding as social cognition". In: *Topics in Cognitive Science* 5.1 (2013), pp. 173–184.

[51] Noah D Goodman and Andreas Stuhlmüller. *The Design and Implementation of Probabilistic Programming Languages*. `http://dippl.org`. Accessed: 2020-1-7. 2014.

[52] H. Paul Grice. "Logic and Conversation". In: *Syntax and semantics, Speech acts*. Ed. by Peter Cole and Jerry Morgan. Vol. 3. New York: Academic Press, 1975.

[53] Hyowon Gweon, Hannah Pelton, Jaclyn A Konopka, and Laura E Schulz. "Sins of omission: Children selectively explore when teachers are under-informative". In: *Cognition* 132.3 (2014), pp. 335–341.

[54] William Donald Hamilton and William Donald Hamilton. *Narrow roads of gene land: evolution of social behaviour*. Vol. 1. Oxford University Press on Demand, 1996.

[55] Adam Harris, Adam Corner, and Ulrike Hahn. "James is polite and punctual (and useless): A Bayesian formalisation of faint praise". In: *Thinking & Reasoning* 19.3 (2013), pp. 414–429. ISSN: 1354-6783, 1464-0708.

[56] Paul Harris, Melissa A Koenig, Kathleen H Corriveau, and Vikram K Jaswal. "Cognitive foundations of learning from testimony". In: *Annual Review of Psychology* 69 (2018), pp. 251–273.

[57] Daniel Hawthorne-Madell and Noah D Goodman. "Reasoning about social sources to learn from actions and outcomes." In: *Decision* 6.1 (2019), pp. 17–60.

[58] Joseph Henrich. *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press, 2015.

[59] Uri Hertz, Raphael Koster, Marco Janssen, and Joel Z Leibo. "Beyond the Matrix: Experimental Approaches to Studying Social-Ecological Systems". In: (2023).

[60] Matteo Hessel, Joseph Modayil, Hado Van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. "Rainbow: Combining improvements in deep reinforcement learning". In: *Thirty-second AAAI conference on artificial intelligence*. 2018.

[61] Sophie Hilgard, Nir Rosenfeld, Mahzarin R Banaji, Jack Cao, and David Parkes. "Learning Representations by Humans, for Humans". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Proceedings of Machine Learning Research. 2021, pp. 4227–4238.

[62] Robin M Hogarth and Hillel J Einhorn. "Order effects in belief updating: The belief-adjustment model". In: *Cognitive Psychology* 24.1 (1992), pp. 1–55.

[63] Carl Iver Hovland, Irving Lester Janis, and Harold H Kelley. *Communication and persuasion.* Yale University Press, 1953.

[64] Anne Hsu, Andy Horng, Thomas L Griffiths, and Nick Chater. "When absence of evidence is evidence of absence: Rational inferences from absent data". In: *Cognitive Science* 41 (2017), pp. 1155–1167.

[65] Anne S. Hsu and Thomas L. Griffiths. "Differential Use of Implicit Negative Evidence in Generative and Discriminative Language Learning". In: *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada.* Ed. by Yoshua Bengio, Dale Schuurmans, John D. Lafferty, Christopher K. I. Williams, and Aron Culotta. Curran Associates, Inc., 2009, pp. 754–762. URL: https://proceedings.neurips.cc/paper/2009/hash/0f96613235062963ccde717b18f97592-Abstract.html.

[66] Edward Hughes, Joel Z. Leibo, Matthew Phillips, Karl Tuyls, Edgar A. Duéñez-Guzmán, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin R. McKee, Raphael Koster, Heather Roff, and Thore Graepel. "Inequity aversion improves cooperation in intertemporal social dilemmas". In: *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada.* Ed. by Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett. 2018, pp. 3330–3340. URL: https://proceedings.neurips.cc/paper/2018/hash/7fea637fd6d02b8f0adf6f7dc36aed93-Abstract.html.

[67] Geoffrey Irving, Paul Christiano, and Dario Amodei. *AI safety via debate.* 2018. arXiv: 1805.00899 [stat.ML].

[68] Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, DJ Strouse, Joel Z Leibo, and Nando De Freitas. "Social influence as intrinsic motivation for multi-agent deep reinforcement learning". In: *International Conference on Machine Learning.* PMLR. 2019, pp. 3040–3049.

[69] Julian Jara-Ettinger, Hyowon Gweon, Laura E Schulz, and Joshua B Tenenbaum. "The naıve utility calculus: Computational principles underlying commonsense psychology". In: *Trends in Cognitive Sciences* 20.8 (2016), pp. 589–604.

[70] Minqi Jiang, Edward Grefenstette, and Tim Rocktäschel. *Prioritized Level Replay.* 2021. arXiv: 2010.03934 [cs.LG].

[71] Arthur Juliani, Samuel Barnett, Brandon Davis, Margaret Sereno, and Ida Momennejad. "Neuro-Nav: A Library for Neurally-Plausible Reinforcement Learning". In: *The 5th Multidisciplinary Conference on Reinforcement Learning and Decision Making*. 2022, pp. 290–294.

[72] Michael J Kahana, Marc W Howard, and Sean M Polyn. "Associative retrieval processes in episodic memory". In: *Psychology* 3 (2008).

[73] Phillipp Kampshoff and Asutosh Padhi. *Autonomous-driving disruption: Technology, use cases, and opportunities*. `https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/autonomous-driving-disruption-technology-use-cases-and-opportunities`. 2017. (Visited on 02/19/2024).

[74] Steven Kapturowski, Georg Ostrovski, John Quan, Remi Munos, and Will Dabney. "Recurrent experience replay in distributed reinforcement learning". In: *International Conference on Learning Representations*. 2018.

[75] Max Kleiman-Weiner, Mark K Ho, Joseph L Austerweil, Michael L Littman, and Joshua B Tenenbaum. "Coordinate to cooperate or compete: abstract goals and joint intentions in social interaction". In: *CogSci*. 2016.

[76] Peter Kollock. "Social dilemmas: The anatomy of cooperation". In: *Annual Review of Sociology* 24.1 (1998), pp. 183–214.

[77] Elias Koutsoupias and Christos Papadimitriou. "Worst-case equilibria". In: *Computer Science Review* 3.2 (2009), pp. 65–69.

[78] Kropotkin Peter Kropotkin. *Mutual aid: A factor of evolution*. Black Rose Books Ltd., 2021.

[79] Tejas D. Kulkarni, Ardavan Saeedi, Simanta Gautam, and Samuel J. Gershman. *Deep Successor Reinforcement Learning*. 2016. arXiv: `1606.02396 [stat.ML]`.

[80] Joel Z Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. "Multi-agent Reinforcement Learning in Sequential Social Dilemmas". In: *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*. 2017, pp. 464–473.

[81] Adam Lerer and Alexander Peysakhovich. *Maintaining cooperation in complex social dilemmas using deep reinforcement learning*. 2018. arXiv: `1707.01068 [cs.AI]`.

[82] Falk Lieder and Thomas L Griffiths. "Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources". In: *Behavioral and Brain Sciences* 43 (2020), e1.

[83] Long-Ji Lin. "Self-improving reactive agents based on reinforcement learning, planning and teaching". In: *Machine Learning* 8.3 (1992), pp. 293–321.

[84] Lola L Lopes. "Procedural debiasing". In: *Acta Psychologica* 64.2 (1987), pp. 167–185.

[85] R Duncan Luce and Howard Raiffa. *Games and decisions: Introduction and critical survey*. Courier Corporation, 1989.

[86] Fengling Ma, Dan Zeng, Fen Xu, Brian J Compton, and Gail D Heyman. "Delay of gratification as reputation management". In: *Psychological Science* 31.9 (2020), pp. 1174–1182.

[87] Marlos C Machado, Andre Barreto, Doina Precup, and Michael Bowling. "Temporal abstraction in reinforcement learning with the successor representation". In: *Journal of Machine Learning Research* 24.80 (2023), pp. 1–69.

[88] Marlos C. Machado, Clemens Rosenbaum, Xiaoxiao Guo, Miao Liu, Gerald Tesauro, and Murray Campbell. "Eigenoption Discovery through the Deep Successor Representation". In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL: https://openreview.net/forum?id=Bk8ZcAxR-.

[89] Michael W Macy and Andreas Flache. "Learning dynamics in social dilemmas". In: *Proceedings of the National Academy of Sciences* 99.suppl_3 (2002), pp. 7229–7236.

[90] Lorenzo Magnani. *Abduction, Reason and Science: Processes of Discovery and Explanation*. Springer Science & Business Media, 2011.

[91] Yiran Mao, Madeline G. Reinecke, Markus Kunesch, Edgar A. Duéñez-Guzmán, Ramona Comanescu, Julia Haas, and Joel Z. Leibo. *Doing the right thing for the right reason: Evaluating artificial moral cognition by probing cost insensitivity*. 2023. arXiv: 2305.18269 [cs.AI].

[92] Antoine Marot, Benjamin Donnot, Karim Chaouache, Adrian Kelly, Qiuhua Huang, Ramij-Raja Hossain, and Jochen L Cremer. "Learning to run a power network with trust". In: *Electric Power Systems Research* 212 (2022), p. 108487.

[93] Antoine Marot, Benjamin Donnot, Gabriel Dulac-Arnold, Adrian Kelly, Aidan O'Sullivan, Jan Viebahn, Mariette Awad, Isabelle Guyon, Patrick Panciatici, and Camilo Romero. "Learning to run a power network challenge: a retrospective analysis". In: *NeurIPS 2020 Competition and Demonstration Track*. PMLR. 2021, pp. 112–132.

[94] Antoine Marot, Benjamin Donnot, Camilo Romero, Balthazar Donon, Marvin Lerousseau, Luca Veyrin-Forrer, and Isabelle Guyon. "Learning to run a power network challenge for training topology controllers". In: *Electric Power Systems Research* 189 (2020), p. 106635.

[95] David Marr. *Vision*. San Francisco, CA: W. H. Freeman, 1982.

[96] Kristy A Martire, Richard I Kemp, M Sayle, and Ben R Newell. "On the interpretation of likelihood ratios in forensic science evidence: Presentation formats and the weak evidence effect". In: *Forensic Science International* 240 (2014), pp. 61–68.

[97] Marcelo G Mattar and Nathaniel D Daw. "Prioritized memory access explains planning and hippocampal replay". In: *Nature Neuroscience* 21.11 (2018), pp. 1609–1617.

[98] Parisa Mazrooei, Christopher Archibald, and Michael Bowling. "Automating collusion detection in sequential games". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 27. 1. 2013, pp. 675–682.

[99] Craig R M McKenzie, Susanna M Lee, and Karen K Chen. "When Negative Evidence Increases Confidence: Change in Belief After Hearing Two Sides of a Dispute". In: *Journal of Behavioral Decision Making* 15.1 (2002), pp. 1–18.

[100] Craig RM McKenzie and Jonathan D Nelson. "What a speaker's choice of frame reveals: Reference points, frame selection, and framing effects". In: *Psychonomic Bulletin & Review* 10.3 (2003), pp. 596–602.

[101] Jörg Meibauer. *The Oxford handbook of lying*. Oxford University Press, 2019.

[102] Candice M Mills and Asheley R Landrum. "Learning who knows what: Children adjust their inquiry to gather information from others". In: *Frontiers in Psychology* 7 (2016), p. 951.

[103] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. *Playing Atari with Deep Reinforcement Learning*. 2013. arXiv: `1312.5602 [cs.LG]`.

[104] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT Press, 2018.

[105] Ida Momennejad. "A rubric for human-like agents and NeuroAI". In: *Philosophical Transactions of the Royal Society B* 378.1869 (2023), p. 20210446.

[106] Ida Momennejad. "Learning structures: Predictive representations, replay, and generalization". In: *Current Opinion in Behavioral Sciences* 32 (2020), pp. 155–166.

[107] Ida Momennejad, A Ross Otto, Nathaniel D Daw, and Kenneth A Norman. "Offline replay supports planning in human reinforcement learning". In: *Elife* 7 (2018), e32548.

[108] Ida Momennejad, Evan M Russek, Jin H Cheong, Matthew M Botvinick, Nathaniel Douglass Daw, and Samuel J Gershman. "The successor representation in human reinforcement learning". In: *Nature Human Behaviour* 1.9 (2017), pp. 680–692.

[109] Andrew W Moore and Christopher G Atkeson. "Prioritized sweeping: Reinforcement learning with less data and less time". In: *Machine Learning* 13.1 (1993), pp. 103–130.

[110] Giuseppe Mosconi and Laura Macchi. "The role of pragmatic rules in the conjunction fallacy". In: *Mind & Society* 2.1 (2001), pp. 31–57.

[111] Martin Nowak and Karl Sigmund. "A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game". In: *Nature* 364.6432 (1993), pp. 56–58.

[112] Daniel J O'Keefe. *Persuasion: Theory and research*. Sage Publications, 2015.

[113] Lauren A Oey, Adena Schachner, and Edward Vul. "Designing good deception: Recursive theory of mind in lying and lie detection". In: *Proceedings of the 41st Annual Conference of the Cognitive Science Society*. 2019, pp. 897–903.

[114] Lauren A Oey and Ed Vul. "Lies are crafted to the audience". In: *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*. 2021, pp. 791–797.

[115] Hee Sun Park, Timothy R Levine, Catherine Y Kingsley Westerman, Tierney Orfgen, and Sarah Foregger. "The effects of argument quality and involvement type on attitude formation and attitude change: A test of dual-process and social judgment predictions". In: *Human Communication Research* 33.1 (2007), pp. 81–102.

[116] Cédric Paternotte. "Minimal cooperation". In: *Philosophy of the Social Sciences* 44.1 (2014), pp. 45–73.

[117] Judea Pearl and Dana Mackenzie. *The Book of Why: the New Science of Cause and Effect*. New York: Basic Books, 2018.

[118] Jorge Peña and Georg Nöldeke. "Cooperative dilemmas with binary actions and multiple players". In: *Dynamic Games and Applications* 13.4 (2023), pp. 1156–1193.

[119] Jing Peng and Ronald G Williams. "Efficient learning and planning within the Dyna framework". In: *Adaptive Behavior* 1.4 (1993), pp. 437–454.

[120] Andrew Perfors, Danielle J Navarro, and Patrick Shafto. "Stronger evidence isn't always better: A role for social inference in evidence selection and interpretation." In: *Proceedings of the 40th Annual Conference of the Cognitive Science Society*. 2018, pp. 864–869.

[121] Julien Pérolat, Joel Z. Leibo, Vinícius Flores Zambaldi, Charles Beattie, Karl Tuyls, and Thore Graepel. "A multi-agent reinforcement learning model of common-pool resource appropriation". In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. Ed. by Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett. 2017, pp. 3643–3652. URL: https://proceedings.neurips.cc/paper/2017/hash/2b0f658cbffd284984fb11d90254081f-Abstract.html.

[122] Richard E Petty. *Attitudes and persuasion: Classic and contemporary approaches*. Routledge, 2018.

[123] Guy Politzer and Laura Macchi. "Reasoning and pragmatics". In: *Mind & Society* 1.1 (2000), pp. 73–93.

[124] Diane Poulin-Dubois and Patricia Brosseau-Liard. "The developmental origins of selective social learning". In: *Current Directions in Psychological Science* 25.1 (2016), pp. 60–64.

[125] Keith Ransom, Wouter Voorspoels, Andrew Perfors, and Daniel Navarro. "A cognitive analysis of deception without lying". In: *Proceedings of the 39th Annual Conference of the Cognitive Science Society*. 2017, pp. 992–997.

[126] Anatol Rapoport, Albert M Chammah, and Carol J Orwant. *Prisoner's dilemma: A study in conflict and cooperation*. Vol. 165. University of Michigan press, 1965.

[127] Evan M Russek, Ida Momennejad, Matthew M Botvinick, Samuel J Gershman, and Nathaniel D Daw. "Predictive representations can link model-based reinforcement learning to model-free mechanisms". In: *PLoS Computational Biology* 13.9 (2017), e1005768.

[128] Stuart Russell. *Human Compatible: Artificial intelligence and the problem of control*. Penguin, 2019.

[129] Arleen Salles, Kathinka Evers, and Michele Farisco. "Anthropomorphism in AI". In: *AJOB neuroscience* 11.2 (2020), pp. 88–95.

[130] George Santayana. *Life of Reason*. Prometheus Books, 2013.

[131] Jennifer Mather Saul. *Lying, misleading, and what is said: An exploration in philosophy of language and in ethics*. Oxford University Press, 2012.

[132] Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. "Prioritized Experience Replay". In: *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2016. URL: http://arxiv.org/abs/1511.05952.

[133] Thomas C Schelling. *The Strategy of Conflict: with a new Preface by the Author*. Harvard University Press, 1980.

[134] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. *Proximal Policy Optimization Algorithms*. 2017. arXiv: 1707.06347 [cs.LG].

[135] G Scontras, Michael Henry Tessler, and M Franke. *Probabilistic language understanding: An introduction to the Rational Speech Act framework*. Accessed: 2020-1-7. 2018.

[136] Patrick Shafto, Noah D. Goodman, and Thomas L. Griffiths. "A rational account of pedagogical reasoning: Teaching by, and learning from, examples". In: *Cognitive Psychology* 71 (2014), pp. 55–89. ISSN: 00100285.

[137] Lloyd S Shapley. "Stochastic games". In: *Proceedings of the National Academy of Sciences* 39.10 (1953), pp. 1095–1100.

[138] Yoav Shoham and Kevin Leyton-Brown. *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. Cambridge University Press, 2008.

[139] Les Sikos, Noortje J Venhuizen, Heiner Drenhaus, and Matthew W Crocker. "Speak before you listen: Pragmatic reasoning in multi-trial language games". In: *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*. 2021.

[140] Tyler Singer-Clark. "Morality Metrics On Iterated Prisoners Dilemma Players". In: (2014).

[141] David M Sobel and Tamar Kushnir. "Knowledge matters: How children evaluate the reliability of testimony as a process of rational inference." In: *Psychological Review* 120.4 (2013), pp. 779–797.

[142] Dan Sperber, Francesco Cara, and Vittorio Girotto. "Relevance theory explains the selection task". In: *Cognition* 57.1 (1995), pp. 31–95.

[143] Kimberly L Stachenfeld, Matthew M Botvinick, and Samuel J Gershman. "The hippocampus as a predictive map". In: *Nature Neuroscience* 20.11 (2017), pp. 1643–1653.

[144] Emma Strubell, Ananya Ganesh, and Andrew McCallum. *Energy and Policy Considerations for Deep Learning in NLP*. 2019. arXiv: 1906.02243 [cs.CL].

[145] Richard S Sutton. "Dyna, an integrated architecture for learning, planning, and reacting". In: *ACM Sigart Bulletin* 2.4 (1991), pp. 160–163.

[146] Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

[147] Ming Tan. "Multi-agent reinforcement learning: Independent vs. cooperative agents". In: *Proceedings of the Tenth International Conference on Machine Learning*. 1993, pp. 330–337.

[148] Ning Tang, Siyi Gong, Minglu Zhao, Chenya Gu, Jifan Zhou, Mowei Shen, and Tao Gao. "Exploring an imagined "we" in human collective hunting: Joint commitment within shared intentionality". In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 44. 44. 2022.

[149] Ning Tang, Stephanie Stacy, Minglu Zhao, Gabriel Marquez, and Tao Gao. "Bootstrapping an Imagined We for Cooperation." In: *CogSci*. 2020.

[150] Joshua B Tenenbaum and Thomas L Griffiths. "Generalization, similarity, and Bayesian inference". In: *Behavioral and Brain Sciences* 24.4 (2001), pp. 629–640.

[151] Joshua B. Tenenbaum. "Bayesian Modeling of Human Concept Learning". In: *Advances in Neural Information Processing Systems 11, [NIPS Conference, Denver, Colorado, USA, November 30 - December 5, 1998]*. Ed. by Michael J. Kearns, Sara A. Solla, and David A. Cohn. The MIT Press, 1998, pp. 59–68. URL: http://papers.nips.cc/paper/1542-bayesian-modeling-of-human-concept-learning.

[152] Edward C Tolman. "Cognitive maps in rats and men". In: *Psychological Review* 55.4 (1948), pp. 189–208.

[153] Michael Tomasello. *The cultural origins of human cognition*. Cambridge, MA: Harvard University Press, 2009.

[154] Michael Tomasello. *Why we cooperate*. MIT press, 2009.

[155] Robert L Trivers. "The evolution of reciprocal altruism". In: *The Quarterly Review of Biology* 46.1 (1971), pp. 35–57.

[156] Jennifer S Trueblood and Jerome R Busemeyer. "A quantum probability account of order effects in inference". In: *Cognitive Science* 35.8 (2011), pp. 1518–1552.

[157] Raimo Tuomela. "What is cooperation?" In: *Erkenntnis* (1993), pp. 87–101.

[158] Harm Van Seijen and Rich Sutton. "Planning by prioritized sweeping with small backups". In: *International Conference on Machine Learning*. PMLR. 2013, pp. 361–369.

[159] Aki Vehtari, Andrew Gelman, and Jonah Gabry. "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC". In: *Statistics and computing* 27.5 (2017), pp. 1413–1432.

[160] Natalia Vélez and Hyowon Gweon. "Integrating incomplete information with imperfect advice". In: *Topics in Cognitive Science* 11.2 (2019), pp. 299–315.

[161] Leander Vignero. "Updating on Biased Probabilistic Testimony". In: *Erkenntnis* (2022), pp. 1–24.

[162] Eugene Vinitsky, Raphael Köster, John P Agapiou, Edgar A Duéñez-Guzmán, Alexander S Vezhnevets, and Joel Z Leibo. "A learning agent that acquires social norms from public sanctions in decentralized multi-agent settings". In: *Collective Intelligence* 2.2 (2023).

[163] Sumio Watanabe. "A widely applicable Bayesian information criterion". In: *Journal of Machine Learning Research* 14 (2013), pp. 867–897.

[164] Christopher JCH Watkins and Peter Dayan. "Q-learning". In: *Machine Learning* 8.3-4 (1992), pp. 279–292.

[165] Joseph Weizenbaum. *Computer Power and Human Reason: From Judgment to Calculation*. W. H. Freeman and Company, 1976.

[166] Stuart A West, Ashleigh S Griffin, and Andy Gardner. "Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection". In: *Journal of Evolutionary Biology* 20.2 (2007), pp. 415–432.

[167] Andrew Whalen, Thomas L Griffiths, and Daphna Buchsbaum. "Sensitivity to shared information in social learning". In: *Cognitive Science* 42.1 (2017), pp. 168–187. (Visited on 02/05/2018).

[168] Richard Willis, Yali Du, Joel Z Leibo, and Michael Luck. *Resolving social dilemmas with minimal reward transfer*. 2023. arXiv: `2310.12928 [cs.GT]`.

[169] Lennart Wittkuhn, Samson Chien, Sam Hall-McMaster, and Nicolas W Schuck. "Replay in minds and machines". In: *Neuroscience & Biobehavioral Review* 129 (2021), pp. 367–388.

[170] Lara A Wood, Rachel L Kendal, and Emma G Flynn. "Whom do children copy? Model-based biases in social learning". In: *Developmental Review* 33.4 (2013), pp. 341–356.

[171] Rabbi Dr Shmuly Yanklowitz. *Pirkei Avot: A Social Justice Commentary*. CCAR Press, 2018.

[172] Erica J Yoon, Kyle MacDonald, Mika Asaba, Hyowon Gweon, and Michael C Frank. "Balancing informational and social goals in active learning." In: *Proceedings of the 40th Annual Conference of the Cognitive Science Society*. 2018, pp. 1218–1223.

[173] Erica J Yoon, Michael Henry Tessler, Noah D Goodman, and Michael C Frank. "Polite speech emerges from competing social goals". In: *Open Mind* 4 (2020), pp. 71–87.

[174] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre M. Bayen, and Yi Wu. "The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games". In: *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*. Ed. by Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh. 2022. URL: http://papers.nips.cc/paper%5C_files/paper/2022/hash/9c1535a02f0ce079433344e14d910597-Abstract-Datasets%5C_and%5C_Benchmarks.html.

[175] Minglu Zhao, Ning Tang, Annya L Dahmani, Yixin Zhu, Federico Rossano, and Tao Gao. "Sharing Rewards Undermines Coordinated Hunting". In: *Journal of Computational Biology* (2022).

Appendix A

Appendix to Chapter 3

A.1 Exclusions and attention checks

Our pre-registered exclusion criteria used two basic attention checks. First, participants were required to complete a comprehension quiz immediately following the task instructions, and we excluded participants who failed to successfully complete this quiz within three attempts. Second, at the end of the experiment, we asked participants to use a slider to indicate the degree of bias they believed each contestant exhibited. These motivations were stated explicitly in the instructions (e.g. "the red contestant will receive $10 if the judge chooses "shorter," otherwise the blue contestant will receive $10") so, although participants may differ in the *degree* to which they thought such incentives would bias the contestants away from neutrality, we took responses in the *opposite* direction of the incentive as indicative of inattentiveness or misunderstanding of task instructions.

We therefore coded bias check responses as "incorrect" if the slider response was inconsistent with the bias given in the instructions (e.g. if the short-biased contestant received a slider rating above the midpoint, $s \geq 50 - \epsilon$, or the long-biased contestant received a slider rating below the midpoint, $s \leq 50 + \epsilon$ where we set $\epsilon = 5$ to allow for the possibility of motor jitter from participants who intended to use the exact midpoint.) In our pre-registered second sample (reported in the main text), 793 participants completed instructions and 723 (91%) of them passed the attention check.

While these pre-registered criteria were designed to ensure that apparent differences in speaker and listener behavior were not simply driven by general attentional factors, it is possible that participants who did not expect the strongest evidence to be shown in the speaker phase (238 participants, or 33%) were still

group	n	both 2AFC and point estimate consistent	generative model also consistent
strongest first	485	0.97	0.89
not strongest first	238	0.96	0.86

Table A.1: Stricter attention check passage rates broken out by speaker group.

systematically less attentive than other participants. To address this concern, we analyzed a series of other measures to assess the degree of attention and task understanding across "speaker expectation" groups. Specifically, we examine internal consistency within several post-test questions, where we asked participants (i) to make a final two-alternative forced choice verdict about whether the sample of sticks is 'longer' vs. 'shorter' than 5 inches, (ii) to provide a point estimate of their best guess of the actual mean on a slider ranging from 1 inch to 9 inches, and (iii) to guess the values of the remaining three sticks that were not revealed, allowing us to impute a "generative" average across the two observed values and the three guessed values (Table A.1).

We say a participant passed the 2AFC check if their binary verdict ('longer' vs. 'shorter') is consistent with the direction of their point estimate. We say a participant also passed the stricter "generative" check if the average imputed from their guesses for the remaining three unobserved sticks matches their 2AFC and point estimates. We observe that rates for the these stricter checks were somewhat lower for participants who expected speakers not to show the strongest evidence first (97% vs. 96%, and 89% vs. 86%, respectively), though neither of these differences was significant, $\chi^2(1) = 0.76, p = 0.38$ and $\chi^2(1) = 1.05, p = 0.31$, respectively. Rates were far above chance for all groups. To ensure robustness, we re-ran our primary analyses on the subset of participants that passed the strictest conjunction of all checks, which is highly improbable under an inattentive null model, and obtained nearly identical results (most crucially, a significant interaction, $t(718) = 5.18, p < 0.001$).

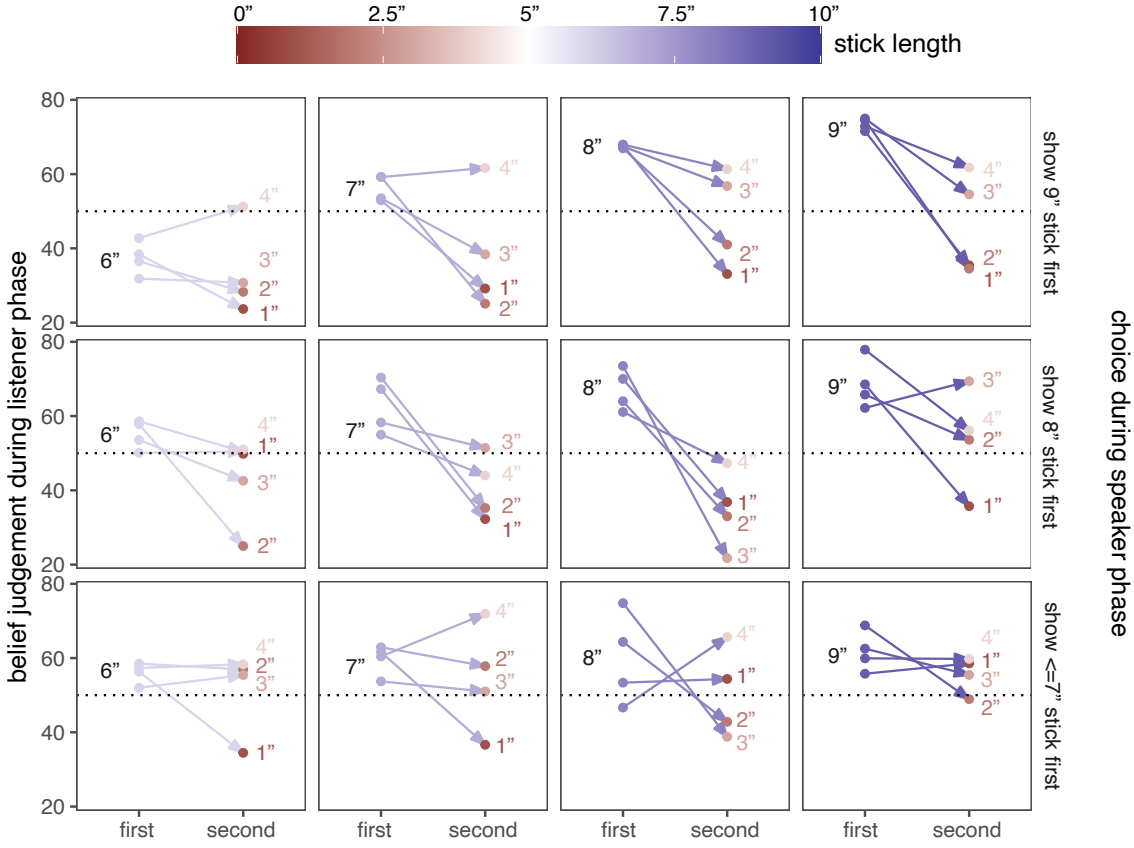


Figure A.1: Participants revised their beliefs after obtaining a second piece of evidence. Each facet represents participants who were given the same initial piece of evidence (blue dots) with each arrow connecting their judgment after the first piece of evidence and the second piece of evidence. In most cases, participants revised their estimates down, although participants who showed a weak evidence effect for the first stick (top column) also displayed a classical weak evidence effect on the second piece of evidence (e.g. in the second row, participants who saw a 7” stick on the first trial were slightly *more* confident the average was longer after seeing a 4” stick).

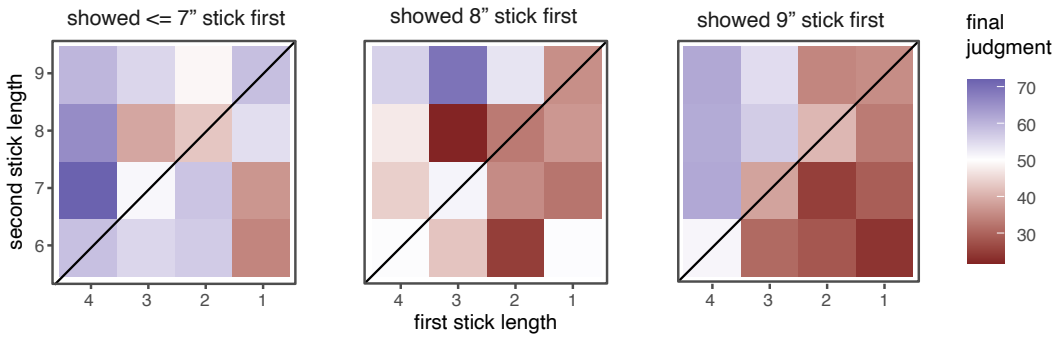


Figure A.2: We found strong order effects, with the belief judgment elicited after the second stick apparently affected by a recency bias. Under perfect averaging, the diagonal would leave the judge with complete uncertainty (denoted on our color scale by white), since the evidence from both the longer side (blue) and the shorter side (red) should cancel out.

A.2 Order effects

While we focus on the first piece of evidence as the clearest weak evidence effect, we also collected a second response after a second piece of evidence was shown by the other speaker. These responses are visualized in Fig. A.1. As expected, we observed a recency effect (more easily observed in the diagonal of Fig. A.2, where evidence from the "short"-biased and "long"-biased speakers were equally strong), where participants weighted the second piece of evidence more strongly.

A.3 Proofs

Theorem 1. *The speaker model using the combined utility Eq. 3.6 simplifies to Eq. 3.8 for the stick contest task.*

Proof. We begin by substituting the combined utility (Eq. 3.6) into the speaker softmax:

$$\begin{aligned} S(u|w,w^*) &\propto \exp\{\alpha \cdot U(u;w,w^*)\} \\ &= \exp\{\alpha \cdot [U_{\text{epi}}(u;w) + \beta \cdot U_{\text{pers}}(u;w^*)]\} \\ &= \exp\{\alpha \cdot U_{\text{epi}}(u;w)\} \cdot \exp\{\alpha \cdot \beta \cdot U_{\text{pers}}(u;w^*)\} \end{aligned}$$

Now, using Eq. 3.3 to expand the first term, note that

$$U_{\text{epi}}(u;w) = \ln P_{L_0}(w|u) = \ln \frac{P(w)\delta_{[\![u]\!](w)}}{\sum_w P(w)\delta_{[\![u]\!](w)}} = \begin{cases} -\ln N & \text{if } [\![u]\!](w) \\ -\infty & \text{o.w.} \end{cases}$$

where N is the number of sticks in the true set ($N = 5$ in our experiment). However, we already assume that the set of possible utterances $\mathcal{U}$ are the true sticks in the underlying set (i.e. the contestants cannot make up sticks, they must choose one of the N sticks in the set), so

$$\begin{aligned} \exp\{\alpha \cdot U_{\text{epi}}(u,w)\} &= \begin{cases} \alpha/N & \text{if } [\![u]\!](w) \\ 0 & \text{o.w.} \end{cases} \\ &= \alpha/N \end{aligned}$$

Because all utterances have the exact same epistemic utility U_{epi}, this term drops out of the soft-max:

$$\begin{aligned} S(u|w,w^*) &\propto \exp\{\alpha \cdot U_{\text{epi}}(u;w)\} \cdot \exp\{\alpha \cdot \beta \cdot U_{\text{pers}}(u;w^*)\} \\ &\propto \exp\{\alpha \cdot \beta \cdot U_{\text{pers}}(u;w^*)\} \\ &= \exp\{\alpha \cdot \beta \cdot \ln L_0(w^*|u)\} \end{aligned}$$

yielding Eq. 3.8. □

Theorem 2. *Persuasiveness monotonically increases as a function of stick length.*

Proof. We say an utterance u is more persuasive than an utterance u' when

$$U_{\text{pers}}(u \mid w^*) > U_{\text{pers}}(u' \mid w^*).$$

Under the stick contest, let $\mathcal{L} = \{l_1, \ldots, l_N\}$ be an partially-ordered set of N stick lengths, such that $l_i \leq l_j$ for any index $i < j$. We denote the mean stick length by $\bar{l} = \frac{1}{N}\sum_i l_i$. Without loss of generality, let the speaker's persuasive goal be $w^* =$ `shorter` $= \bar{l} < 5$ (the argument follows analogously for `longer`). Take two utterances $u = l_i$ and $u' = l_j$ such that $l_i \leq l_j$ (i.e. such that u is just as short or shorter than u'). First,

we expand the utility:

$$\begin{aligned} U_{\text{pers}}(u \mid \texttt{shorter}) &= \ln L_0(\texttt{shorter} \mid u) \\ &= \ln P(\bar{l} < 5 \mid l_i) \\ &= \ln P\left(\frac{l_i + \sum l_{-i}}{N} < 5\right) \\ &= \ln P\left(\sum l_{-i} < 5N - l_i\right) \end{aligned}$$

Now, let X be a random variable representing the sum of the $N-1$ still-unknown sticks, $X = \sum l_{-i}$. Then we recognize this as the cumulative distribution function (CDF), $F_X(x) = P(X < x)$. Because the underlying set of sticks $\mathcal{L}$ is assumed to be i.i.d., note that the random variable $X = \sum l_{-i}$ does not depend on the original choice of i. Critically, we know that the cumulative distribution function is monotonic increasing in x, i.e. $F_X(a) \leq F_X(b)$ for $a \leq b$. Hence if $l_i \leq l_j$ then $5N - l_i \geq 5N - l_j$ and $F_X(5N - l_i) \geq F_X(5N - l_j)$:

$$\begin{aligned} U(u \mid \texttt{shorter}) &= \ln P\left(\sum l_{-i} < 5N - l_i\right) \\ &= \ln F_X(5N - l_i) \\ &\geq \ln F_X(5N - l_j) \\ &= U(u' \mid \texttt{shorter}) \end{aligned}$$

□

A.4 Results from original sample

The results reported in the main text are based on a pre-registered replication we conducted during the revision of the manuscript (May 2022). In this appendix, we report the corresponding results from our original sample (February 2020). The only methodological difference between the original study and the internal replication was the way we counter-balanced the order of the "long"- vs. "short"-biased contestants. In our original study, the "long"-biased contestant always presented their evidence first; in our replication, the order of the contestants was randomized. Additionally, in our replication, we added the following clarification to the instructions: "Sticks ranging in length from 1 to 9 inches are equally likely to appear in the set." Participants in the initial sample were recruited on the Prolific platform, with no restriction on country. Of

the 784 participants who successfully completed the instructions, 708 passed the second attention check.

Our regression model was the same as in the main text, except we did not include a fixed effect of "long" vs. "short": all participants were shown evidence from the "long"-biased speaker. As in the study reported in the main text, we found a significant interaction between speaker expectations and evidence strength on beliefs about the underlying mean, $t(704) = 5.9, p < 0.001$. For participants who expected the speaker to provide the strongest evidence (421 participants or 60% of our sample), the weak evidence provided by a six inch stick backfired, leading them to instead expect that the mean stick length was significantly less likely to be longer than five inches, $m = 37.5$, 95% CI: $[33.1, 41.9]$, $t(98) = -5.7$, $p < 0.001$. Meanwhile, for participants who expected to be shown the second-longest stick (40% of the sample), no weak evidence effect was found, with the 'longest stick' group significantly different from the other groups, $t(167) = -5.5, p < 0.001$.

A.5 Model fitting details

A.5.1 RSA model

We used the following priors for our Bayesian data analysis:

$$
\begin{aligned}
y &\sim \text{Gaussian}(\mu + o, 0.3) \\
p_z &\sim \text{Unif}[0, 1] \\
\beta &\sim \text{Unif}[0, 10] \\
o &\sim \text{Unif}[-0.5, 0.5]
\end{aligned}
$$

where p_z is the mixture weight used for heterogeneous models, $\mu = P_{L_i}(\texttt{longer}|u) \in [0, 1]$ is the RSA listener model's posterior belief, and o is a uniform offset included to allow for systematic response biases in use of the slider. Intuitively, Gaussian$(\mu + o, 0.3)$ can be viewed as a simple way of scoring the error between the model prediction $\mu + o$ and the participant's response y. For the speaker-dependent model, we used independent priors depending on the participant's choice of stick j: $p_z^{(j)} \sim \text{Unif}[0, 1]$. Because there were relatively fewer participants who expected the `longer` speaker to choose 0.2 or 0.4 (sticks that were in the opposite direction of their goal; and vice versa for the `shorter` speaker), we collapsed these participants together, forming three groups: those who expected the strongest evidence to be presented first (e.g. who selected $\{0.2, 0.9\}$ for the *short* and *long* biased speakers, respectively), those who expected the second-strongest to be presented first

(e.g. who selected $\{0.4, 0.8\}$, respectively), and those who expected less strong evidence. However, our findings are robust to whether we collapse these groups or not.

A.5.2 Belief-adjustment models

In the notation of McKenzie, Lee, and Chen [99], Eq. 3.9 is written:

$$C_k = C_{k-1} + w_k \cdot (s(e_k) - R), \tag{A.1}$$

where $C_k \in [0, 1]$ is the degree of belief in a particular claim after being presented with evidence e_k, $s(e_k)$ is the *independently judged* strength of evidence e_k, R is a reference point, and $w_k \in [0, 1]$ is an adjustment weight for evidence e_k. In the *adding* variant of the belief-adjustment model, Hogarth and Einhorn [62] argue that the evidence should be encoded in an absolute manner, letting $R = 0$ and $s(e_k) \in [-1, 1]$, and assuming that if $s(e_k) \leq R$ then $w_k = C_{k-1}$, otherwise $w_k = 1 - C_{k-1}$.[1] To allow the reference point for evidence to be more demanding than neutrality, McKenzie, Lee, and Chen [99] proposed replacing the reference point R with a Minimum Acceptable Strength (MAS) threshold $(m \mid e)$, that depends on the evidence previously presented. We can therefore rewrite Eq. A.1 as

$$C_k = C_{k-1} + w_k \cdot (s(e_k) - (m_k \mid e_1, ..., e_{k-1})). \tag{A.2}$$

To fit this class of models to our data, we follow Trueblood and Busemeyer [156], assuming a mapping between stick length and evidence strength given by a centered logistic function:

$$\text{strength}(u) = \frac{1}{1 + \exp\left(-B \cdot (u - 5)\right)} - 0.5, \tag{A.3}$$

where the logistic growth rate B is fit to the data (we used a uniform prior $B \sim \text{Unif}[0, 10]$). This function satisfies several desiderata: it is monotonically increasing in the size of the stick, it is bounded in the interval $[-1, 1]$, and it is centered in line with the prior over stick lengths, so that a stick of length 5 inches has a strength of 0.5.

For the anchor-and-adjust (AA) variant, we fix the reference point as $R = 0$, and for the minimum

[1] The *averaging* variant, in which evidence is encoded in relationship to the current belief in the hypothesis, is more suited for *estimation* tasks involving some kind of moving average [62], whereas the Stick Contest is better described as an *evaluation* task in which a single hypothesis is under consideration ("is the sample long?"). We also found empirically that the adding variant provided a better fit to the data than the averaging variant.

acceptable strength (MAS) variant, we infer a reference point with prior $R \sim \text{Unif}[-1, 1]$. We consider *homogeneous* variants in which the entire population is assumed to share the same model with the same parameters, as well as a *heterogeneous* model, in which we assume *a priori* that participants are a convex combination of the two models. As in the RSA models, we infer the mixture weight p_z that best explains the population-level mixture (marginalizing over latent variable assignments z).

A.5.3 Higher levels of reasoning and the strong evidence effect

While our cover story explicitly provided participants with the motivations of speakers, in terms of their financial incentives, these motivations are less obvious in most real-world scenarios. They must be *inferred* from what the speaker is saying. This is straightforwardly derived in our framework by allowing the listener to jointly infer the true state of the world w *and* the speaker's bias β:

$$P_{L_1}(w, \beta \mid u) \propto P_{S_1}(u \mid w, \beta) \cdot P(w) \tag{A.4}$$

Our formulation raises a natural question about how speakers would behave if they were *aware* judges were making such inferences. This emerges at the next level of recursive reasoning:

$$P_{S_2}(u \mid w, \beta) \propto \exp\Big(|\beta| \ln\big(P_{L_1}(w^* \mid u) - w_c \cdot C(u)\big)\Big). \tag{A.5}$$

where $C(u)$ represents some cost associated with being perceived as biased by the judge:

$$C(u) = \mathbb{E}_{\beta \sim P_{L_1}(\cdot \mid u)}\big[|\beta|\big], \tag{A.6}$$

and $w_c \geq 0$ is a parameter specifying the degree of the cost. We included a L_2 model who reasons about this listener in our model comparison (i.e. allowing participants to be explained by a convex combination of all three levels) and found that this three-level speaker-dependent model leads to improved performance over the two-level speaker-dependent model (max likelihood = 16.2, WAIC= -18.3 ± 8.9, PSIS-LOO= -9.2 ± 8.9.) We conjecture that this formulation is required to account for the *strong evidence effect* [120], in which the desire to appear unbiased leads a speaker to choose weaker evidence in spite of the presence of stronger alternatives, but leave further investigation for future work.

A.6 Transcript of the Experiment

The written instructions for our experiment are reproduced below. Note that the task can be seen exactly as participants experienced it (e.g. with images) using the code released in our repository: `https://github.com/s-a-barnett/bayesian-persuasion`.

> In this task, you will serve as the judge for a heated game between these two contestants. The two contestants in this game have been given a set of sticks ranging in length from very long ones to very short ones. Sticks ranging in length from 1 to 9 inches are equally likely to appear in the set. One contestant (shown in pink) will be rewarded handsomely if they can convince you that the average length of these sticks is shorter than 5in (see dotted line). The other (shown in blue) will get paid if if they can convince you that the average length of these sticks is longer than 5in (see dotted line). In this case, the average length is 6in, so the position that this person was arguing for was true. As the judge, however, you will not be able to see the full set of sticks: you will only see what the contestants choose to show you. They will each get to show exactly one of the five sticks to convince you. After you see each stick, you will use this slider to report how strongly you are leaning in your decision. If you think the stick average is more likely to be shorter than 5in, click further to the left. If you think it is more likely to be longer than 5in, click further to the right.

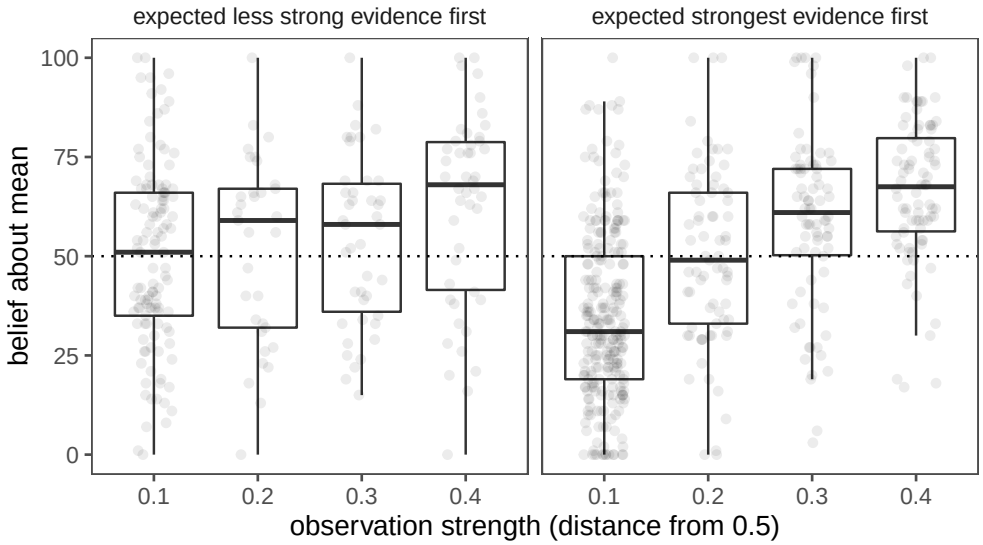


Figure A.3: The raw data distribution of responses for the listener phase, where each individual (jittered) point is a different participant and the boxplot represents the median (dark line) and first and third quartiles (top and bottom of box) of the response distribution.

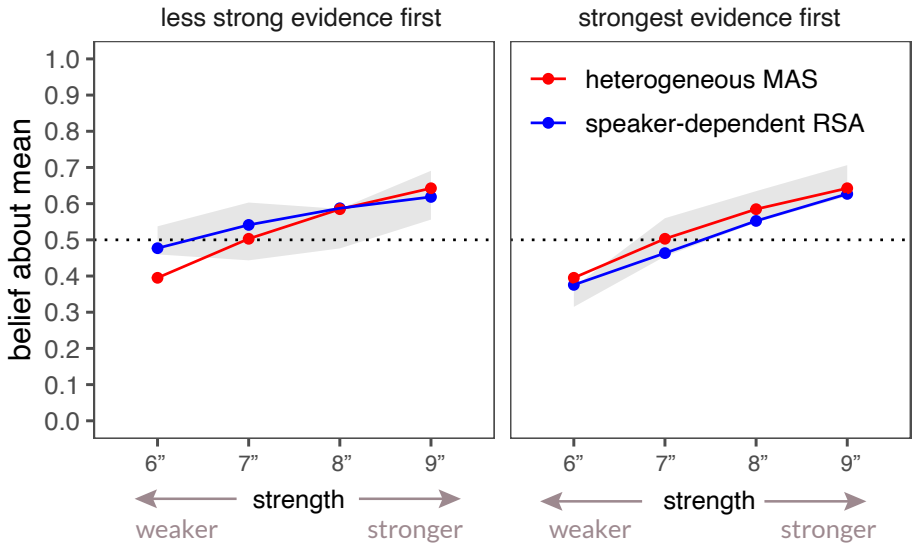


Figure A.4: We visualized the posterior predictives for the speaker-dependent RSA model (blue) and heterogeneous MAS model (red). The facets represent which stick was expected to be chosen first in the speaker phase, and the grey region represents the 95% confidence interval of the empirical data.

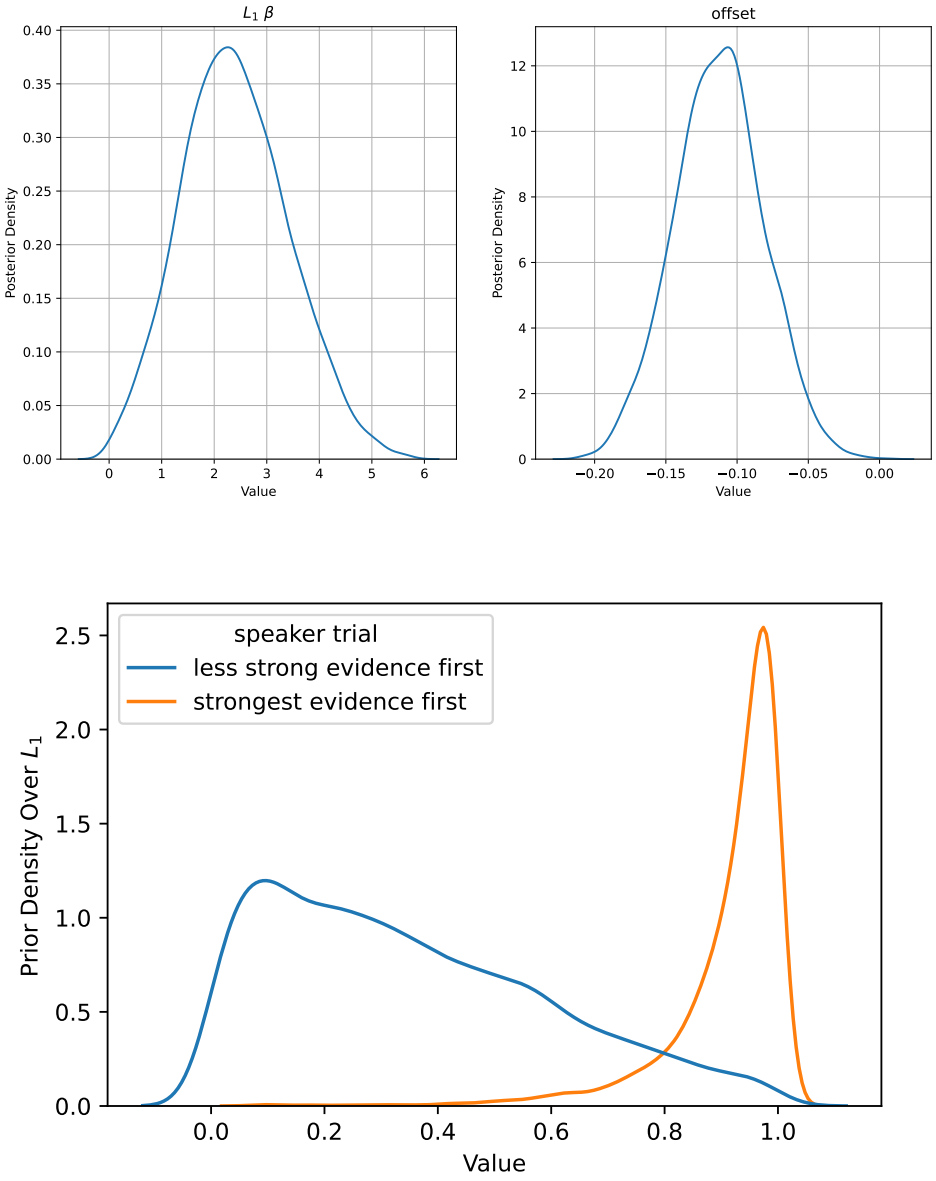


Figure A.5: Full Bayesian posteriors for the parameters of the speaker-dependent RSA model. In the top panel, the MAP parameter values are found to be $(\beta^*, o^*) = (2.26, -0.11)$. The bottom panel shows the posteriors over mixture weights p_z for the different speaker groups. The MAP parameter values for are $p_z = 0.10$ for the less strong evidence group and $p_z = 0.97$ for the strongest evidence group.

group	**What was your strategy for selecting sticks as a speaker?**
strongest evidence	If I need them to believe more than 5 inch i'd choose the biggest and opposite for below 5 inch — *either choose the longest if I am blue, or the shortest if I am red* — I picked the longest or shortest stick based on what I wanted the judge to believe — *Trying to show the extremes for each argument so the judge thinks the average is more likely to be closer to those* — I think it's best to show the longest/shortest stick you own - to make it appear that they're all very long/short — *i guess it was to show extremes of the sizes of sticks i had, show the smallest on or the tallest one* — my strategy was to create the illusion that the average lenght is bigger in the case I am the blue contestant by showing the longest sticks only, and the same with the red one only showing the shortest. — *Pick the shortest or longest one to bump up or reduce the avearge*
weaker evidence	Selected slightly towards where the first stick suggested — *I actually want to avoid the highest or lowest if I can at first to give the impression that you yourself have picked a more "average" stick.* — Not going too far either way, but just enough to seem less obvious. — *To show a slightly longer or shorter length than the average to try persuade the judge otherwise.* — show some variation to gain trust — *try to keep them guessing*

Table A.2: Participants were presented with a free-response text field to explain their reasoning at the end of both phases. Here we provide sample responses from the end of the *speaker* phase, from both participants who expected the *strongest* evidence and those who expected less strong evidence.

group	**How did you reach your decision as a judge?**
strongest evidence	6 is not very much over the average that their trying to prove - which makes me think that al the other sticks are even shorter than that." — *I was thinking that the pink player would choose the shortest stick, whilst the blue would choose the longest* — if 4cm was the shortest stick available then the maximum number of sticks above 5cm would be 4 — *blue showed me a very long stick meaning there would have to be an opposite short stick to average it out. pink however did not show a very short stick suggesting there aren't any.* — the blue would have shown a longer one if it was there — *I assume that blue would be likely to pick the longest possible stick as they have an incentive to make me think the average is above 5in; if they only present a 6in stick, it is likely that the average is under 5in.*
weaker evidence	tried to do a average — *Felt the pink player was bluffing* — thge average of the 2 sticks was shorter than 5 — *Looking at the average of the values I'd been given so far* — seemed similar to how I played it so assumed there were more long ones to come like in my strategy — *the contestant is likely to trick you*

Table A.3: Sample responses from the end of the *judge* phase, from both participants who expected the *strongest* evidence and those who expected less strong evidence.

Appendix B

Appendix to Chapter 4

B.1 Full results across multiple welfare functions

We evaluate the cooperativeness measure on each game using three different welfare metrics. Let $V_1, V_2, ..., V_N$ denote the values for N agents in the environment. These metrics are then defined as:

- Total Value: $\sum_{i=1}^{N} V_i$.
- Minimum Value: $\min_{i=1,2,...,N} V_i$.
- Equality: $1 - \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} |V_i - V_j|}{2N \sum_{i=1}^{N} V_i}$.

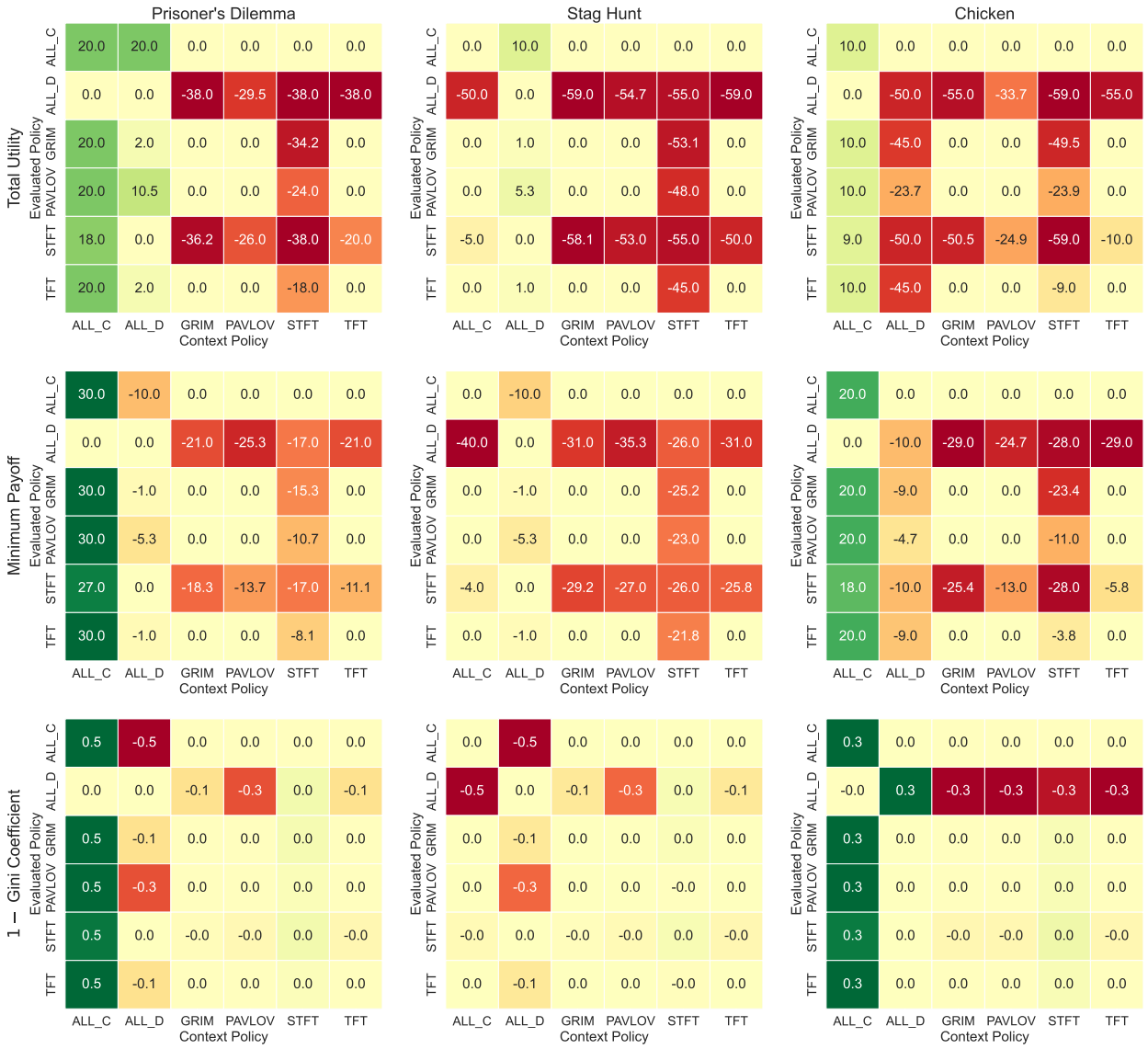


Figure B.1: Cooperativeness of six common deterministic policies in the Iterated Prisoner's Dilemma, in the context of each other policy. The cooperativeness is valued on the initial state, with a discount factor $\gamma = 0.9$.

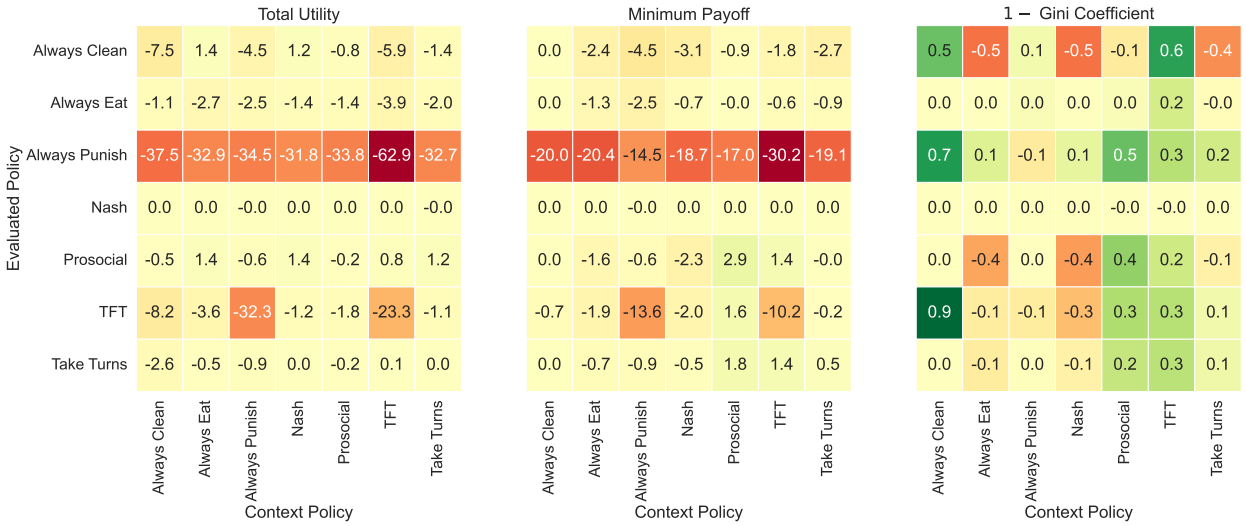


Figure B.2: Cooperativeness of seven deterministic policies in the 2-player Tabular Cleanup, in the context of each other policy, for all three welfare functions. The cooperativeness is valued on the initial state, with a discount factor $\gamma = 0.9$.

Figure B.3: Cooperativeness of eight deterministic policies in the 3-player Tabular Cleanup, in the context of two players playing the same context policy, for all three welfare functions. The cooperativeness is valued on the initial state, with a discount factor $\gamma = 0.9$.