A Computational Perspective on Incentives in Multi-Agent Systems

David Hyland

St Cross College

University of Oxford

A dissertation submitted to the Department of Computer Science at the University of Oxford in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

April 2026

Supervised by Michael Wooldridge and Julian Gutierrez

Abstract

This dissertation develops a computational perspective on the analysis and design of incentives in multi-agent systems. The main question that motivates this work is to understand how self-interest can be harnessed to achieve desirable outcomes, a question of increasing importance in a world populated by both human and artificial agents. The contributions of this work are structured in two parts, beginning with (1) a theoretical investigation into how methods from formal verification and synthesis can be applied to incentives for rational agents, and subsequently (2) developing models that account for real-world tradeoffs and constraints in decision-making from the perspective of information theory and cognitive science.

Part I investigates incentives for perfectly rational agents interacting strategically with one another to establish the fundamental possibilities, computational complexity, and limits of incentive design under different assumptions. We begin by exploring a version of the principal-agent problem under partial observability, where a principal must design contracts to align the outcomes of agents' decision making with their own objective. This analysis is then expanded to temporally extended concurrent games, where we study the power of different incentive schemes, including static and dynamic incentives, to achieve outcomes that satisfy specific properties. We then develop a general, logic-based framework for the automatic synthesis of incentive schemes in a quantitative model of concurrent games. Finally, this part explores the strategic considerations of endogenous incentives, where resource-bounded agents incentivise each other through mutual transfers of limited resources.

Part II shifts the focus to boundedly rational agents, motivated by the fact that perfect rationality, though a useful idealisation, is also often an intractable and descriptively inadequate model. Here, we develop novel frameworks for modelling decision making, belief updating, influence, and strategic interactions in the context of boundedly rational agents. We introduce the *Path Divergence Objective* (PDO), a model of decision making that captures the trade-off between

maximising utility and minimising the cognitive cost of updating from prior beliefs and habits. We then show that this framework provides a principled basis for explaining several cognitive phenomena such as curiosity, confirmation bias, and motivated reasoning. Following this, we explore a less traditional incentive pathway, modelling how an agent's choices can be influenced not by altering their values or beliefs, but by strategically directing their attention and the salience of different decision attributes. Extending the PDO to model strategic interactions between multiple agents, we introduce *Free-Energy Equilibria* as a new solution concept for boundedly rational agents in partially observable game theoretic settings and study some of its properties. Finally, we discuss the relevance of this work for the design of incentives for boundedly rational agents and conclude with a discussion of some of the open problems and concerns in this area.

To my parents, Charles & Samantha, and my brother, Christopher.

Acknowledgements

As far as it is possible to express gratitude through words on a page, I am immeasurably grateful for the people with whom I have had the fortune of travelling, however briefly, along this journey.

Firstly, to my supervisors, Mike and Julian, I am deeply thankful for the joy and privilege of knowing and working with you. Without your support, coaching, kindness, and unreserved willingness to help me through times of need throughout this entire adventure, the opportunity to bring this document into being would not have existed. Your extraordinary encouragement through both times of disappointment and success has provided me with the motivation to keep going countless times.

I am sincerely appreciative of the countless others who have contributed to my academic journey, especially the collaborators on different projects that I have had the honour of working alongside and learning from.

To Tomáš Gavenčiak, Lancelot Da Costa, and Lewis Hammond, I am deeply grateful for your innumerable contributions to my growth on this academic journey. You have been wonderful colleagues, mentors, and friends to me throughout my time here, encouraging me through some of the most difficult times and wholeheartedly sharing your resources, wisdom, and support whenever I was in need of them.

To Sarit Kraus, I am very thankful for your guidance, generosity, and patience in working together on extremely challenging but fulfilling projects. You have greatly inspired me to improve in all aspects of my work.

On top of this, to Alessandro Abate, Conor Heins, Edith Elkind, Giuseppe Perelli, James Fox, Jan Kulveit, Jiarui Gan, Mahault Albarracin, Muhammad Najib, Munyque Mittelmann, Nello Murano, Paul Harrenstein, Shankara Narayanan Krishna, Vojta Kovarik, and Will Lee, I owe an immense debt of gratitude to you all. It has been a great pleasure to work with and learn from you throughout various projects.

I would also like to say a heartfelt thank you to Jonas Hallgren for providing feedback on a draft of this thesis and many stimulating interactions, Ran Wei for our numerous enlightening and deep conversations which have provided me with a great deal of value of information, Dave Waters for your encouragement and support as my college advisor, Adam Goldstein for several illuminating discussions and incredible opportunities, and Pedro Ortega for the inspiring discussions and laying the foundation for the second part of this thesis.

To my examiners, Dave Parker and Alessio Lomuscio, thank you for your invaluable feedback on the content of this thesis and for the stimulating discussions during the viva voce examination, which have greatly improved the quality of this work.

To those who have enriched my time along the way, particularly Adam Safron, Adelaide Pitcock, Alessandra Yu, Ana Isabel Sousa, Anna Gautier, Arnau Quera-Bofarull, Athena Chow, Ayush Chopra, Bernhard Hilpert, Chloe Huang, Damian Pang, Dan Kwok, Daniel Ornia, Emanuel Tewolde, Emin Berker, Fernando Rosas, Gio Epifani, Guillem Monso, Harald Fredriksson, Heon Jin, Jirko Rubruck, Joel Dyer, Jonathon Liu, Katie Dock, Laila Shah, Laura Csuka, Leo Lind, Macken Murphy, Marco Gaspari, Maria Ambrosio, Matt MacDermott, Nick Bishop, Nicholas Teh, Nikki Vellidis, Peter Waade, Simon Lam, and Stefano Fronzoni, thank you for your friendship and support. The conversations, meals, activities, and events that we have shared are some of the most precious memories that I have made during my time in the DPhil.

To my dear friends from my time in Australia whom I miss dearly: Alec Zhang, Brendan Trinh, Brendon Wang, Colin Huang, David Gao, Harry Luong, Jack Ma, Johnny Lim, Leo Shu, Lloyd Morrison, Michael Li, Michael Zhao, Monica Lee, Sammy Wu, Sarah Chow, Tom Luo, Wei Wei Wang, Wynona Ly, and Yuweng Xue, I look back on our times together with particular fondness and appreciation. You have each been there for me and contributed to my growth in countless ways, and I give you my heartfelt thanks for that.

To my closest friends from home: Big Krishnamra, Mard Patana-Anake, Michael Laungjessadakun, Nicky Bain, Patrick Chou, and Tommy Boonchuaysream, thank you for all of the times and adventures that we have shared together. I have been extremely fortunate to know and grow with you, and I count it a great privilege to call you my friends.

To my teachers, mentors, and role models from high school and university who have made a profound impact in shaping the trajectory of my formative

years: Andrew Healey, Billy Molloy, Carl Turland, Dava Romyanond, Gauthier Voron, Hasindu Gamaarachchi, Jake Dodds, Jeff Rothwell, Karen Prout, Mark Hensman, Mick Farley, Paul Michael, Peter Kim, Vincent Gramoli, and Zhou Zhou, thank you for all that you have taught me, for believing in me, and for encouraging me. I would not have made it here without your efforts.

To my parents, Charles and Samantha, I am eternally grateful to you for your abundant love in providing me with spiritual, personal, and intellectual support throughout my life. Thank you very much for everything you have taught me, especially the values that I hold dear and for equipping me with so many of the tools that I have needed to navigate life.

Finally, to my brother Christopher, words cannot express how much I appreciate you, and what a blessing it is to have you in my life. You have unfailingly been there for me, encouraged me to develop in every aspect of my life, taught me so many things, and always looked out for me. Thank you from the bottom of my heart.

Contents

List of Tables

List of Figures

Notation

Basic Mathematical Sets, Symbols, and Conventions

$\mathbb{N}$	Natural numbers $\{0, 1, 2, \ldots\}$
$\mathbb{Z}$	Integers
$\mathbb{Q}$	Rational numbers
$\mathbb{R}$	Real numbers
$\mathbb{B}$	Boolean values $\{\top, \bot\}$ or $\{0, 1\}$
$\Delta(X)$	Probability distributions over the set X
2^X	Power set of X (set of all subsets of X)
$\lvert X \rvert$	Cardinality of the set X
$\emptyset$	Empty set
$\vec{x}$	Tuple $\vec{x} = (x_1, \ldots, x_n)$, usually indexed by players in a game
$\vec{x}_I$	Tuple $(x_i)_{i \in I}$ of elements of $\vec{x}$ indexed by elements in I
$\vec{x}_{-I}$	Tuple $(x_i)_{i \notin I}$ of elements of $\vec{x}$ indexed by elements not in I
$\sim$	Distributed as or sampled according to

Agents and Actions

N	Set of agents/players, typically $\{1, \ldots, n\}$
i, j	Individual agents, where $i, j \in N$
A_i	Set of actions for agent i
$\vec{A}$	Joint action space, $\vec{A} = \prod_{i \in N} A_i$
a_i	Action for agent i
$\vec{a}$	Action profile/joint action, $\vec{a} = (a_1, \ldots, a_n)$

Strategies and Policies

σ_i	Deterministic strategy for agent i
$\vec{\sigma}$	Strategy profile, $\vec{\sigma} = (\sigma_1, \ldots, \sigma_n)$
π	Policy for an agent
π_i	Policy for agent i
$\vec{\pi}$	Independent policy profile
$\vec{\mu}$	Correlated policy profile
Π_i	Set of policies for agent i

Utilities, Preferences, and Deviations

u_i	Utility function for agent i
$\mathcal{U}_i$	History/belief utility function
c_i	Cost function for agent i
c_i^*	Maximum cost for agent i
$\succeq_i$	Preference relation for agent i
$\succ_i$	Strict preference relation for agent i
$\sim_i$	Indifference relation for agent i
$\rightharpoonup_i$	Initial deviation for agent i
$\rightarrow_i$	Inducible deviation for agent i
$\leftrightarrows_i$	Soft deviation for agent i
$\rightrightarrows_i$	Hard deviation for agent i

Solution Concepts

$\text{NE}(G)$	Set of Nash equilibria of game G
$\text{CORE}(G)$	Core of game G
$\text{CCE}(G)$	Set of Coarse Correlated Equilibria of game G
$\text{DSE}(G)$	Set of Dominant Strategy Equilibria of game G
$\text{RE}_m(G)$	Set of m-Resilient Equilibria of game G
$\text{IFEE}(G)$	Set of Independent Free-Energy Equilibria of game G
$\text{CCFEE}(G)$	Set of Coarse Correlated Free-Energy Equilibria of game G

Boolean Games

B	Boolean game
Φ	Set of Boolean variables $\{p_1, p_2, \ldots, p_m\}$
Φ_i	Set of Boolean variables controlled by agent i
γ_i	Goal of agent i (propositional formula)
v_i	Valuation/strategy for agent i

Observability and Contracts

$\mathcal{O}$	Observable set (subset of variables)
o	Observation
Ω_i	Set of possible observations for agent i
O_i	Observation function for agent i
$=_{\mathcal{O}}$	Observational equivalence relation
$\neq_{\mathcal{O}}$	Observational distinguishability relation
$\|_{\mathcal{O}}$	Restriction to observable variables
κ	Contract $\kappa = (\kappa_1, \ldots, \kappa_n)$
κ_i	Payment function for agent i

Concurrent and Stochastic Games

Game Structures

$\mathcal{A}$	Concurrent game/Reactive Modules Game Arena
$\mathcal{G}$	Concurrent Game/Reactive Modules Game
S	Set of states
x^t	State of x at time t
x_i^t	Component i of x at time t
$x^{s:t}$	Sequence of values of x from time s to t inclusive, with $s \leq t$
$x^{\geq s}$	Sequence of values of x from time s onwards
$x^{\leq s}$	Sequence of values of x up to time s
I	Initial state distribution
T	Time horizon
$\mathcal{T}$	State transition/transformer function
p	Probabilistic transition function

(Partially Observable) Markov Decision Processes

$\mathcal{M}$	Markov Decision Process (MDP) or POMDP
h	History
$h^{0:t}$	History from time 0 to t
$s^{0:t}$	State trajectory from time 0 to t
$a^{0:t}$	Action trajectory from time 0 to t
$o^{0:t}$	Observation trajectory from time 0 to t
$\mathbb{H}$	Set of histories
$\mathbb{O}$	Set of observation histories

Temporal Logic

φ, ψ	Propositional or temporal logic formulae
γ	Goal formula
$\models$	Satisfaction relation
$\top$	True/tautology
$\bot$	False/contradiction
$\neg$	Negation
$\wedge$	Conjunction (and)
$\vee$	Disjunction (or)
$\rightarrow$	Implication
$\leftrightarrow$	Bi-implication
$\mathbf{X}$	Next operator
$\mathbf{F}$	Eventually operator
$\mathbf{G}$	Always operator
$\mathbf{U}$	Until operator
ρ	Run/execution sequence
$\rho(\vec{\sigma})$	Run induced by strategy profile $\vec{\sigma}$

Information Theory and Probability

$\mathbb{E}_P[X]$	Expectation of X under distribution P
$\mathbb{E}_P[X\|Y]$	Conditional expectation of X given Y under distribution P
$p(\cdot)$	Probability distribution/mass function
$p(\cdot\|\cdot)$	Conditional probability
$\mathrm{D}_{\mathrm{KL}}[P \mid\mid Q]$	Kullback-Leibler divergence from Q to P
$H[X]$	Shannon entropy of X
Z	Partition function
β	Rationality parameter/inverse temperature

Bounded Rationality

$G(Q; Q^0)$	Path Divergence Objective (PDO)
Q	Agent's world model
$\mathcal{Q}$	Set of possible world models for an agent
$\tilde{P}$	Preference model
π^0	Prior policy
$\mathcal{V}$	Value function

Complexity Classes

P	Polynomial time
NP	Non-deterministic polynomial time
coNP	Complement of NP
PSPACE	Polynomial space
EXPTIME	Exponential time
NEXPTIME	Non-deterministic exponential time
k-EXPTIME	k-exponential time
Σ_k^p	k-th level of the polynomial hierarchy
Π_k^p	Co-Σ_k^p

1

Introduction

How can agents coordinate their behaviours in service of a particular purpose or function? This question can be read in at least two ways. On one hand, it can be read as a scientific question about the mechanisms that govern interactions between agents. On the other hand, it can be read as an engineering question about how we can design systems of interaction that promote or discourage particular outcomes. What is the language that we might use to articulate this question more precisely and make progress on it? I would argue that the language we are looking for is that of *incentives*.

Incentives are stimuli, conditions, or events that causally motivate an agent to act in a particular way. They can be influenced by existing physical, technological, economic, social, cultural, and political structures, as well as by the actions of the very agents who are influenced by them. Incentives are crucial to our understanding of how economies, organisations, and biological systems operate, and what behaviours they might be more or less likely to display. They have also been instrumental to the formation and sustainment of human civilisation itself, from the localised societies of our ancestors to the highly interconnected global one that we live in presently. As the capabilities of artificial agents rapidly expand in the wake of the machine learning revolution, these agents are increasingly being integrated into society. They are being granted more autonomy and are becoming increasingly capable of making decisions that have significant consequences for the well-being of others, including the capability to interact with the World Wide Web on their users' behalf [1–3]. This ongoing integration

of agents into human society is giving rise to what has been described as *Human-Agent Collectives* [4], which are characterised by "many distinct stakeholders, each with their own resources and objectives, [which] means participation is motivated by a broad range of incentives." As these collectives grow in size and diversity to include a range of heterogeneous agents with differing embodiments, preferences, resources, and capabilities, our ability to understand and guide the emergent behaviours of these systems is diminishing rapidly, with significant consequences for the various ecosystems that we depend upon every day.

Incentives are a powerful lens through which to understand human-agent collectives because fundamentally, it is not within the capability of any entity to control all of the agents that make up the collective. Absent the ability to manipulate the neuronal firing patterns within people's brains and gain control over the source code of deployed autonomous agents, the primary means by which influence can be exerted on these agents is through the use of incentives. Incentives shape the affordances perceived by agents, which influences the likelihood with which they take particular courses of action or arrive at particular beliefs. When sufficiently advanced agents make decisions in the presence of other agents, the landscape of decision making for a given agent depends in general not only on their surrounding environment, but also on the decisions of those other agents. This can significantly complicate the problem of inferring how agents will respond to different affordances, and hence the problem of designing incentives for *multi-agent systems* (MAS) is often significantly more challenging than it is for single-agent systems.

What lies before us in light of recent advances in Artificial Intelligence (AI) is the prospect of an agentic web, characterised by the proliferation of autonomous goal-driven agents interacting with both humans and each other [5], which could significantly transform the nature of how information, work, and resources are distributed and deployed in society. The successful realisation of such a web could result in significant advances in coordination, cooperation, and productivity throughout the world. The promise of this vision is apparent: automating the design, implementation, dissemination, execution, monitoring, and evaluation of incentives could significantly expand the scope and scale of coordination among the vast and growing societies of agents that exist. Moreover, the programs used in each stage of the pipeline could be transparently audited and verified by multiple parties along a number of dimensions, including for example safety, justice, fairness, and regulatory compliance. The development and refinement of such technology might culminate in an *'incentive compiler'*,

which could automatically translate goal and system specifications written in a high-level language into verifiably effective and safe incentive schemes, along with agents that work to automate its implementation and maintenance.

While the prospect of this technology is enticing, it overlooks many important considerations that the real world and our nature as human beings impose on us. Firstly, given our current technological capabilities, it is currently infeasible to adequately model the real world and our nature as human beings as inputs to a computer program. Even if such a feat were possible, as the results in this thesis indicate, the computational difficulty and cost of running such a program would be prohibitively high for any reasonably large-scale systems, limiting their applicability primarily to small-world domains. Moreover, even in the cases where such a program could be run, this would still leave open the question of *how* we determine to what extent different outcomes are (un)desirable, and how to precisely specify this for a machine[1]. Nonetheless, the process of studying idealised and simplified models can yield insights into what the important considerations are in the science and engineering of incentives. In addition, they may allow us to understand the trade-offs that arise between, for instance, the expressiveness and the complexity of different models of incentives and agents.

On the other side of this positive vision, the implementation of autonomous goal-directed multi-agent systems without sufficient safety measures in place could lead to many kinds of catastrophes [7]. Indeed, a phenomenon recently described as *Moloch's bargain* [8] highlights how the optimisation of Large Language Models (LLMs) in competitive settings can lead to an emergent form of misalignment of incentives between the models and the humans whom they interact with. In their study, the authors found that optimising models to produce convincing sales pitches or campaign statements led to an increased propensity to generate misleading or false information in order to persuade potential customers or voters. This foregrounds a crucial point that needs to be considered in multi-agent systems: the presence of misaligned incentives between different agents can lead to novel failure modes and negative externalities. Thus, the question of how to identify, design, and verify incentives for agents in an increasingly technologically augmented society is becoming more important to ensure that desired outcomes are achieved and that undesired outcomes are avoided [9, 10].

The motivating reason behind this study of incentives is to *understand and design systems of interaction that achieve better outcomes than would otherwise be*

[1]This is also known as the *outer alignment problem* [6].

attainable without the interventions. Some examples of what 'better' might mean could include promoting greater resilience and efficiency [11], encouraging internalisation of the externalities produced by one's actions [12], and achieving mutually beneficial (or 'win-win') outcomes [13]. For instance, a firm might successfully deploy performance-related pay to encourage higher employee productivity, driving up profits and the wages that workers take home. In this sense, incentives serve as mechanisms contributing to motivation that channel self-interest towards a broader goal or desired outcomes. Crucially, incentives can take many different forms, ranging from material (money, goods) to social (reputation, status) and internal psychological rewards (meaning, satisfaction, purpose). These different kinds of incentive often overlap and interact with each other in real-world decisions, making a full analysis of the incentives that bear on any particular situation extremely difficult.

What these different different kinds of incentives have in common is that they reliably *motivate* people to act in particular ways, which allows systems to be built that utilise them to organise behaviour. Understanding this from the foundations of economics, Adam Smith famously wrote in 1776 [14]:

> *It is not from the benevolence of the butcher, the brewer, or the baker that we expect our dinner, but from their regard to their own self-interest. We address ourselves not to their humanity but to their self-love, and never talk to them of our own necessities, but of their advantages.*
>
> *— Adam Smith, The Wealth of Nations (p. 26-27) [14]*

Of course, this is not to say that virtues such as benevolence and generosity do not motivate people, but that by means of appropriately designed incentive structures, one need not hope for the goodwill of others to ensure that a mutually beneficial outcome is achieved; instead, one can, with greater confidence, *rely on the self-interest and agency of others* within the context of an appropriate system of incentives to achieve these aims. Moreover, self-interest can be harnessed towards productive ends, such as the discovery of new knowledge and technologies through the process of competition [15].

A central message of this dissertation is that, to the extent that it is feasible and reasonable to do so, incentives can be viewed as a programmable and verifiable substrate of the interactions between agents, both natural and artificial. Failing to make use of the affordances that computers provide to engage in this task

would be to squander the opportunity to approach some of the largest collective opportunities and challenges that we face thoughtfully and deliberately, and would leave us ill-equipped to navigate the increasingly complex and agentic world that we are building. However, due to the breadth of the topic and the range of disciplines that it intersects with, a fully general treatment of all kinds of incentives is beyond the scope of this thesis. For this reason, we will focus on a particular subset of the field, specifically the factors that give rise to *limitations* on the ability of incentive designers and agents to achieve their desired outcomes.

1.1 Thesis Overview

This thesis investigates two main topics: how to harness self-interest subject to informational, computational, and strategic limitations to achieve desirable outcomes in multi-agent systems, and modelling the ways in which real-world agents such as humans respond in the presence of incentives.

Chapter 2: Background

Chapter 2 presents an overview of the central concepts and formal models related to this thesis. We also briefly introduce some of the core ideas related to the central topics of this thesis such as formal verification, multi-agent systems, game theory, and incentive design.

Part I: Incentives for Rational Agents

In the first part of this thesis, we aim to understand the factors that contribute to our ability to effectively automate processes of incentive design and verification using computers. We explore various settings of incentive design for rational agents, with particular emphasis on multi-agent systems. Crucially, this introduces the complication of the incentive designer having to take into consideration the interactions between agents in their responses to incentives.

Chapter 3: Principal-Agent Boolean Games

Chapter 3 introduces and studies a simple hidden action multi-agent incentive design problem, where the agents' actions are specified by setting the values

of a set of Boolean variables under their control. Agents are assumed to have lexicographic preferences, which are characterised by a logical formula over the Boolean variables and a cost function that assigns to each realisation of the variables a cost that the agents desire to minimise. Agents will always prefer outcomes that satisfy their logical goal over those that do not, and will secondarily prefer to minimise costs. The principal possesses only partial observability over the variables, in that they are only able to observe a subset of them. The principal's challenge is then to design a contract, that is, a mapping from observed outcomes to payments for each agent such that, if the agents are rational (i.e., make choices that form a game-theoretic equilibrium), then, firstly, the principal's goal will be accomplished, and secondly, the principal can formally and automatically verify that the goal is accomplished. We first study the computational complexity of the problem of verifying whether the principal's objective has been achieved under a given contract. We classify different equilibria based on the principal's capability to eliminate them by the choice of an appropriate contract. Finally, we study the computational complexity of the problem of Nash contractibility, which comes in two flavours. The existential (E) and universal (A) Nash contractibility problems ask whether a contract exists such that the principal can verifiably ensure that their objective, specified as a propositional logic formula, has been satisfied in some or all Nash equilibria (i.e., stable outcomes) induced by the contract.

Chapter 4: Incentive Engineering for Concurrent Games

Chapter 4 expands some of the concepts in Chapter 3 to a setting where agents interact in a much richer game model known as a concurrent game, which is played over an infinite number of rounds. Here, as in Chapter 3, agents are assumed to have lexicographic preferences, where this time the logical goal is specified by a linear temporal logic formula, which is a highly expressive language that allows the succinct representation of temporal goals such as "eventually p will be true" or "Q will always be true". Their secondary quantitative goals are a limiting average of the costs incurred along a run of the game. Here, the principal can modify the costs associated with taking different actions in different states by applying a *taxation scheme*, which can be static or dynamic. Static taxation schemes are applied at the beginning of the game and remain fixed throughout its play. In contrast, dynamic taxation schemes permit the memoryful application of incentives based on the history of the game. We again characterise the different kinds of equilibria that arise based on their eliminability, and study the existential and universal Nash implementation problems, which ask whether

a taxation scheme exists that renders the resulting equilibria acceptable to the principal based on the satisfaction of their goal, which is specified as a linear temporal logic formula.

Chapter 5: A General Logic-Based Approach to Automatic Incentive Design

Chapter 5 introduces a general approach to automatic incentive verification and design, harnessing the expressivity of $\mathsf{LTL}[\mathcal{F}]$, a variant of linear temporal logic that allows one to reason quantitatively over traces in a system's execution. Instead of assuming lexicographic preferences here, we assume that agents' goals are specified by $\mathsf{LTL}[\mathcal{F}]$ formulae, which can naturally encode both logical and quantitative preferences. Here, incentives are interpreted as assignments of satisfaction values to a set of propositional variables, and as with Chapter 4, we consider both static and dynamic incentive schemes. We study the computational complexity of the problems of: (1) *incentive verification*, which is the problem of verifying whether the satisfaction value of a given $\mathsf{LTL}[\mathcal{F}]$ formula exceeds a certain threshold, and (2) *incentive synthesis*, which is the problem of designing an incentive scheme such that the same holds for all strategy profiles that satisfy a particular solution concept. We show that these problems can be solved by model checking appropriately constructed $\mathsf{SL}[\mathcal{F}]$ formulae, which are an extension of $\mathsf{LTL}[\mathcal{F}]$ with strategy quantifiers, on a transformed version of the game. This allows the use of extensively developed model checking algorithms to determine the feasibility of incentive verification and synthesis in a general way.

Chapter 6: Endogenous Incentives: Agents Incentivising Each Other

Chapter 6 concludes Part I by taking a different perspective on incentives which eschews the top-down approach of having a principal design and impose incentives on agents, and instead studies how agents incentivise each other by making offers of resource transfers to one another. We build on top of the model of *Reactive Modules Games* [16], which are a succinct representation of concurrent games, by introducing energy budgets, which are used to model resource-bounded agents. We then introduce *Endogenous Energy Reactive Modules Games*, where agents can offer to transfer their energy to other agents prior to the beginning of the game. This introduces a two-stage game, where the notion of stability of a profile of offers made by the players rests upon the feasible outcomes in the induced game. We study the computational complexity of the rational verification problems in

both the cooperative and non-cooperative settings for our game model, as well as the problem of determining whether a stable offer profile exists for different combinations of solution concepts in the different stages of the game.

Part II: Incentives for Boundedly Rational Agents

In Part II of this thesis, we develop and explore a foundation for modelling both natural agents such as humans and artificial agents, and identify some of the resulting differences that arise in the incentive design problem as a result of bounded rationality.

To begin with, we present a framework for modelling bounded rationality using concepts from information theory. Then, we explore the implications of this model for policy selection and belief change. Following this, we present a novel perspective on how attention-driven preferences can be used to influence the behaviour of an agent under a simple idealised model of how attention influences decision making. Finally, to conclude, we extend the information-theoretic models developed previously to capture interactions between multiple boundedly rational agents in partially observable environments and explore the subsequent implications of such models for incentive design.

Chapter 7: A Model of Bounded Rationality

Chapter 7 introduces a framework for modelling boundedly rational agents with world models in partially observable environments. We build on top of an information-theoretic model of bounded rationality and introduce the *Path Divergence Objective* to capture the trade-offs between information-processing costs and utility that real-world agents face. We study decompositions of this objective and relate it to the pragmatic and epistemic motivations described in active inference, which is a biologically inspired framework for modelling the decision making of embodied agents. This decomposition also reveals different motivational pathways that can be utilised to incentivise agents. We provide an algorithm for optimising the Path Divergence Objective in perfect recall environments that, while exponential in the horizon, is optimal given the agent's informational constraints, and study its behaviour compared to expected value and expected free energy optimisation.

Chapter 8: Boundedly Rational Belief Change

Chapter 8 applies the model of bounded rationality developed in Chapter 7 to belief change. Historically, the study of bounded rationality has been mostly focused on limitations in decision making and policy selection. Here we propose and demonstrate a simple model showing how the application of the same principles to belief change can give rise to a number of so-called biases that we observe in human epistemics. Drawing on a number of recent works highlighting the role of motivation and utility in human belief change, we implement a set of simple computational experiments to show qualitatively how phenomena such as confirmation bias and attitude polarisation can arise from the model of bounded rationality proposed thus far. Since decision making is commonly decomposed into beliefs and preferences, understanding the way that incentives and informational costs affect belief updating has significant implications for effective incentive design, particularly for systems involving humans.

Chapter 9: Attention and Influence

Chapter 9 studies a model of multi-alternative, multi-attribute decision making, where an agent allocates their attention across different attributes of the alternatives they are choosing between. We begin by presenting a representation theorem for preferences that are influenced by attentional weighting. We then provide an algorithm for deciding whether or not a given alternative can be made the most preferred one under some salience distribution. Finally, we introduce and demonstrate a simple Bayesian model in which a principal can induce preference shifts in an agent without changing their subjective valuations or beliefs about the alternatives. This highlights another important pathway of influence over agents that can be achieved in the absence of material incentives or information asymmetries.

Chapter 10: Incentives and Interactions Between Boundedly Rational Agents

Chapter 10 extends the model of bounded rationality presented in Chapter 7 to capture interactions between multiple boundedly rational agents in partially observable environments, bridging the game-theoretic concepts explored earlier in this thesis with the model of real-world decision making proposed in Chapter 7. We introduce the concept of *Free Energy Equilibria*, a new class of game-theoretic solution concepts, which are 'equilibria' in which each agent minimises their

Path Divergence Objective. We study properties of this class of solution concepts under different assumptions and explore the additional considerations that a multi-agent setting introduces to the problem of incentive design in the context of boundedly rational agents.

1.2 Contributions

This thesis contains work reviewed and presented at the following conferences and workshops:

- Chapter 3: **Hyland, D.**, Gutierrez, J., Wooldridge, M. Principal-Agent Boolean Games. *International Joint Conference on Artificial Intelligence*, 2023.
- Chapter 4: **Hyland, D.**, Gutierrez, J., Wooldridge, M. Incentive Engineering for Concurrent Games. *Theoretical Aspects of Rationality & Knowledge*, 2023.
- Chapter 5: **Hyland, D.**, Mittelmann, M., Murano, A., Perelli, G., Wooldridge, M. Incentive Design for Rational Agents. *International Conference on Principles of Knowledge Representation and Reasoning*, 2024.
- Chapter 6: Gutierrez, J., **Hyland, D.**, Najib, M., Perelli, G., Wooldridge, M. Endogenous Energy Reactive Modules Games - Modelling Side Payments Among Resource-Bounded Agents. *International Joint Conference on Artificial Intelligence*, 2024.
- Chapter 7: Gavenčiak, T.*, **Hyland, D.***, Da Costa, L., Wooldridge, M., Kulveit, J. Path Divergence Objective: Boundedly Rational Decision Making in Partially Observable Environments. *NeurIPS NeuroAI Workshop*, 2024.
- Chapter 8: **Hyland, D.**, and Albarracin, M. On the Variational Costs of Changing our Minds. *International Workshop on Active Inference*, 2025.
- Chapter 10: **Hyland, D.***, Gavenčiak, T.*, Da Costa, L., Heins, C., Kovarik, V., Gutierrez, J., Wooldridge, M., Kulveit, J. Free Energy Equilibria: Toward a Theory of Interactions Between Boundedly Rational Agents. *ICML Workshop on Models of Human Feedback for AI Alignment*, 2024.

Chapter 9 contains an earlier version of unpublished work in progress:

* Equal contribution.

- **Hyland, D.**, Kraus, S., Wooldridge, M. Salient Preference Dynamics: A Model of Attention-Driven Preference Change. 2025.

Work completed during the course of my DPhil that is not presented here:

- Lee, W.C.*, **Hyland, D.***, Abate, A., Elkind, E., Gan, J., Gutierrez, J., Harrenstein, P., Wooldridge, M. k-Prize Weighted Voting Games. *International Conference on Autonomous Agents and Multiagent Systems*, 2023.
- Abate, A., Almulla, Y., Fox, J., **Hyland, D.**, Wooldridge, M. Learning Task Automata for Reinforcement Learning using Hidden Markov Models. *European Conference on Artificial Intelligence*, 2023.
- **Hyland, D.**, Gutierrez, J., Krishna, S.N., Wooldridge, M. Rational Verification with Quantitative Probabilistic Goals. *International Conference on Autonomous Agents and Multiagent Systems*, 2024.
- Da Costa, L., Gavenčiak, T., **Hyland, D.**, Samiei, M., Dragos-Manta, C., Pattisapu, C., Razi, A., Friston, K. Possible principles for aligned structure learning agents. arXiv preprint arXiv:2410.00258, 2024.
- Albarracin, M., de Jager, S., and **Hyland, D.**, The Physics and Metaphysics of Social Powers: Bridging Cognitive Processing and Social Dynamics, a New Perspective on Power Through Active Inference. *Entropy*, 2025.

2
Background

To provide a broad overview of the relevant literature, it will be helpful to break down the problem that we are interested in studying into some of its key constituent components. Firstly, when presented with any formally specified system, we will need to be able to ask questions about the kinds of behaviour that can be expected of the system and/or its components. We will thus begin by describing core ideas in formal verification and synthesis. Following this, the primary objects of study are multi-agent systems, so we will begin with a summary of some of the key components needed to model multi-agent systems. This will begin with a discussion of the central models of agency that recur throughout this thesis, as well as descriptions of the environments that agents inhabit. Subsequently, we dive deeper into the field of game theory and explore the relevant branches of this area to the present text. Finally, we present a broad overview of the aspects of the study of incentives that are related to the work contained herein, primarily focusing on the economics and computer science literature. We will defer formal definitions of concepts relevant to specific chapters in this dissertation to those chapters, only including a selection of the core formalisms that recur throughout.

2.1 Formal Verification and Synthesis

Modelling Frameworks

Research questions are often conceptualised and studied in the context of a model. In computer science, we are interested in a formal model of the system to be studied, which includes specifying its syntax and semantics. Two broad classes of approaches to modelling multi-agent systems are:

1. *Analytical approaches*, where properties of the model (e.g., decidability, computational complexity, approximation guarantees) are formally proven by reasoning from the initial assumptions;

2. *Data-driven approaches*, where either real-world data is collected under controlled or natural experiments [17], or a parameterised model of the system is simulated [18, 19]. Then, conclusions are drawn based on the observed data [20].

Both frameworks have their advantages and disadvantages, trading off crucial properties such as the ability to provide formal proofs and guarantees of correctness/runtime/approximation bounds, verisimilitude of the model, and the ability to test and fit hypotheses with real-world data. Naturally, analytical approaches are more helpful in advancing our understanding of MAS, while a more balanced mixture will be needed in building and deploying such systems.

Model Checking

First, we will discuss some of the core ideas and methods used in model checking, which forms the foundation for many of the ideas developed later in the context of MAS. At its core, model checking is a formal method for verifying that the behaviour of some model of a system satisfies particular properties [21, 22]. The two main under-specified components of this definition are a model and a property. Indeed, the general model checking question can be stated slightly more formally as: given a model M of a system and a behavioural property φ, does M verify φ ($M \models \varphi$)?

One of the central motivating forces behind the invention (or discovery) of models to represent concurrent systems was the need to understand and manage

them. As networked computer systems became increasingly ubiquitous with the rapid development of more affordable and powerful digital computers, the complexity of these interconnected systems also grew so rapidly that people recognised the need to find better ways of guaranteeing that these systems operated 'correctly' [23]. Prior to the development of these techniques, methods for the formal verification of such models relied on humans manually proving that some logical formula is satisfied under some or all of the possible executions of the system [24, 25]. With the advent of algorithmic procedures for model checking, the verification of whether certain classes of systems satisfy given properties could be done automatically [23]. However, this does not imply that model checking has ushered in an era of bug-free software. Indeed, the two core components of any model-checking procedure also pose the most significant challenges to the method's universal adoption, which are (1) state-space explosions when modelling real-world programs and systems [26] and (2) the difficulty in formally specifying what it means for a system to be 'correct.' Nonetheless, model checking has proved to be an invaluable tool for assisting system designers in verifying that their creations satisfy the most important desirable properties.

Models of Concurrent Systems

In some of the first works discussing model checking, models were used to formally describe *finite-state concurrent systems* [27, 28]. Two of the most prominent semantic models used to do so are known as *labelled transition systems* (LTS) and *Kripke structures* (KS). Both models share many common features by representing a system using (1) a set of the system's possible states or configurations, (2) a set of transition rules that specify, given a state of the system, which states the system may transition to next, (3) a set of possible initial states of the system, (4) a set of propositional (Boolean) variables, and (5) a labelling function that assigns truth values to the set of propositional variables, either in each transition of the system (LTS) or in each state of the system (KS) [29].

Definition 2.1. A *Kripke Structure* over a set Φ of atomic propositional variables is a 4-tuple $K = (S, I, \mathcal{T}, \mathcal{L})$, where

1. S is a finite set of states;
2. $I \subseteq S$ is a set of initial states;

3. $\mathcal{T} \subseteq S \times S$ is a left-total transition relation; and

4. $\mathcal{L} : S \to 2^{\Phi}$ is a labelling function, that specifies which atomic propositions are true in each state of the system.

The execution of both models occurs by the repeated application of some rule that selects, given a finite sequence of previous states and transitions of the system, which state the system will transition to next. This amounts to a resolution of the possible nondeterminism encoded by the model, which can be achieved by having some agent(s) *deterministically choose* the next state, by *randomly sampling* the next state according to some probability distribution, or some combination of both [21]. The two models are equivalent in the sense that either one can be encoded in terms of the other [30]. However, depending on the system being modelled, it may be more natural to reason about sequences of states rather than transitions or vice versa. Going forward, we will primarily discuss model checking in the context of KS with the understanding that most of these ideas can also be understood and formulated using LTS.

Logics for Model Checking

Now, we focus on the second important component of any model-checking problem – the property to be checked. There are two main classes of methods used in the model-checking literature to specify properties – logical formulae and finite automata. Firstly, we will briefly discuss some of the key instantiations of these methods in the context of model-checking.

Recall that one component of a KS is a labelling function that assigns truth values to a fixed set of propositional variables in each state of the system. Because an execution (aka a *run*) of the system corresponds to an infinite word of state-transition pairs, every run induces an infinite sequence of valuations over all of the propositional variables under consideration. Formally, given an alphabet of symbols A, an *infinite word* is an infinite sequence of symbols from A, which we typically write as $\rho = \alpha^0\alpha^1 \ldots \alpha^n \ldots$, where $\alpha^i \in A$ for all $i \in \mathbb{N}$. A *run* of a KS is an infinite word $\rho = s^0 s^1 \ldots s^n \ldots$, where $s^i \in S$ and $(s^i, s^{i+1}) \in \mathcal{T}$ for each $i \in \mathbb{N}$. To reason about infinite sequences of truth assignments, extensions of the classical propositional logic had to be developed. Out of this necessity, the family of *temporal logics* was born, which has had a tremendous and growing impact on the field of computer science. Two particularly influential temporal logics that were developed early on are *Linear Temporal Logic* (LTL) [31] and *Computation*

Tree Logic (CTL) [27, 32]. Both of these logics are types of *propositional temporal logics* which extend classical propositional logic using temporal operators, but they differ in their syntax and semantics. In particular, their semantics encode two different views on time as either a linear progression of events or as a branching tree of alternatives. In the linear-time viewpoint, temporal reasoning is performed with respect to infinite sequences of valuations of the propositional variables that are induced by runs of the system. In the branching-time view, one instead considers possible executions of the system from a given state as a tree labelled with valuations.

Linear Temporal Logic Throughout Part I of this dissertation, we make heavy use of the standard language of *Linear Temporal Logic* (LTL) to express properties of infinite words [31, 33, 34]. As described above, runs of a KS give rise to an infinite word where each element of the word represents a valuation of the propositional variables under consideration, which we denote by the set Φ. Hence, we can interpret formulae of LTL with respect to runs of a KS through the infinite word of valuations that it generates. LTL extends classical propositional logic with tense operators $\mathbf{X}$ ("in the next state…"), $\mathbf{F}$ ("eventually…"), $\mathbf{G}$ ("always…"), and $\mathbf{U}$ ("…until…") [34]. Formally, the syntax of LTL is defined with respect to a set of Boolean variables Φ by the following grammar:

$$\varphi ::= \top \mid p \mid \neg\varphi \mid \varphi \vee \varphi \mid \mathbf{X}\varphi \mid \varphi \,\mathbf{U}\, \varphi$$

where $p \in \Phi$. The remaining classical connectives ("$\bot$", "$\wedge$", "$\rightarrow$", "$\leftrightarrow$") are defined in terms of $\neg$ and $\vee$ in the conventional way. The LTL operators $\mathbf{F}$ and $\mathbf{G}$ are defined in terms of $\mathbf{U}$ as follows: $\mathbf{F}\varphi = \top \,\mathbf{U}\, \varphi$, and $\mathbf{G}\varphi = \neg\mathbf{F}\neg\varphi$. Given a set of variables Φ, let $LTL(\Phi)$ be the set of LTL formulae over Φ; where the variable set Φ is clear from the context, we simply write LTL.

We interpret formulae of LTL with respect to pairs (ρ, t) where ρ is a run of a KS and $t \in \mathbb{N}$ is a temporal index into ρ, interpreting ρ as a sequence of valuations over the variables in Φ. We will write $(\rho, t) \models \varphi$ to mean that $\varphi \in LTL$ is true at time $t \in \mathbb{N}$ on run ρ. Additionally, we write $\rho[t]$ to denote the state of the KS on run ρ at time t. The semantics of LTL (i.e., the rules specifying when

formulae are true) are defined as follows:

$$
\begin{array}{lll}
(\rho, t) \models \top & & \\
(\rho, t) \models x & \text{iff} & x \in \mathcal{L}(\rho[t]) \qquad \text{(where } x \in \Phi\text{)} \\
(\rho, t) \models \neg\varphi & \text{iff} & \text{it is not the case that } (\rho, t) \models \varphi \\
(\rho, t) \models \varphi \vee \psi & \text{iff} & (\rho, t) \models \varphi \text{ or } (\rho, t) \models \psi \\
(\rho, t) \models \mathbf{X}\varphi & \text{iff} & (\rho, t+1) \models \varphi \\
(\rho, t) \models \varphi \,\mathbf{U}\, \psi & \text{iff} & \text{for some } t' \geq t : (\rho, t') \models \psi \text{ and} \\
 & & \text{for all } t \leq t'' < t' : (\rho, t'') \models \varphi.
\end{array}
$$

We write $\rho \models \varphi$ as a shorthand for $(\rho, 0) \models \varphi$, in which case we say that ρ *satisfies* φ. A formula φ is *satisfiable* if there is some word satisfying φ. The *size* of an LTL formula φ, written $|\varphi|$, is given by the number of subformulae in φ. Checking satisfiability for LTL formulae is known to be PSPACE-complete [35]. Moreover, checking synthesis for LTL is 2EXPTIME-complete [36].

Many other logics have been proposed over the years to capture different properties of the systems being modelled, such as CTL*, which is a branching-time logic that subsumes both CTL and LTL [37], *Alternating-time Temporal Logic* (ATL) which is also an extension of CTL to reason more naturally about multiple agents in a system [38], and the *modal μ-calculus* [39] which is an extension of modal logic to include fixed-point operators that can be used to encode CTL* formulae [29].

Another more general approach to specifying properties of interest is to consider the class of ω-regular languages, which are a generalisation of the regular languages to infinite words/runs [40]. Just as with regular languages, the class of ω-regular languages has a correspondence with the languages accepted by a certain class of finite-state automata whose accepting conditions are specified in terms of infinite runs [33]. These are all referred to as *ω-automata*, and the canonical ones are known as (nondeterministic) Büchi, parity, Rabin, Streett, and Muller automata. These types all share the same underlying state and transition model and differ only in their acceptance conditions, which are specified as different rules regarding which states can or must be visited a finite or infinite number of times in any accepting infinite execution. We will not go into the details of how each differs, as the critical relevant property to our discussion on model checking that is common to all of these automata is that they can be used to recognise ω-regular languages. Thus, the problem of model checking ω-regular languages (of which LTL defines a subset) can be reduced to checking

for non-emptiness or containment of the languages accepted by an appropriate ω-automaton [29]. Similarly, for branching-time temporal logics like CTL and CTL*, the process of model-checking a formula in these logics can be translated into queries about the languages accepted by a specific alternating automaton [41, 42].

Finally, we briefly summarise the computational complexity of the model-checking problem for the various specification logics and languages outlined here. Firstly, LTL model checking is PSPACE-complete, and CTL model-checking is P-complete [43]. Additionally, model-checking is P-complete for ATL [38] and is PSPACE-complete for both CTL* and the linear modal μ-calculus [43, 44].

Specification Models

In determining the degree to which a given system's behaviour is desirable, the three main approaches are qualitative, quantitative, and mixed specifications. Qualitative specifications typically take the form of logical formulae defined over the set of possible outcomes of the system (see Section 2.4). Quantitative measures of outcomes are mathematical functions that map executions of the system to real-valued numbers or vectors. A typical example of such a measure is an aggregating function of the agents' utilities, such as a social welfare measure [45, 46]. Alternatively, a utility/reward function may be defined for the system analyst to measure some other property of the system that they are trying to optimise. Finally, mixed specifications incorporate both qualitative and quantitative aspects of specifications, including extensions of Linear Temporal Logic (LTL) [47] or Strategy Logic (SL) [48] to reason about the quality of an execution.

2.2 Agent Models

There have been many proposals over the years as to what constitutes an agent, and the definition settled on in [49] is that "*an autonomous agent is a system situated within and a part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to effect what it senses in the future*." Another widely adopted definition of an agent in [50] takes the approach of defining an agent by its essential properties, which consist of (1) *autonomy* (operation without direct intervention), (2) *social ability* (interaction with other agents), (3) *reactivity* (ability to perceive the environment and respond promptly to changes therein), and (4) *pro-activeness* (ability to act in the pursuit of goals). Following [49], we

generally take an agent to be a system that is situated in an environment, who can be usefully described as being able to act on that environment in order to achieve its goals[1]. One characteristic of agents that we find to be useful in ascribing agency is that they are able to model their own future states and those of their environment and take actions to realise some subsets of those futures according to some criterion. This ability may be present in varying degrees in different systems, suggesting that a graded conception of agency may be more useful as opposed to a binary one.

The space of agent architectures has been richly populated by several decades of work in the design of intelligent agents. Agent architectures can be classified as deliberative, reactive, or hybrid architectures [51]. Deliberative architectures aim to explicitly capture the notion that agents contain a symbolically represented model of the world and make decisions on the basis of symbolic reasoning over this model. One widely renowned example of a deliberative architecture is the *Belief-Desire-Intention* (BDI) agent [52], which was introduced in the mid-90s and still features prominently in many works today [53–55]. In contrast, reactive architectures describe agents that engage in an ongoing interaction with their environment and respond to changes that occur in a timely manner. Finally, hybrid architectures, as the name suggests, incorporate elements of both deliberative and reactive behaviour.

Rationality and its Limitations

Rationality has been distinguished across several dimensions due to its varied usage in different contexts. These correspond to different standards that one may hold for rational behaviour, or different purposes for which the term is used. What is common among different usages is the assumption that there is an agent who has preferences over a set of choices, among which a selection is to be made. Here, we outline four key distinctions that will be useful for clarifying our usage of the term. A first distinction, made by Herbert Simon, is between *normative* and *descriptive* conceptions of rationality [56]. The normative approach describes how agents "should behave in order to achieve certain goals under certain conditions" [56], and is the commonly understood usage among researchers in economics, decision theory, and AI. In contrast, the descriptive

[1] In particular, we make no ontological claims about the existence of a bright line between systems that may be classified as agents or not, though perhaps one day such a line may be found.

approach is concerned with how agents "do, in fact, behave" [56], and is more commonly adopted in the behavioural sciences and psychology.

A second distinction made by Simon was between *substantive* and *procedural* rationality [57], which shifts the emphasis to what standards one should use to ascribe the term "rational" to a given behaviour. Substantive rationality refers to behaviour that is "appropriate to the achievement of given goals within the limits imposed by given conditions and constraints" [57]. On this view, rationality is judged with respect to the outcomes produced by the behaviour in question. On the other hand, procedural rationality refers to behaviour that is the "outcome of appropriate deliberation" [57]. Here, the standard for rationality is instead defined by the *processes* that produced the observed behaviour, which allows one to take into account the costs associated with executing those processes.

Thirdly, a distinction can be made concerning the target of rationality between *instrumental* and *epistemic* objectives, as highlighted by Eliezer Yudkowsky [58]. Instrumental rationality is concerned with choosing and implementing actions in the world that successfully steer the future towards one's preferred outcomes. In contrast, epistemic rationality refers to the systematic improvement of the accuracy of one's beliefs. The relevant normative mathematical frameworks for accomplishing such objectives are decision theory and probability theory, respectively.

Finally, closely mirroring and elaborating on the normative/descriptive distinction, Gerd Gigerenzer has articulated a separation between *axiomatic* and *ecological* rationality [59]. Axiomatic rationality is imputed to an agent on the basis of its fulfilment of a specific set of axioms characterising rational behaviour in a given context, e.g., under von Neumann and Morgenstern's or Savage's axiomatisations [60, 61]. Gigerenzer conjectures that such an approach is only valid within the context of small worlds, i.e., decision making scenarios in which the agent is perfectly aware of the "exhaustive and mutually exclusive set of future states S and their consequences C" [59]. Situations where either of these sets are not known are referred to as situations of uncertainty. As an alternative to axiomatic rationality, Gigerenzer proposes ecological rationality to encompass decision-makers facing uncertainty, which insists that the context or conditions under which a behaviour is exhibited need to be taken into account to determine its degree of rationality. This allows for the possibility that, in certain contexts, the reliance of a decision-maker on "fast and frugal" heuristics can be said to be ecologically rational, given the constraints, trade-offs, and alternatives available.

Reinforcement Learning

Reinforcement learning (RL) is a paradigm for training agents by incrementally modifying their behaviours based on reward signals that measure progress towards a particular goal [62]. It is arguably the most powerful and successful method we have for training agents to perform arbitrary specific tasks. A key distinction of RL from standard supervised or unsupervised learning is the involvement of action, the separation and acknowledgement of the agent-environment boundary, and the translation of scalar reward or penalty signals into modifications of the agent's behaviour.

The core of the RL process involves a repeated feedback loop involving five primary elements: the agent, the environment, actions, observations, and rewards. This reward signal is the primary mechanism through which the environment and reward designer communicate the task's objective to the agent. The agent is designed to learn a policy that maximises the expected discounted cumulative sum of these rewards over time. The iteration of this feedback loop allows the agent to gradually acquire information about the environment and the objective, and to adjust its behaviour accordingly.

RL is intimately connected to questions around incentive design. Indeed, the problem of finding an appropriate reward function to promote particular policies and discourage others is a central question in RL research, particularly in the era of Large Language Models (LLMs) which are often post-trained using RL, e.g., using reinforcement learning from human feedback [63]. This training procedure, which is applied to models after they have been pre-trained to capture patterns in and behind large corpora of data, is what allows us to shape their outputs according to our preferences. In *multi-agent reinforcement learning*, a collection of agents is typically trained to jointly achieve a common goal, and one of the main challenges in this domain is to design effective and efficient reward structures that induce agents to successfully coordinate to achieve the common goal [64]. As we will see later in this chapter, RL has also become an increasingly important toolkit in approaches to adaptively designing incentives for agents in dynamic environments.

Active Inference and the Free Energy Principle

Active inference is a framework for designing and modelling adaptive agents based on a concept in theoretical and cognitive neuroscience known as the *Free*

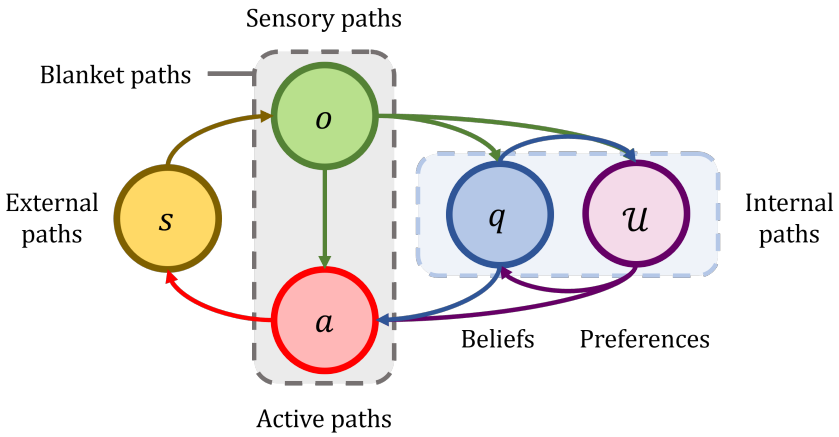

Figure 2.1: A depiction of the causal influences of different components of the agent-environment pair on each other. External paths s represent states of the world external to a system or agent under consideration. This system possesses a Markov blanket, which separates internal from external paths and can itself be divided into sensory and active paths. We further assume that internal paths consist of two distinct components: beliefs and preferences, which mutually influence each other, are influenced by sensory paths, and in turn influence active paths.

Energy Principle (FEP) [65], which articulates how stochastic dynamical systems possessing a boundary or Markov blanket can be described as engaging in a process of approximate Bayesian inference [66–69]. A *Markov blanket* refers to the set of states or trajectories of a system that mediate the influence of an agent on its environment and vice versa. This blanket can be partitioned into two kinds of paths: sensory paths, which are affected by but do not affect the environment, and active paths, which affect but are not affected by the environment. Hence, if systems possessing a Markov blanket are equipped with both 'sensory' and 'active' paths, then this means they can be seen as performing both 'perceptual inference' to optimise beliefs about external/hidden paths, as well as actively controlling their environments through a form of predictive control, i.e., *active* inference. Crucially, both these processes (external path estimation and control) rest on fulfilling the same objective: minimising the surprise associated with sensory paths. Figure 2.1 illustrates the Markov blanket partition that is typically assumed in active inference.

Despite its theoretical underpinnings in the physics of stochastic systems, active inference also can be practically applied to the design of Bayesian decision making agents. In the context of cognitive neuroscience, active inference models are used to model intelligent behaviour, e.g., humans or animals making decisions or planning under uncertainty [70]. Active inference agents are distinguished

from other Bayesian models of decision making by the fact that both goal-directed action and information-seeking behaviour follow from the single imperative of active inference agents: namely, to minimise an upper bound on surprisal (variational free energy). In this way, active inference agents solve both passive state estimation (perceptual inference) and control (active inference) as a variational inference problem. Core to the specification of any active inference model is a *world model*, which represents an agent's knowledge of its world in terms of probabilistic beliefs about the causal structure of its environment and its own actions, which ultimately cause sensory observations. Part II of this thesis is heavily inspired by and draws on many concepts in active inference.

2.3 Multi-Agent Systems

Multi-agent systems are characterised by the presence of multiple agents who interact with a shared environment and one another [51, 71]. Agents may be natural, such as biological organisms, or artificial, such as computer programs, and can both be constituted themselves by agents as well as members of multiple group agents. Different agents may have conflicting, similar, or aligned preferences giving rise to the potential for competition or cooperation. The standard mathematical framework for representing and analysing multi-agent systems is game theory, which is the central modelling approach employed throughout this thesis.

What is sometimes abstracted away in game-theoretical descriptions of agent interactions is a description of the environment in which they are situated. Models of the shared environment within which agents interact can be represented explicitly by using a relatively compact symbolic description, e.g., Reactive Modules Games (see Section 2.4) or process calculi [72, 73], or a graphical representation of the state of the system along with a set of rules for how the state is updated [74, 75]. Fundamentally, these different representations make trade-offs in the expressiveness, compactness of representation, and ease of analysis which they afford to the system modeller.

2.4 Game Theory

At its core, game theory is a mathematical discipline for modelling and studying *strategic interactions* between agents. The field was more or less established by

the pioneering work of von Neumann and Morgenstern in the 1940s [60] and has found applications in economics [76–78], political science [79–81], biology [82–84], computer science [85–89], and many other disciplines. Such widespread adoption was primarily due to the discipline's power in answering two important questions about interacting agents, which are (1) what outcomes might we *expect* to result from strategic interactions between agents? and (2) what strategies *should* agents choose in these interactions in order to conform to some notion of rationality? These two interpretations of what game theory has to offer, known as the *descriptive* and *normative* interpretations, respectively, are crucial in both explaining the past and suggesting a path forward into the future. Much of humanity's endeavours have been aimed towards these pursuits, and it is therefore no surprise that game theory has become an increasingly important tool in studying interactions between agents. For the purposes of this review, we will focus on the computational aspects of game theory and consider different classes of games in increasing levels of temporal abstraction.

Strategic Form Games

Strategic form games (SFGs) are games typically represented by a finite set of players $N = \{1, \ldots, n\}$, a set of *policies/strategies*[2] Π_i (which are generally defined as a mapping from agent states to distributions over a set of possible actions) for each player, and a set of utility functions $u_i : \Pi_1 \times \Pi_2 \times \cdots \times \Pi_n \to \mathbb{R}$ mapping joint strategies to a utility value for each player [90]. SFGs can be played in two ways, depending on whether or not players are allowed to form binding agreements with each other to play in a certain way. If this is possible, then the game is known as a *cooperative game*; otherwise, it is known as a *non-cooperative game*. Additionally, players can select their policies in an independent or correlated manner. An independent *policy profile* is a tuple $\vec{\pi} = (\pi_1, \ldots, \pi_n)$, where each $\pi_i \in \Pi_i$ is a policy of player i. A *partial policy profile* is a tuple of policies $\vec{\pi}_A = (\pi_i)_{i \in A}$ from $\vec{\pi}$ indexed by the players in A. A policy (or a policy profile) is said to be *pure* when each distribution π_i is deterministic.[3] A *correlated policy profile*

[2] In SFGs, the set of policies is usually just a finite set of alternatives to choose from. However, in temporally extended games, these can be more complex objects, such as the set of finite-state machines with output or the set of functions from observed histories to probability distributions over action trajectories.

[3] In Chapters 4-6 of this dissertation, we use σ_i to denote an agent i's strategy and Σ_i to represent their set of possible strategies, which are assumed to be deterministic and typically represented by a finite state machine with output. In Part II of this dissertation, we instead use π_i to denote a player's policy, which is taken to be probabilistic.

$\vec{\mu}$ is a probability distribution over pure policy profiles.[4] Such correlations can be implemented in many different ways, but formal definitions of this concept typically rely on the presence of a correlation device which makes use of a shared source of randomness to give players information about which actions to take in a coordinated manner. For a correlated policy profile $\vec{\mu}$, we will use $\vec{\mu}_{-i}$ to denote the corresponding marginal distribution over $\bigtimes_{j \neq i} \Pi_j$.

Solution Concepts

Turning towards the normative interpretation of game theory, the answer to the question of what outcomes rational players might be expected to choose was first studied in the context of two-player zero-sum games with the celebrated minimax theorem [60, 91] stating that the maximin $(\max_{\pi_1 \in \Pi_1} \min_{\pi_2 \in \Pi_2} u_1(\pi_1, \pi_2))$ and minimax values $(\min_{\pi_2 \in \Pi_2} \max_{\pi_1 \in \Pi_1} u_1(\pi_1, \pi_2))$ in two-player zero-sum games coincide. Later, Nash introduced the concept of a *Nash equilibrium* (NE) for n-player games, which describes a strategy profile such that no single player can strictly benefit by unilaterally deviating to another strategy [92, 93].[5]

Definition 2.2. A *Nash Equilibrium* (NE) is an *independent* policy profile $\vec{\pi}$ such that, for all $i \in N$ and $\pi_i' \in \Pi_i$,

$$u_i(\vec{\pi}) \geq u_i((\pi_i', \vec{\pi}_{-i})). \tag{2.1}$$

For a game $\mathcal{G}$, we write $\text{NE}(\mathcal{G})$ to denote the set of NE of $\mathcal{G}$. Nash's seminal work showed that every finite game has a NE in mixed strategies. Almost six decades later, it was shown that the complexity of computing a NE in two-player general-sum games is PPAD-complete [94, 95]. Even when restricting attention to only pure NE (deterministic strategies), the problem of deciding whether a game has a pure NE is known to be NP-complete [96].

The idea of a Nash equilibrium is arguably the first instance of a whole family of what are now known as *solution concepts*, which are formally specified rules to capture one's assumptions about what policy profiles are 'rational' or likely to occur in a given game. While the Nash equilibrium has become the cornerstone solution concept in game theory, many others have been proposed since then. One

[4] Note that the definitions relevant to correlated policies make sense even for probability distributions over profiles of non-deterministic policies.

[5] Such a unilateral deviation which improves a player's (expected) utility is commonly known as a *beneficial deviation*.

important implicit assumption that was recognised by Robert Aumann is that the Nash equilibrium assumes that the players choose their actions independently of each other [97]. Questioning this assumption, the concept of correlated policies was developed, leading to a generalisation of the Nash equilibrium known as the (coarse) *correlated equilibrium*.

Definition 2.3. A *Coarse Correlated Equilibrium* (CCE) is a *correlated* policy profile $\vec{\mu}$ such that, for all $i \in N$ and $\pi'_i \in \Pi_i$,

$$u_i(\vec{\mu}) \geq u_i((\pi'_i, \vec{\mu}_{-i})). \tag{2.2}$$

For a game $\mathcal{G}$, we write $\text{CCE}(\mathcal{G})$ to denote the set of CCEs of $\mathcal{G}$. One key assumption upon which this solution concept rests is that only unilateral deviations are considered. However, this does not capture the possibility presented in many real-world scenarios for agents to reach binding agreements regarding what strategies they will select, so as to coordinate towards achieving mutually improved outcomes. One common solution concept which captures this is the notion of the *core*, which considers deviations not just by single agents, but by coalitions of agents. In order to define this concept, we first have to answer the question of what it means for a group to profitably deviate from a given policy profile. Given a set of agents $A \subseteq N$ and a policy profile $\vec{\pi}$, we say that a partial policy profile $\vec{\pi}'_A$ is a *cooperative deviation* for A if for all $\vec{\pi}'_{-A} \in \Pi_{-A}$ and all $i \in A$, it holds that $(\vec{\pi}'_A, \vec{\pi}'_{-A}) \succ_i \vec{\pi}$. Intuitively, a beneficial deviation for a group of agents from a strategy profile $\vec{\pi}$ is a partial strategy profile selected by the group in which *all* agents in the group achieve a strictly greater utility than they did under $\vec{\pi}$, *no matter what strategies the remaining agents select*.[6] The *core* of a game is then the set of policy profiles that admit no cooperative deviations.

Boolean games Before proceeding to the next relevant level of temporal abstraction, we will briefly discuss a compact representation of SFGs that draws a connection between propositional logic and game theory: *Boolean games* [99]. Boolean games are non-cooperative games in which players each control a finite set of Boolean variables, and the pure strategies available to a player correspond to the set of possible assignments of truth or falsity to those variables. Preferences are captured by assigning each player a propositional logic formula that the player desires to see satisfied.

[6]The implicit assumption here is that each agent takes into account only whether they can be guaranteed to improve their utility when considering whether to form a coalition [98].

Formally, these games are represented by a tuple $B = (N, \Phi, (\Phi_i)_{i \in N}, (\gamma_i)_{i \in N})$, where $N = \{1, \ldots, n\}$ is a set of players, Φ is a set of Boolean variables, $(\Phi_i)_{i \in N}$ is a partition of Φ into n disjoint subsets, with the interpretation that each player $i \in N$ "controls" the value of the variables in Φ_i, and $(\gamma_i)_{i \in N}$ is a tuple consisting of the propositional logic goal of each player over Φ. Boolean games are played by each player $i \in N$ selecting a *valuation* $v_i \in \mathbb{B}^{|\Phi_i|}$ (i.e., an assignment of truth or falsity to each of the Boolean variables under their control), where $\mathbb{B} \coloneqq \{\top, \bot\}$ is the set of Boolean literals. The resulting *outcome* or strategy profile $\vec{v} = (v_1, \ldots, v_n)$ is then a complete assignment of truth values to all of the variables in Φ. Given an outcome $\vec{v}$, the utility of a player $i \in N$ is 1 if $\vec{v} \models \gamma_i$ and is 0 otherwise. The problem of checking whether a given outcome forms a pure-strategy NE is co-NP-complete and the problem of checking whether a Boolean game has a pure strategy NE is Σ_2^p-complete [100]. In the case of mixed strategies, the problem of checking whether a Boolean game has a mixed NE where every player is guaranteed a certain payoff is NEXPTIME-hard [101].

Extensive Form and Repeated Games

Extensive Form Games The immediate next step in introducing a temporal component into games is to consider the case where players take turns moving sequentially. This class of games is known as *extensive form games* (EFGs), which are defined by a finite set N of players, a finite tree (V, E, v_0), an ownership function assigning to each node $v \in V$ a single player $i \in N$ who solely makes a decision at the node v, an action function associating each edge $e \in E$ with an action, a utility function u_i for each player $i \in N$ assigning a utility to player i for each leaf in the tree, and a partition of the set of decision nodes of each player into a collection of information sets, which describe which nodes a player cannot distinguish between [90, 102]. If the information sets of all players are singletons, then the game is known as an EFG of *perfect information*; otherwise, it is known as an EFG of *imperfect information*. In what is considered the first known work in game theory, Zermelo studied the game of chess, which is an EFG of *perfect information*, and showed that in every game of chess, either white can force a win, black can force a win, or both sides can at least force a draw [103, 104]. This was later generalised to deterministic EFGs of perfect information and a backward induction algorithm known as Zermelo's algorithm was used to show that (1) every deterministic EFG of perfect information has a pure NE, and (2) Pure NE in such games can be computed in polynomial time [105, 106]. Building on top of this, many refinements of NE have been proposed in both the perfect and

imperfect information cases to address counterintuitive consequences of applying the concept of NE to EFGs [107, 108]. One of the key refinements of NE in the context of EFGs is the Subgame Perfect Nash Equilibrium (SPNE) [109], which requires that a strategy profile must not only be an NE in the full game, but also an NE in every subgame (i.e., EFGs starting from different internal nodes of the tree). In the case of deterministic EFGs with perfect information, an SPNE always exists, and Zermelo's algorithm can again be applied to compute one in polynomial time.

Repeated Games Instead of considering a single game in which players move sequentially, another approach is to consider the same SFG being played repeatedly. The study of repeated games can be categorised into two parts: finitely repeated games [110] and infinitely repeated games. The case of finitely repeated games is fairly straightforward to analyse as these are a type of EFG, so techniques developed in the EFG literature can be straightforwardly applied to these games. In the case of infinitely repeated games, this is not possible in general, and it becomes unclear how one should represent strategies and calculate utilities. In these games, strategies can be encoded by *Moore machines* (finite-state machines with state-dependent outputs) so that players' decision making can be conditional on what other players have played in previous rounds of the game. In the context of utilities, one approach is to consider the limit-inferior of means of the payoff accrued in each round of the game [111] and another is to use geometric discounting on utilities obtained as the game progresses [112]. In the limit-average payoff case, the celebrated Nash folk theorem states that in an infinite game, every outcome in which all players achieve at least their security value (the best payoff they can guarantee no matter what other players do) can be achieved in a NE. Moreover, algorithms have been developed to compute Nash equilibria in infinitely-repeated two-player games with limit-average payoffs in polynomial time [111, 113].

Another way that repeated games can be played is by assuming that players *learn* over successive rounds of play. In this setting, policies are typically taken to be simple probability distributions over the actions available to players in the game, and these distributions are updated based on the observed histories of play according to some learning rule [114, 115].

Iterated Boolean Games The idea of infinitely repeating an SFG has also been applied to Boolean games, in what is known as *Iterated Boolean Games* (IBGs) [116]. In the same way that utilities were defined for one-shot Boolean games through the satisfaction of propositional goal formulae, players' preferences in IBGs are defined with respect to the satisfaction of an agent's goal formula, which is expressed as an LTL formula over the sequence of labels induced by a *run* of the game. Runs are themselves induced by strategy profiles $\vec{\sigma} = (\sigma_1, \ldots, \sigma_n)$ chosen by the agents at the beginning of the game, which are encoded as Moore machines that determine what truth values to assign to each variable under an agent's control at each timestep. Some of the main decision problems introduced concerning IBGs are:

> IBG MODEL-CHECKING:
> *Given*: IBG $\mathcal{G}$, strategy profile $\vec{\sigma}$, LTL formulae φ.
> *Question*: Is it the case that $\rho(\vec{\sigma}) \models \varphi$ (i.e., the run induced by $\vec{\sigma}$ satisfies φ)?
>
> IBG MEMBERSHIP:
> *Given*: IBG $\mathcal{G}$, strategy profile $\vec{\sigma}$.
> *Question*: Is $\vec{\sigma}$ a Nash equilibrium of $\mathcal{G}$?
>
> IBG NON-EMPTINESS:
> *Given*: IBG $\mathcal{G}$.
> *Question*: Is there a pure-strategy NE in $\mathcal{G}$?

It is shown in [116] that IBG MODEL-CHECKING is in PSPACE, IBG MEMBERSHIP is PSPACE-complete, and IBG NON-EMPTINESS is 2EXPTIME-complete. Other important decision problems in the setting of IBGs will be discussed under the Rational Verification heading in Section 2.4.

Infinite Games

In the most general two-player definition, a (two-player) infinite game $G(X)$ on an alphabet A is simply a subset $X \subset A^\omega$ of infinite words on A [33, 117]. A play can be represented as an element $x \in A^\omega$, and player 1 is said to win the game if $x \in X$ or otherwise, player 2 wins the game. Determinacy results (guarantees that some player has a winning strategy) have been proven for very broad classes of games, but this is not the case for all infinite games [33].

Turn-based Games Whilst the definition of infinite games above abstracts away all of the unnecessary features that we typically associate with games, a more natural interpretation model is that of infinite games played on graphs (IGGs) [118]. An IGG consists of a directed bipartite graph where every vertex has an outgoing edge. The partition of the graph's nodes allows us to assign ownership over the nodes to players 1 and 2 in the same way as is defined in EFGs. The game starts off in one of the nodes and players take turns choosing the next node for the game to transition to, where allowable moves are defined by the directed edges. In some of the original work on IGGs, Büchi and Landweber showed that all infinite games definable by an ω-regular specification are determined, that is, some player has a *finite* strategy that guarantees their victory [119].

The value of such games is that any infinite game where the set X of words for which player 1 wins defines an ω-regular language can be viewed as an IGG whose game graph is the ω-automaton that recognises X [33]. Thus, reasoning about winning strategies in games has a close connection with model-checking ω-regular properties. Additionally, IGGs can be used to derive algorithms and prove complexity results in multiplayer (concurrent) games [120, 121].

The original formulation of IGGs has also been extended in several directions to allow for multiple players and other objectives apart from ω-regular sets. In the case of multiplayer turn-based IGGs where all players have ω-regular winning conditions, the problem of checking whether there exists a mixed NE where each player i's probability of winning lies within some range (x_i, y_i) is NP-complete for Streett, parity, or co-Büchi objectives and is in P for Büchi objectives [122, 123]. Another approach to modelling objectives in IGGs is to assign rewards to different transitions or states, and then *quantitative objectives* over infinite sequences of observed rewards can be defined, such as maximising one's limit-average payoff. Such an objective is commonly referred to as a *mean-payoff* (MP) objective. These have the elegant property that memoryless strategies (strategies that only depend on the current state) are sufficient for optimal play in two-player, turn-based mean-payoff games [124]. Later, it was shown that the complexity of computing such optimal strategies is in NP $\cap$ coNP [125]. Other approaches combine qualitative (ω-regular) objectives and quantitative (mean-payoff) objectives typically by assuming that agents have a lexicographic preference structure – agents will prefer to satisfy their qualitative objective no matter what their mean-payoff, but *ceteris paribus*, agents will prefer a higher mean-payoff [126–130].

Concurrent Games Traditionally, models of infinite games played on graphs have players take turns in deciding what action to take. In a concurrent game, play continues for an infinite number of rounds, where at each round, all players choose an action simultaneously and independently in each node of the game graph [38, 131], and the next state of the game is determined by the joint actions $\vec{a} = (a_1, \ldots, a_n)$ of all the players.[7] Concurrent games are a strict generalisation of IGGs and SFGs in the sense that any of these games can be described by a concurrent game. As in the case of turn-based games, concurrent two-player games are not determined for arbitrary winning conditions [132]. However, it is shown in [132] that concurrent games satisfying specific properties (race-free, bounded-concurrent, and Borel winning conditions) are determined.

In more general multiplayer concurrent games, extensive work has been done in determining the complexity of computing pure NE (see [120]) under a variety of different ω-regular winning conditions. In the case of mean-payoff objectives, the problem of checking whether a NE exists with limit-average payoffs lying between given bounds is NP-complete for pure strategies, in PSPACE for (mixed) memoryless strategies, and undecidable for mixed arbitrary strategies [133].

Stochastic Games In the game models considered so far, transitions are mostly assumed to be deterministic – the next state of the game (if there is one) depends solely on the actions of the players. A very obvious extension of this is to relax the deterministic assumption by considering *probabilistic* games, which was first introduced by Shapley as *stochastic games* in the two-player setting [134] and later extended to the general n-player setting [135–137]. In stochastic games, the best that one can do in general is to maximise the probability of winning, as opposed to ensuring that one can win. Decision problems involving computing the set of states in which a given player has a strategy to guarantee victory (with probability 1) are known as *qualitative problems*, whereas problems which involve computing the highest probability of satisfying their objective from each state are known as *quantitative problems* [138, 139].

Rational Verification

Having canvassed the landscape of relevant areas in the model checking and game theory literature, we are now in a good position to discuss *rational verifica-*

[7] A formal definition of concurrent games is given in Chapter 4.

tion [140, 141], which lies at the intersection of these two fields. When one designs or encounters a multi-agent system, one of the first questions to ask is: *what are the possible outcomes that this system could give rise to?* This question is hardly straightforward to answer and fundamentally depends on what behaviours and objectives the system's components possess. In the most general case, one may decide not to make *any* assumptions of the agents within the system, in which case there is a whole wealth of techniques from the model checking literature outlined in Section 2.1 to draw upon. As we have seen, in traditional approaches to formal verification, one is often interested in determining whether a given property, expressed as a temporal logic formula, holds on some or all of the possible executions of a system [21]. However, supposing that one has some insight into the objectives of the agents in the system and assumes them to be rational, a simplifying assumption would be to only consider the set of outcomes that could arise in a game-theoretic equilibrium. In the spirit of model checking, one can then ask whether a given temporal logic formula φ is satisfied in some or all game-theoretic equilibria of a game. These questions lie at the heart of the rational verification paradigm and are known as forms of *equilibrium checking* or *rational verification* problems. More formally, the key decision problems are:

E-NASH:
Given: Game $\mathcal{G}$, temporal logic formula φ.
Question: Does there exist a strategy profile $\vec{\sigma} \in \text{NE}(\mathcal{G})$ such that $\rho(\vec{\sigma}) \models \varphi$?

A-NASH:
Given: Game $\mathcal{G}$, temporal logic formula φ.
Question: Is it the case that for all strategy profiles $\vec{\sigma} \in \text{NE}(\mathcal{G})$ we have $\rho(\vec{\sigma}) \models \varphi$?

In other words, *Rational verification* refers to the set of problems of establishing which temporal logic properties will be satisfied by a multi-agent system, under the assumption that agents in the system choose strategies that form a game-theoretic equilibrium [121, 141, 142]. Thus, rational verification enables us to verify which desirable and undesirable behaviours could arise in a system through individually rational choices.

Other related decision problems include the MEMBERSHIP problem (deciding whether a given strategy profile is a NE) and the NON-EMPTINESS problem (deciding whether a given game has a NE). The former is closely related to the

standard model checking problem and the latter is a special case of the E-NASH problem where the formula φ is simply $\top$. Also note that the E- and A-Nash problems are dual to each other, in the sense that the A-Nash question can be solved by simply asking whether E-Nash is false on the formula $\neg\varphi$.

These two central questions have been studied in a range of different contexts. Here, we will briefly discuss some of the game models, temporal logic specifications and solution concepts studied in the rational verification literature.

One of the first game models to be studied in rational verification was IBGs [116, 143, 144], which we introduced in section 2.4. Because every infinite run of an IBG consists of an infinite sequence of valuations over the variables in Φ, it is an eminently reasonable idea to define temporal logic formulae over these valuations. Taking LTL goals and specifications as our starting point, it was shown in [116] that the E-NASH problem can be reduced to a closely related problem known as *rational synthesis* [142], which is concerned with the synthesis of a strategy profile in a game similar to IBGs such that the resulting run is a Nash equilibrium and satisfies a particular LTL formula φ. This problem is known to be 2EXPTIME-complete [142], giving us an upper bound on the complexity of the E-NASH problem. Next, it was shown that a related problem known as the *LTL realisability* problem [36], which is also 2EXPTIME-complete, can be reduced to the IBG NON-EMPTINESS problem. Recalling that the NON-EMPTINESS problem is a special case of E-NASH with the specification $\varphi = \top$, this gives us 2EXPTIME-hardness for the E-NASH problem as well.

Shortly after this, another expressive yet succinct model to study concurrent games known as *Reactive Modules Games* (RMG) was studied, which models concurrent game structures using a simplified version of the Reactive Modules Language [145] known as the *Simple Reactive Modules Language* (SRML) [16, 146]. The idea behind RMGs is that each player is represented in SRML as a *module* in which a set of Boolean variables is initialised in some way and then subsequently controlled by the player. However, players are not simply allowed to choose any valuation they like for the variables under their control – a module also comes with a set of *guarded commands*, which specify conditions (as propositional logic formulae) under which a particular setting of the variables is permitted. In the same manner as IBGs, infinite runs of an RMG define an infinite sequence of valuations over the union of the set of variables controlled by each player and hence, temporal logic formulae for players' goals and for specifications can be defined in terms of these. One variant of RMGs studied in [147] are known as *Weighted Reactive Modules Games* (WRMG), in which each player is additionally

Game	Player Goals	Spec.	x	E-x	A-x	Refs.
IBG	LTL	LTL	NASH	2EXP-c	2EXP-c	[116]
IBG	Prop. Safety	LTL	NASH	PSPACE-c	PSPACE-c	[116]
IBG_F	LDL_F	LDL_F	NASH	2EXP	2EXP	[144]
RMG	LTL	LTL	NASH	2EXP-c	2EXP-c	[16]
RMG	CTL	CTL	NASH	2EXP-h	2EXP-h	[16]
RMG_F	LDL_F	LDL_F	NASH	2EXP-c	2EXP-c	[148]
WRMG	MP	ω-regular	NASH	NEXP	NEXP	[147]
CGS	LTL	LTL	NASH	2EXP-c	2EXP-c	[121, 149]
CGS	GR(1)	LTL	NASH	PSPACE-c	PSPACE	[150]
CGS	GR(1)	GR(1)	NASH	FPT	FPT	[150]
CGS	MP	LTL	NASH	PSPACE-c	PSPACE	[150]
CGS	MP	GR(1)	NASH	NP-c	coNP	[150]
CGS	MP	ω-regular	NASH	NP-c	co-NP-c	[147]
CGS	LTL	LTL	CORE	2EXP-c	2EXP-c	[151]

Table 2.1: Summary of complexity results for various equilibrium checking decision problems. Notation used is as follows: **E/A-x** = Existential or universal rational verification problem under solution concept **x**; IBG = Iterated Boolean Game; IBG_F = IBGs with LDL_F goals; RMG = Reactive Modules Games; WRMG = Weighted Reactive Modules Game; CGS = Concurrent Game Structure; Prop. Safety = Propositional Safety LTL formula; LDL_F = Linear Dynamic Logic over finite traces; 2EXP = 2EXPTIME; NEXP = Nondeterministic exponential time; FPT = Fixed-Parameter Tractable; -c = -complete; -h = -hard.

assigned a weight function mapping commands to integers. Utilities are then defined as the mean-payoff achieved on the infinite run.

In the context of *Concurrent Game Structures* (CGS), we have already seen how labelling functions mapping states to propositional variable assignments in a concurrent game can give rise to infinite sequences of valuations over which temporal logic formulae can be defined. A natural extension of this is known as *Concurrent Stochastic Games* (CSGs), in which the state transition function is probabilistic. CSGs have become increasingly favoured both in the multi-agent reinforcement learning literature [152–154] and in the probabilistic model checking literature [155–160], but with a focus mainly on players with reward-based utilities.

Most of the temporal logic formulae used for players' goals and specifications to be checked have been covered, including LTL and CTL formulae and ω-regular specifications. Other types of logical specifications have also been considered including Linear Dynamic Logic over finite traces (LDL_F) which is equivalent to monadic second-order logic [144], propositional safety formulae

which are a subset of LTL taking on the form $\mathbf{G}\varphi$ (where φ is a propositional logic formula) and GR(1) formulae which are also a subset of LTL taking on the form $(\mathbf{GF}\psi_1 \wedge \ldots \wedge \mathbf{GF}\psi_m) \rightarrow (\mathbf{GF}\varphi_1 \wedge \ldots \wedge \mathbf{GF}\varphi_n)$, where each subformula ψ_i and φ_i is a propositional logic formula [150]. Finally, the introduction of quantitative agent objectives has also been considered, such as MP objectives [147] and lexicographic LTL and MP (Lex(LTL,mp)) objectives [129]. In the latter case, the NON-EMPTINESS problem was studied for the set of Finite-State ε-Nash Equilibria (FSNE$^{\varepsilon}$), because NE may require infinite memory strategies in these settings [126].

Finally, an extension to the setting of cooperative game theory has also been studied [151], wherein players are allowed to form binding agreements to cooperate with each other and play an agreed-upon set of strategies.Table 2.1 presents a summary of the complexity for the primary rational verification decision problems outlined above.

2.5 The Study of Incentives

The final area that we will cover in this chapter is of course the central topic of investigation in this thesis: incentives.[8] Broadly speaking, the study of incentives can be broken down into questions about: (1) what incentives are, (2) what effect they have, (3) how we can identify them, and (4) how they can be designed or reshaped to achieve specific outcomes.

Incentive Models

Incentive models can vary along several axes, one of the most fundamental ones being *who does the incentivising*? The first and most common option is for an external agent, commonly known as a *principal*, to make observations about the behaviours of the agents and use their knowledge to design incentive schemes to align the agents' local incentives with their own objectives [161]. Alternatively, agents may also be able to incentivise each other on the basis of observing others' actions [162, 163].

Once it is clear who is doing the incentivisation, the next decision to be made is what form the incentives can take. This can be guided by a taxonomy

[8] For an excellent complementary review of the study of incentives, we refer the reader to [11].

of incentive schemes, e.g., by breaking down incentives into trust-based and trade-based incentives [164]. Trust-based incentives seek to make use of indirect coordination mechanisms by having agents become members of a collective or having reputation-based mechanisms that encourage them to contribute towards the desired global objective. On the other hand, trade-based incentives are a more direct appeal to agents' material interests, including offers of direct exchanges of goods and services or deferred actions in return.

Mechanism Design

Incentive design in the field of economics is also broadly known as *mechanism design* [165, 166]. Broadly speaking, mechanism design is the study of how game structures and payoffs can be designed for players to elicit desirable outcomes under the assumption that agents are rational. Over the decades since its inception, the field has coalesced around a few key desirable properties that mechanisms ought to have, which are (non-exhaustively) summarised in the following list:

1. **Incentive Compatibility**: A mechanism is incentive-compatible if every participant achieves the best outcome for themselves by acting truthfully according to their private information (e.g., bidding their true value in an auction).

2. **Individual Rationality**: A mechanism is individually rational if no participant is worse off by participating in the mechanism than they would be by not participating at all. This property guarantees that agents have a rational incentive to participate in the mechanism.

3. **Strategy-proofness**: This is a stronger form of incentive compatibility. A mechanism is strategy-proof if telling the truth is a participant's best strategy, regardless of what other participants do. It is a very powerful and desirable property because it removes strategic uncertainty for the agents.

4. **Efficiency**: This property relates to how well a mechanism allocates resources to maximise a certain objective. The most common form is Pareto efficiency, where no other outcome exists that would make at least one agent better off without making any other agent worse off.

5. **Equilibrium**: An equilibrium is a strategy profile in the game induced by the mechanism where no single participant or group of participants stand to gain by 'deviating', given the strategies of all other participants. A good mechanism should have desirable equilibrium outcomes.

6. **Computational Tractability**: This means that the outcome of the mechanism can be computed by the designer in a reasonable amount of time and using a reasonable amount of resources. If a mechanism produces a great outcome but takes centuries to calculate, it is not practical.

7. **Robustness**: A mechanism is robust if it performs well even when the designer's assumptions about the environment (e.g., the agents' valuation distributions or strategic capabilities) are not perfectly accurate.

8. **Budget Balance**: This property ensures that the mechanism does not run at a loss or generate an unintended surplus, and typically applies in the setting of taxation and public spending. In a weakly budget-balanced mechanism, the total payments collected are greater than or equal to the total payments made. In a strongly budget-balanced mechanism, the payments collected exactly equal the payments made.

9. **Simplicity**: A simple mechanism is easy for participants to understand, use, and decide what to do under. If the rules are too complex or the optimal/equilibrium strategies are too difficult to figure out, agents may make mistakes or be unwilling to participate, undermining the mechanism's other desirable properties.

10. **Fairness**: This property relates to the equitable treatment of participants. This can be defined in many ways, such as ensuring that no agent envies the allocation of another (envy-freeness) or that all agents are treated 'equally', according to some definition of equality.

11. **Privacy**: This property ensures that the mechanism protects the private information of its participants (e.g., their preferences or actions) from being revealed to other participants or the public.

12. **Verifiability**: From the perspective of participants, a mechanism is verifiable if participants (or an external auditor) can confirm that the outcome was determined correctly according to the stated rules. This builds trust in the system. From the perspective of the mechanism designer, verifiability may refer to the property that the intended outcomes that the mechanism was designed to achieve are indeed the ones that are actually achieved.

13. **Welfare Maximisation**: This is a specific type of efficiency where the primary goal is to maximise the sum of all participants' utilities or values. It is also known as social welfare maximisation, and has sometimes been criticised for making the strong assumption that interpersonal utilities are directly comparable.

14. **Designer Preference Satisfaction**: This is a broad category referring to the specific goals of the mechanism's designer, which might differ from general welfare maximisation. For example, a government allocating radio spectrum might prioritise public good over profit, while a company running an auction would prioritise maximising revenue.

The Principal-Agent Problem

One aspect of mechanism design problems that has been extensively studied is that of information asymmetry between a principal and an agent, which is broadly referred to as the *principal-agent problem*. The principal-agent problem is a classic problem in economics. It arises when an agent – the *principal* – would like to contract out a task to an agent, but has imperfect information about the agent that would allow them to ensure that the task has been carried out satisfactorily and the appropriate wages have been paid to the agent [161, 167–170]. The two main classes of principal-agent problems are *adverse selection* [171] and *moral hazard* problems [167].

In an adverse selection problem, the principal only has partial information about the agent's type, which influences the quality of their output. For example, when an employer hires a new employee, they only have partial information about the skills and aptitude of the employee. Their skills and aptitude will then play a part in the performance of the delegated task. In the moral hazard problem, the principal knows the agents' types but is unable to fully observe the *actions* that the agents took. Instead, they are only able to observe a noisy signal of the true action taken. In particular, the principal may be unable to directly verify that the terms of the contract have been respected. For example, suppose we want to hire a cleaner to ensure that a particular building is kept clean, although we are only able to inspect the building at infrequent intervals: if the cleaner desires to minimise effort, why would they then not simply clean up only when an inspection is imminent?

Investigations of moral hazard problems were initially motivated by the desire to better understand the relationship between risk and incentives in insurance contracts [172, 173]. Early models of moral hazard sought to study how risks can be shared optimally between principal and agent using contracts [161, 167, 174, 175].[9] Extensions of these models to settings with *multiple agents* and a single principal were also developed, in which the presence of competition between agents must be taken into account [176–178]. More recently, computational approaches to contract design have extended the original models in various ways and have taken an algorithmic lens on computing optimal contracts [179–185]. Multi-agent contracts have also been studied from this perspective, which introduces the challenge of reasoning about combinatorially large action spaces [186–190].

The key challenge in both adverse selection and moral hazard problems is to design a contract for the agent that aligns the incentives of the agent with those of the principal, so that the principal can have confidence that the agent will be rationally motivated to expedite the task at hand, even if the principal is unable to directly verify that this is indeed the case. In other words, the task of the principal is to design a contract that maps observed outcomes to payments for the agent to maximise the principal's utility, which is typically an increasing function in the 'quality' of the observation and decreasing in the expected payment made to the agent. This setting has also been generalised to cases with multiple agents, but such works focus primarily on quantitative utility functions [176, 178, 191–193]. More recent work in the area of contract design considers dynamic, temporally-extended interactions between principals and agents [194, 195].

Computational Approaches to Incentive Design

The classical approach to approaching problems of mechanism design is for a designated person (or group of people) to identify: (1) the set of possible participants, (2) the set of possible or feasible mechanisms that could be implemented, (3) an appropriate subset of the properties that a satisfactory mechanism should satisfy, and (4) a mechanism that provably satisfies the identified properties. In general, it may be proven that no mechanism exists in a given context that satisfies some combination of criteria, that there is a unique one that does so, or that multiple mechanisms satisfy the criteria, in which case additional criteria may be included to decide which one to use. However, the design of mechanisms by hand can be a painstaking process, and formally proving

[9] See [169] for an overview of moral hazard studies.

their suitability can impose significant additional burdens on the designer. To ameliorate these issues, recent computational approaches to mechanism design have proposed the application of formal methods to the verification and synthesis of mechanisms [196–200], under the broad heading of *automated mechanism design*. The idea behind these approaches is to model the mechanism design problem using computational modelling languages and then use formal methods to automatically synthesise and verify mechanisms that satisfy the desired properties, or to identify that no such mechanism exists given the specified constraints. One of the central challenges with these approaches is the computational complexity of the problem, which is increasingly exacerbated as the number of participants, mechanisms, and properties to consider increases. Another related but distinct area, known as *algorithmic mechanism design*, studies the use of algorithms and mechanism design to incentivise self-interested agents to approximately solve computational problems such as task scheduling and the shortest path problem [201]. This has since been expanded to incorporate topics such as mechanism design without money, combinatorial auctions, cost sharing, and questions around efficiency, distributedness, profit maximisation, and online algorithms in mechanism design [202].

Game Modifications

An implicit assumption in many approaches to mechanism design is that the designer has full control over the interaction structure between the agents. In many settings however, the underlying structure that describes the agents' interactions with the environment and other agents is already known or partially given. While those systems may not comply with the designer's objective, a complete redesign is not always feasible. For instance, a country may enact environmental legislation but fail to achieve its sustainability objectives because the incentives of private individuals and companies are not aligned with these higher-level objectives. In such cases, policymakers may, for example, resort to establishing taxes based on a company's pollution rate, or subsidising public transportation fees to incentivise its use. This can be more generally described as the designer making *modifications to an existing game structure* [203].

The underlying structure can take many forms, such as a *Markov Decision Process* (MDP), where the incentive designer can impose additional costs or rewards on an agent's actions to influence an agent's optimal policy so that they are motivated to satisfy the designer's objective [204, 205]. In multi-agent settings,

Regulatory system	Prescriptive	Prohibitive	Compliance	Violation
Recommendation	✓		Reward	No penalty
Dissuasion		✓	Reward	No penalty
Obligation	✓		No reward	Penalty
Prohibition		✓	No reward	Penalty
Incentive system	✓	✓	Reward	Penalty

Table 2.2: Different classifications of regulatory systems by the consequences of compliance and violation.

the structure is some form of game, and as alluded to earlier, the principal is typically interested in making sure that *equilibrium* outcomes of the game satisfy desirable properties. For example, the specification repair problem asks how the objectives of players can be modified so that the game admits a stable outcome such as a Nash equilibrium [206]. Beyond merely stabilising a game however, the principal usually aims to ensure that the resulting equilibria have desirable properties, such as being Pareto-optimal or socially desirable [207–209].

Normative Systems In the multi-agent systems literature, the term *normative system* typically refers to mechanisms that aim to regulate the behaviours of agents by means of specifying which actions players are allowed or recommended to take at any given point in time. This stands in contrast to other approaches to incentivising agents, which may be based on the state of the game/environment or the decision-making processes (i.e., policies) themselves, as opposed to the actions taken.

Approaches to studying the design of norms typically centre around the use of logic for describing norms [210, 211], the synthesis of dynamic norms that update the set of actions available to agents [212–216] or modify the sanctions imposed for the violation of different norms [217, 218].

Several methods of categorising the different approaches to implementing norms have been proposed in the literature [219–223]. In particular, [222] define a useful categorisation of what they term *regulative* norms, in which actions taken by the players or states of the system are prescribed, permitted, or prohibited. According to the classification of a given norm, the regulative agent is able to impose penalties or rewards based on the compliance or violation of such norms. The term *recommendation norm* is introduced to describe actions which are incentivised by rewarding agents for compliance. Other approaches, which

have been more common in the literature, include obligations or prohibitions which punish violating agents with a penalty, or permissions which involve the (temporary) removal of punishments and rewards associated with norms. However, to fully capture the range of possible incentives available, a regulatory agent should be able to impose both prescriptive and prohibitive norms and apply both rewards and penalties for compliance or violation. We refer to such structures as *incentive systems* to highlight this flexible capability which is distinguished from the previously mentioned classifications. Table 2.2 summarises the different classifications of regulatory systems by the consequences of compliance and violation.

Equilibrium Design Another class of computational approaches to incentive design in multi-agent settings is known as *equilibrium design* [207–209], which aims to understand how incentives can be designed to shape the game theoretic equilibria that arise in a multi-agent system so that the resulting set of equilibria in the game satisfies a certain desirability criterion, specified as a logical formula or quantitative specification. Early work on this problem began with the study of taxation schemes in one-shot Boolean games [224], where agents are assumed to have lexicographic qualitative and quantitative objectives. In this setting, agents' primary objectives are to satisfy their logical goals, given as a propositional formula. Additionally, actions that players take are associated with costs, and agents would secondarily prefer to minimise the costs they incur. Subsequent work on this model characterised the sets of *soft* and *hard* equilibria of a game [225], which are the sets of NE in the costless version of the game that can and cannot be eliminated by the introduction of a taxation scheme. In general, however, it is insufficient to eliminate a single undesirable NE from a game as there may be several such equilibria. This motivated subsequent study which characterised when *sets* of equilibria can be eliminated or induced in a game [226].

Causal Approaches to Understanding Incentives

Understanding what incentives are present in the context of decision making is a crucial first step towards identifying and mitigating potential failure modes that might arise. This can be achieved by understanding the causal structure of the system, and identifying the variables that are most likely to influence the decision-maker's behaviour. To this end, *causal influence diagrams* (CIDs) and *structural causal influence models* (SCIMs) have been demonstrated to be an

invaluable tool for representing and reasoning about the factors influencing the decision making of agents, allowing one to categorise and identify the presence of different kinds of incentives that may be present in a system [227, 228].

Initial work in this direction sought to understand which nodes in a CID an agent has an incentive to observe and control by measuring the values of information and control, respectively [227]. The *value of information* quantifies the degree to which an agent is able to improve their expected utility by observing the value of a particular node in the CID, while the *value of control* quantifies the degree to which an agent is able to improve their expected utility by controlling the distribution of a particular variable in the CID. A refinement to the value of control known as an *Instrumental Control Incentive* was later introduced to capture which nodes an agent can influence with their decisions [229]. These models were also used to propose a normative framework for removing instrumental control incentives over 'delicate' states of a system through the use of *Path-Specific Objectives*, which capture the causal effect of an agent's actions on its objective along a given path under the 'natural' distribution of other variables [230]. Additionally, the concept of a *Response Incentive* was proposed to capture which changes in the environment affect an optimal policy of the agent [229]. Recognising the subtleties between intentional and unintentional incentives, Carey et al. later generalised the instrumental control incentive by introducing the concept of an *Impact Incentive* to capture which nodes an agent can influence intentionally or otherwise [231]. The power of CIDs in formalising such concepts is that they allow one to define precise graphical criteria to identify the presence of such incentives.

Incentive Design For and By Learning Agents

The proliferation of techniques in machine learning to solve large-scale, high-dimensional optimisation problems has naturally found use in the field of incentive design. These developments have given rise to an important set of questions and challenges in the field, and manifest in two important ways. Firstly, how can learning algorithms be used to adaptively design and modify incentives for agents? Secondly, how can incentives be designed specifically for learning agents?

A great deal of work has been devoted to studying the first question, typically by formulating incentive design as a learning problem where the designer learns through interaction with an agent or multiple agents to modify the incentives that they provide in order to optimise a particular objective or to approximately satisfy

certain properties [162, 232–239]. In these approaches, it is usually assumed that the agents will predictably act to optimise their utility function or play some form of game theoretic equilibrium.

The second question has also received a significant amount of attention, typically by specifying some learning algorithm that agents are assumed to use to update their behaviours and then formulating the incentive design problem taking this into account [240–244]. In these approaches, one is typically interested in identifying theoretically optimal incentives in the presence of learning agents.

Given these two approaches, it is natural to wonder about settings where both the agent(s) *and* the incentive designer are learning, and indeed, this space has been richly explored particularly in recent years [233, 245–256]. A common technique in these settings is to view them as bilevel optimisation problems, which involve an outer optimisation problem of incentive design subject to an inner optimisation problem of agent learning.

Other work in this area involves the study of settings where agents learn to incentivise each other [162, 163], and where an agent aims to use incentives to learn about another agent [257] or the payoff structure of a game [258].

This remains an active area of research, and though the work presented in this thesis barely ventures into this domain, there is a wide space of opportunities to expand the work presented herein to incorporate learning agents. In particular, the application of machine learning and stochastic approximation techniques is likely to lead to more scalable methods of incentive design, though potentially at the cost of formal guarantees about the correctness or optimality of the resulting incentives. However, this may well be worth the trade-off in many cases, as Part II of this dissertation will highlight the inherent limitations of the strict rationality assumption in the context of incentive design.

2.6 Discussion

To summarise, this chapter has provided an overview of the most pertinent concepts, formalisms, and areas of inquiry to the study of incentives and multi-agent systems. We began with a brief summary of the vast literature on formal verification and synthesis, covering model checking, concurrent systems, and logic. We then discussed some of the key ideas related to models of agency, including the concepts of (bounded) rationality, reinforcement learning, and

active inference. We then surveyed the relevant literature on multi-agent systems, particularly from the perspective of game theory and rational verification. Finally, we introduced some of the central concepts in the study of incentives, touching on some of the core ideas from the economics literature, but mostly focusing on computational perspectives.

One of the major gaps that the first part of this thesis aims to address is the application of tools from formal synthesis and verification to questions of incentive design. Subsequently, though interest has been rising in the concept of bounded rationality, many foundational questions around how to best model it remain unanswered, and we attempt to make progress in this direction in Part II of this thesis. The implications of bounded rationality for the study of incentives have received relatively scarcer attention compared to standard conceptions of rationality and basic reinforcement learning models of agents, and it is this area that we have aimed to contribute to as well as sketching out the directions for future investigation.

Part I

Incentives for Rational Agents

3

Principal-Agent Boolean Games

> *It is the interest of every man to live as much at his ease as he can; and if his [earnings] are to be precisely the same, whether he does, or does not perform some very laborious duty, it is certainly his interest ... either to neglect it altogether, or ... to perform it in [as] careless and slovenly a manner as that authority will permit.*
>
> Adam Smith, *The Wealth of Nations*, (p. 760) [14]

In this chapter, we introduce and study a computational version of the *principal-agent problem* – a classic problem in economics that arises when a principal desires to contract an agent to carry out some task, but has incomplete information about the agent or their subsequent actions. The key challenge in this setting is for the principal to design a contract for the agent such that the agent's preferences are then aligned with those of the principal. We study this problem using a variation of Boolean games [99], where multiple players each choose valuations for Boolean variables under their control, seeking the satisfaction of a personal goal, given as a Boolean logic formula. In the variant of Boolean games that we adapt in this chapter, secondary preferences are defined by costs associated with assignments of truth or falsity – a player first seeks to satisfy their goal, and secondarily seeks to minimise costs [224].

In this setting, the principal can only observe some subset of these variables, and the principal chooses a contract which rewards players on the basis of the

assignments they make for the variables that are observable to the principal. The principal's challenge is to design a contract so that, firstly, the principal's goal is achieved in some or all Nash equilibrium choices, and secondly, that the principal is able to verify that their goal is satisfied. In this chapter, we formally define this problem and completely characterise the computational complexity of the most relevant decision problems associated with it. More specifically, we derive complexity results for two classes of problems: 1) can the principal formally verify whether or not their objective, given as a propositional logic formula over the variables in the game, is satisfied on some/all observations, and 2) can the principal design a contract such that their objective is satisfied on some/all possible observations?

Classical work in the area of moral hazard in teams fails to take into account the potential lexicographic or qualitative nature of agent preferences. To our knowledge, our work is the first to study multi-agent moral hazard problems in a game where agents have lexicographic qualitative and quantitative preferences. In our model, a principal has a goal they desire to see accomplished, expressed, as with the agents' goals, as a propositional formula. The principal chooses a contract that rewards individual agents on the basis of their observable variables. The fundamental question we ask is whether it is possible for the principal to design a contract such that, if agents rationally make choices (i.e., make choices that form a game-theoretic equilibrium), then, firstly, the principal's goal will be accomplished, and secondly, the principal can formally and automatically *verify* that the goal is accomplished, under the assumption that agents have acted rationally.

Common among most models of moral hazard is the assumption that agent preferences are quantitative, and partial observability is primarily modelled as a noisy signal of the agents' actions, which are often represented as a quantitative degree of effort [167]. In contrast, our model allows agents to have qualitative goals specified as propositional formulae, which are prioritised over their quantitative objectives and constrain the agents' decision making to a discrete action space. Furthermore, the principal's objective in our model is a propositional logic formula, as opposed to a function of the observation signal.

An example of a setting where this may be more appropriate is a scenario in which a set of tasks are divided among a group of autonomous agents, each of whom is capable of reliably completing their assigned tasks. For instance, these tasks could be maths or programming problems where a definite binary outcome can be associated with the successful completion of the task. In this scenario each

task might be associated with a cost of completion corresponding to the amount of compute required to complete the task, and the agents must decide which tasks to complete based on their objectives and the contract payments that are offered under the principal's limited observational abilities.

3.1 Related Work

Computational Approaches to Mechanism Design There are computational approaches to mechanism design in which approximation algorithms [201], program synthesis [199], and satisfiability checking [196, 197], among others, have been proposed to automatically construct optimal mechanisms. These works assume a stronger degree of control by the principal over the agents' interactions and their utilities. Also closely related to our model is the *combinatorial contracts* model [190, 259], where a principal designs a contract for a set of agents who choose a subset of actions to take, which can be modelled using the Boolean games framework. Also similarly to our model, actions are associated with costs, which are assumed to be additive across actions. In contrast, our model allows for the cost function to be an arbitrary function of the setting of all of the variables in the game. This naturally increases the complexity of the problem and means that cost functions in our model may not have succinct representations, but the tradeoff is that our model can capture dependencies between different variables in the game, which is not possible in the combinatorial contracts model.

Boolean Games Finally, there have been several studies devoted to understanding solution concepts in games with lexicographic qualitative and quantitative preferences [126–130], but these do not consider problems of incentive design. Other previous work has studied the idea of a principal introducing a taxation scheme into games where players have a very similar preference structure to the one we consider [224–226, 260]. It was assumed that the principal did this in order to rationally motivate players to make choices so as to satisfy the principal's goal, which was expressed as a propositional logic formula. Additionally, Gutierrez et al. [207] study a related problem called 'equilibrium design', but they consider players with only quantitative objectives. A key assumption in such previous works was that the principal can *fully observe* the outcome $\vec{v}$ selected by the agents. However, the principal does not have this luxury in many real-world situations, and it is this issue that motivates the study of the principal-agent problem. Other studies of incomplete information in Boolean games introduce

partial observability at the level of the agents, as opposed to an external principal who seeks to influence their behaviour [261, 262].

3.2 Preliminaries

Notation Where Φ is a set of Boolean variables, we let $\mathbb{B}^\Phi$ denote the set of possible *valuations*, which are combinations of truth assignments to each variable in Φ. Where Ψ is also a set of Boolean variables and $v \in \mathbb{B}^\Phi$ is a valuation in Φ, we will let $v|_\Psi$ denote the *restriction* of v to Ψ, which is the Boolean valuation $v' \in \mathbb{B}^{\Phi \cap \Psi}$ that agrees with v on all variables in $\Phi \cap \Psi$.

Boolean Games In this study, we will model a multi-agent moral hazard problem in the context of one-shot Boolean Games with costs, as presented in [224]. A *Boolean Game with costs* (henceforth "Boolean game" or simply "game") is a structure

$$B = (N, \Phi, (\Phi_i)_{i \in N}, (\gamma_i)_{i \in N}, (c_i)_{i \in N}),$$

where:

- $N = \{1, \ldots, n\}$ is a set of *agents*;
- $\Phi = \{p_1, p_2, \ldots, p_m\}$ is a finite set of $m \geq n$ *Boolean variables*;
- For each $i \in N$, the set Φ_i is the non-empty set of *Boolean variables uniquely controlled by agent* i, such that the collection $(\Phi_i)_{i \in N}$ forms a partition of Φ;
- For each $i \in N$, formula γ_i is the *goal* of agent i, which is represented as a propositional formula over Φ;
- For each $i \in N$, with $c_i : \mathbb{B}^\Phi \to \mathbb{Q}_+$ we represent the *cost function* of i, which assigns to each valuation $\vec{v}$ of the variables in Φ a non-negative cost $c_i(\vec{v})$, indicating the cost incurred by i under the valuation. We let c_i^* denote the maximal cost that i can incur under their c_i.

Figure 3.1 depicts a simple example of a Boolean game with two agents controlling sets of three and two variables respectively and example goals for each agent, along with an illustration of how a valuation gives rise to costs for each agent.

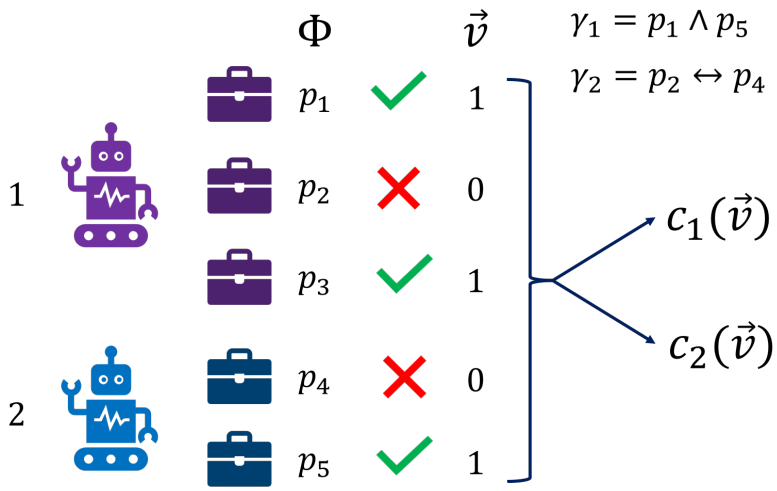

Figure 3.1: An example of a Boolean game with costs. Here, two robots (1 and 2) are able to select which of the tasks that are assigned to them to complete, which are represented by sets of three and two Boolean variables respectively. An example valuation of $\vec{v} = (1, 0, 1, 0, 1)$ is chosen, which gives rise to costs $c_1(\vec{v})$ and $c_2(\vec{v})$ for robots 1 and 2, respectively. In this example, the goal of robot 1 is $\gamma_1 = p_1 \wedge p_5$ and the goal of robot 2 is $\gamma_2 = p_2 \leftrightarrow p_4$ and hence, both robots' goals are satisfied in this instance.

A *strategy* $v_i : \Phi_i \to \{\top, \bot\}$ for agent i is defined as an assignment of truth values to every variable in Φ_i. Here, we will use 1 and 0 interchangeably with $\top$ and $\bot$, respectively, in the context of strategies. For each agent $i \in N$, let V_i be their set of possible strategies, and let $V := \prod_{i=1}^{n} V_i = \mathbb{B}^{\Phi}$. A *strategy profile* or *outcome* is then a tuple of strategies $\vec{v} = (v_1, \ldots, v_n)$, where agents select their strategies simultaneously in the absence of communication and knowledge of what strategies other agents selected. Additionally, given an agent $i \in N$, a strategy profile $\vec{v}$, and an individual strategy $v_i' \in \mathbb{B}^{\Phi_i}$ for i, we let $(\vec{v}_{-i}, v_i')$ be the strategy profile obtained by replacing v_i in $\vec{v}$ with v_i', i.e., the strategy profile $(v_1, \ldots, v_{i-1}, v_i', v_{i+1}, \ldots, v_n)$. Given a game B, we also define its *costless* version as the game $B^0 = (N, \Phi, (\Phi_i)_{i \in N}, (\gamma_i)_{i \in N}, (c_i^0)_{i \in N})$ which is the same as game B, except that $c_i^0(\vec{v}) = 0$ for all $\vec{v} \in \mathbb{B}^{\Phi}$ and $i \in N$.

For a strategy profile $\vec{v}$ and a propositional logic formula φ over the set of Boolean variables Φ, we write $\vec{v} \models \varphi$ to mean that $\vec{v}$ satisfies φ. If it holds that $\vec{v} \models \gamma_i$ for some agent $i \in N$, we say that agent i's goal is satisfied under $\vec{v}$ and i is a *winner* under $\vec{v}$. Any agent whose goal is not satisfied under a strategy profile $\vec{v}$ will be referred to as a *loser* under $\vec{v}$. We write $W(\vec{v})$ and $L(\vec{v})$ to denote the

sets of winners and losers under a given strategy profile $\vec{v}$, respectively.

Utilities, Preferences, and Nash Equilibrium In order to model both qualitative *and* quantitative preferences, we consider a model where agent preferences are defined according to a lexicographic ordering on goal satisfaction and their incurred costs. Specifically, an agent i prefers any outcome in which their goal γ_i is satisfied over any outcome in which it is not, no matter what cost they incur. Secondarily, each agent prefers to minimise the cost that they incur. Formally, we define the *utility function* u_i for each agent $i \in N$ given a strategy profile $\vec{v}$ as follows:

$$u_i(\vec{v}) = \begin{cases} 1 + c_i^* - c_i(\vec{v}) & \text{if } \vec{v} \models \gamma_i \\ -c_i(\vec{v}) & \text{otherwise.} \end{cases}$$

Observe that if an agent i achieves their goal, then their utility will lie in the range $[1, c_i^* + 1]$; if their goal is not achieved, then their utility will lie in the range $[-c_i^*, 0]$. Preference relations $\succeq_i$ over outcomes for each $i \in N$ are defined in the obvious way: $\vec{v}_1 \succeq_i \vec{v}_2$ if and only if $u_i(\vec{v}_1) \geq u_i(\vec{v}_2)$, with indifference relations $\sim_i$ and strict preference relations $\succ_i$ defined in the usual way.

The primary solution concept we will work with is the pure strategy Nash equilibrium. If a strategy $v_i' \in V_i$ exists such that $(\vec{v}_{-i}, v_i') \succ_i \vec{v}$ for an agent $i \in N$, we say that i has a *beneficial deviation* from $\vec{v}$ to $(\vec{v}_{-i}, v_i')$. Where B is a game, we write $\text{NE}(B)$ to denote the set of Nash equilibria of B. Additionally, if φ is a Boolean logic formula over Φ, we let $\text{NE}_\varphi(B) = \{\vec{v} \in \text{NE}(B) \mid \vec{v} \models \varphi\}$ denote the set of Nash equilibria of B which satisfy formula φ. Figure 3.2 depicts an example of the relationship between different sets of strategy profiles and an agent's preference relation over members of each set.

3.3 The Principal-Agent Verification Problem

In this study, we aim to formulate and analyse the multi-agent moral hazard problem in the context of Boolean games with costs, where hidden actions are modelled by the principal only being able to observe the values of some subset $\mathcal{O} \subseteq \Phi$, known as the *observable set*. An *observation* by the principal is defined to be an assignment of truth values to all variables in its observable set $\mathcal{O}$, and is denoted by $\vec{o}$.[1] In general, a single observation $\vec{o}$ may correspond to more than

[1] A special case of this is when we have $\mathcal{O} = \emptyset$, where the principal cannot observe the values of any of the Boolean variables in the game. In this case, we can say that the principal makes a *null observation*, which is written as $\vec{o} = \varepsilon$.

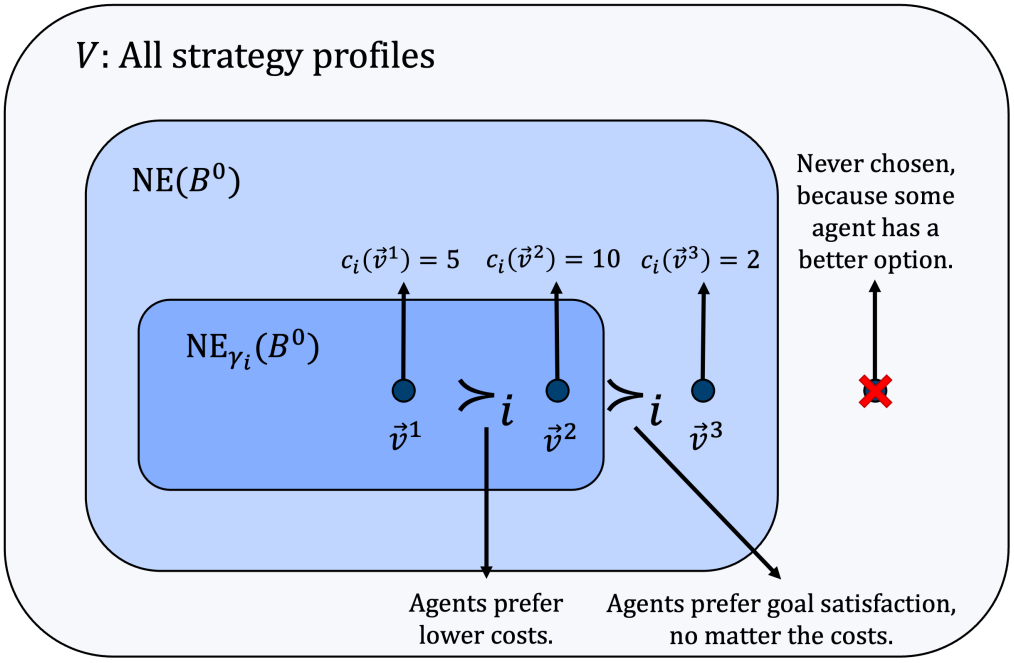

Figure 3.2: An example of an agent's preferences over outcomes in a Boolean game with costs. Given that agents are assumed to act rationally, we rule out any non-equilibrium strategy profiles from being chosen. Within the set of initial equilibria (i.e., equilibria of the game B^0 whose costs are all zero) of a game, a player i always prefers equilibria in which their goal is achieved over those in which it is not. Secondarily, agents prefer equilibria in which their costs are lower.

one underlying strategy profile $\vec{v}$ if not all of the variables are fully observable. Thus, the observable set $\mathcal{O}$ naturally gives rise to the possibility for strategy profiles to be *indistinguishable* from one another. This can be expressed formally as an observational equivalence relation $=_{\mathcal{O}}$, defined over the set of observations $\mathbb{B}^{\mathcal{O}}$ such that for all strategy profiles $\vec{v}, \vec{v}' \in V$, we say that $\vec{v}$ is *indistinguishable* from $\vec{v}'$, written $\vec{v} =_{\mathcal{O}} \vec{v}'$, if it is the case that $\vec{v}|_{\mathcal{O}} = \vec{v}'|_{\mathcal{O}}$. If it is *not* the case that $\vec{v} =_{\mathcal{O}} \vec{v}'$, then we say that strategy profile $\vec{v}$ is *distinguishable* from strategy profile $\vec{v}'$ and write $\vec{v} \neq_{\mathcal{O}} \vec{v}'$ in that case.

The problem faced by the principal is to design a contract so that they are able to ensure/verify that their objective has been accomplished on some or all of the Nash equilibria under the contract. To begin with, we will first analyse the simple case where the principal does not intervene, but simply asks whether they can *verify* whether or not some or all Nash equilibria consistent with any observation will satisfy their goal, given as a propositional logic formula φ. The value of studying this problem is that it provides a baseline against which the principal can assess the benefit of any potential contract that is designed.

In order to formally state these problems, we first make precise the notion of consistency. Formally, we say that a strategy profile $\vec{v}$ is *consistent* with an observation $\vec{o}$ if it is the case that $\vec{v}|_{\mathcal{O}} = \vec{o}$. Then, we define the *consistent set* of an observation $\vec{o}$ in a given Boolean game B in the following way: $\text{CONS}(B, \vec{o}) = \{\vec{v} \in \text{NE}(B) \mid \vec{v} \text{ is consistent with } \vec{o}\}$. Finally, we define the set $\text{ENV}(B, \vec{o}, \varphi) = \text{NE}_{\varphi}(B) \cap \text{CONS}(B, \vec{o})$ to be the set of Nash equilibria in game B that satisfy φ and are consistent with $\vec{o}$. The first decision problem of interest is then stated as follows:

E-NASH VERIFIABILITY:
Given: Game B, observation $\vec{o} \in \mathbb{B}^{\mathcal{O}}$, formula φ.
Question: Is it the case that $\text{ENV}(B, \vec{o}, \varphi) \neq \emptyset$?

Whilst an answer to the E-Nash Verifiability problem provides useful information to the principal about the *possibility* of their goal being achieved in some Nash equilibrium that is consistent with an observation, it does not make any guarantees as to whether or not the principal's goal has actually been achieved, given what they have observed. To obtain such a guarantee, we require the natural counterpart to this problem – the *A-Nash Verifiability* problem[2]:

A-NASH VERIFIABILITY:
Given: Game B, observation $\vec{o} \in \mathbb{B}^{\mathcal{O}}$, formula φ.
Question: Is it the case that $\text{CONS}(B, \vec{o}) \subseteq \text{NE}_{\varphi}(B)$?

If the answer to the A-Nash Verifiability problem is "yes", then the principal can be assured that any Nash equilibrium consistent with their observation satisfies their goal. The following complexity results for the E- and A-Nash Verifiability problems can be readily derived from existing results relating to Boolean games [224]:

Proposition 3.1. E-NASH VERIFIABILITY *is* Σ_2^p*-complete, and* A-NASH VERIFIABILITY *is* Π_2^p*-complete.*

Proof. We consider E-NASH VERIFIABILITY first. For membership in Σ_2^p, the problem of verifying whether a candidate strategy profile $\vec{v}$ is a Nash equilibrium of B is coNP-complete [224]. Checking whether $\vec{v}$ is consistent with $\vec{o}$ and satisfies φ can be done in polynomial time, and thus, membership in Σ_2^p follows.

[2]The notations E- and A- are used to indicate existential and universal reasoning about the set of Nash equilibria in a game respectively.

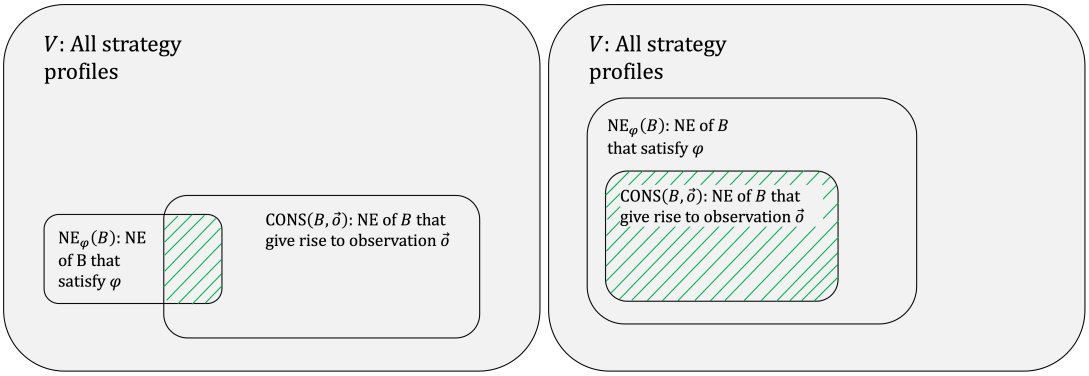

Figure 3.3: Venn Diagrams depicting the regions of strategy space whose properties the E- and A-Nash Verifiability problems check. Left: E-Nash Verifiability problem, which checks whether there exists a Nash equilibrium satisfying φ that is consistent with an observation $\vec{o}$. Right: A-Nash Verifiability problem, which checks whether all Nash equilibria satisfying φ are consistent with $\vec{o}$.

Hardness follows from the fact that the problem of checking whether a Boolean game with no costs has a pure-strategy Nash equilibrium is Σ_2^p-complete [100] – simply set $\mathbb{B}^{\mathcal{O}} = \emptyset$, $\vec{o} = \varepsilon$, and $\varphi = \top$ in a Boolean game with no costs but the same structure as the one in which we are interested in determining the existence of Nash equilibria.

Now we consider A-NASH VERIFIABILITY. For membership, simply note that the A-Nash Verifiability problem is the complement of the E-Nash Verifiability problem, where the set $\text{ENV}(B, \vec{o}, \neg\varphi) \neq \emptyset$ if and only if the answer to A-Nash Verifiability is "no." Similarly, hardness follows by reducing the E-Nash Verifiability problem to the complement of A-Nash Verifiability. □

Example 3.1. To illustrate the differences between the E-Nash and the A-Nash Verifiability problems, consider the game B_1 consisting of two players with $\Phi_1 = \{p_1\}$, $\Phi_2 = \{p_2\}$, goals $\gamma_1 = \top$, $\gamma_2 = \neg p_1$, cost function as given by the table on the left in Figure 3.4, and observable set $\mathcal{O} = \{p_1\}$. Now suppose that the principal's objective is $\varphi = p_2$ and consider the E- and A-Nash Verifiability problems for each observation $\vec{o} \in \mathbb{B}^{\mathcal{O}} = \{0, 1\}$. Firstly, we can identify the Nash equilibria in B_1 as being $(0, 0)$ and $(0, 1)$, with only the latter satisfying φ. Under the observation $o_1 = (1)$, there is no Nash equilibrium in B_1 that is consistent with o_1, so the answer to E-NASH VERIFIABILITY is "no" and the answer to A-NASH VERIFIABILITY is "yes" for o_1. However, under the observation $o_2 = (0)$, it is the case that $\text{CONS}(B, \vec{o}) \cap \text{NE}_\varphi(B) \neq \emptyset$ but not the case that $\text{CONS}(B, \vec{o}) \subseteq \text{NE}_\varphi(B)$. Thus, the answer to E-Nash verifiability for o_2 is "yes" while it is "no" for A-Nash verifiability.

	1	0
1	$(2,0)$	$(3,0)$
0	$(1,1)$	$(1,1)$

	1	0
1	$(10,1)$	$(10,2)$
0	$(1,2)$	$(1,1)$

Figure 3.4: Two cost matrices representing the agents' costs incurred under different strategy profiles in two different Boolean games. The game represented by the left figure is referred to as B_1 and discussed in Example 3.1. The game represented by the right figure is referred to as B_2 and discussed in Example **??**. Player 1 (the row player) controls variable p_1 and player 2 (the column player) controls variable p_2.

3.4 Contract Design

The verification problem studied thus far allows the principal to obtain helpful information about whether or not they can be confident that their goal was satisfied under any possible observation that is consistent with at least one Nash equilibrium. However, in the classical formulations of the moral hazard problem, a principal is tasked with *designing* a contract to align the agents' incentives with their own. In this study, the principal's limitation lies not in an inability to accurately observe the outcome of its observables, but rather in their ability to only observe some subset of the actions chosen, as well as the agents' lack of willingness to compromise on satisfying their goal, no matter what the incentives are. For this, we first introduce the concept of contracts, which amount to an alteration of the costs incurred by the agents for taking various actions. This is similar to the notion of taxation schemes introduced in [224], but with the critical distinction that in our setting, the principal can only reward agents based on the outcomes of *observable* variables. Formally, a contract $\kappa = (\kappa_1, \ldots, \kappa_n)$ is a tuple of functions $\kappa_i : \mathbb{B}^{\mathcal{O}} \to \mathbb{Q}_{\geq 0}$ for each agent $i \in N$ that map each possible observation by the principal to a non-negative rational number. Let $\mathcal{K}$ denote the set of all contracts for a given game and for a given contract κ, let κ_i^* denote the highest possible payment that the principal offers to each agent $i \in N$ over all possible observations. Then, given a game B and an observable set $\mathcal{O}$, a contract κ gives rise to a new utility function u_i^κ for each agent $i \in N$ given a strategy profile $\vec{v}$:

$$u_i^\kappa(\vec{v}) = \begin{cases} 1 + c_i^* + \kappa_i(\vec{v}\,|_{\mathcal{O}}) - c_i(\vec{v}) & \text{if } \vec{v} \models \gamma_i \\ -\kappa_i^* + \kappa_i(\vec{v}\,|_{\mathcal{O}}) - c_i(\vec{v}) & \text{otherwise.} \end{cases}$$

Defining the utility of agents under a given contract in this way captures two desirable properties. Firstly, it preserves the lexicographic preferences of agents – each agent is guaranteed to achieve a strictly positive utility if their goal is satisfied, whereas they can achieve a utility of at most zero if it is not. Secondly, regardless of whether an agent's goal is satisfied, their utility is strictly increasing

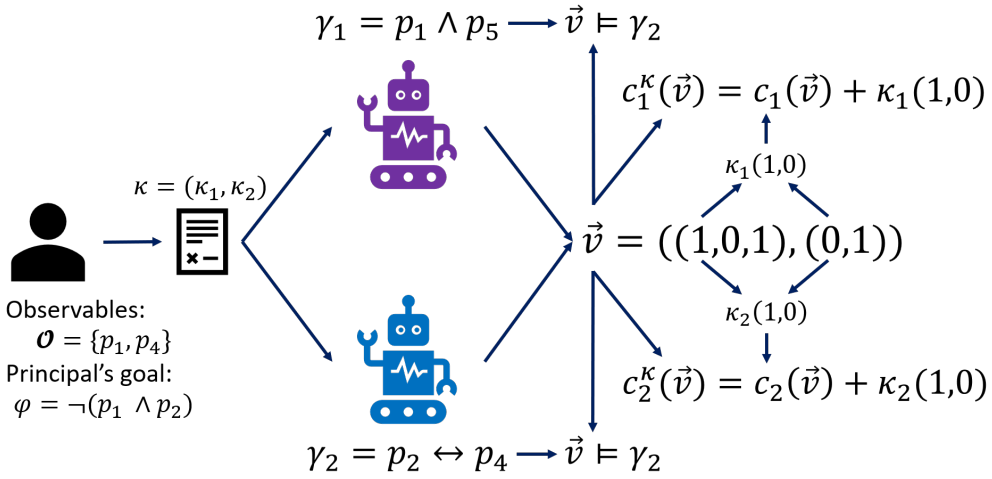

Figure 3.5: The workflow of our computational model of the contract design problem in the context of Boolean games with costs. The principal is able to observe a subset of the set of all Boolean variables, and has a goal specified as a propositional logic formula. A contract, mapping observed outcomes by the principal to payments for each agent is applied to the game and the players choose a setting of the Boolean variables under their control. This leads to an outcome $\vec{v}$ giving rise to a cost $c_i^\kappa(\vec{v})$ incurred by each agent i as well as a satisfaction value for their logical goal γ_i.

in the payment received from the principal and strictly decreasing in the costs they incur. Given a game B, an observable set $\mathcal{O}$ and a contract κ, we define the *Boolean game induced by* κ to be the game B^κ which is as game B, but where each agent i's utility is given by u_i^κ. Given a contract κ, an agent $i \in N$ and two strategy profiles $\vec{v}^1, \vec{v}^2$, we write $\vec{v}^1 \succeq_i^\kappa \vec{v}^2$ if and only if $u_i^\kappa(\vec{v}^1) \geq u_i^\kappa(\vec{v}^2)$ and define $\succ_i^\kappa$ in the obvious manner.

Inducing and Eliminating Equilibria

Due to the partial observability of the agents' actions and the fact that agents will always prefer to achieve their goal than not to do so, there are limits on the ability of a principal to either introduce or eliminate equilibria from a given game. As a special case, note that if all agents' actions are fully observable ($\mathcal{O} = \Phi$), then the contract problem can be characterised by the game's hard and soft equilibria – respectively those that can not be eliminated under any contract and those that can [224–226]. Thus, we will focus on the case where $\mathcal{O} \subset \Phi$. In this scenario, it is not in general possible to characterise the manipulability of a given game B by analysing the qualitative structure of the underlying costless Boolean game, as the principal is no longer able to eliminate the cost considerations for

all agents by designing a contract so that each agent's quantitative payoff is constant regardless of their actions. Thus, the ability of the principal to induce or eliminate equilibria by means of contract design will be critically dependent on the initial cost functions of the agents.

Example 3.2. As an example of the importance of initial costs in this model, consider a game B_2 consisting of two players with goals $\gamma_1 = \gamma_2 = \top$, with a cost function as given by the table on the right in Figure 3.4, and observable set $\mathcal{O} = \{p_1\}$. Now suppose again that the principal's objective is $\varphi = p_2$. We write $\bar{p}_x$, $x \in \{1, 2\}$, as an alternative for $\neg p_x$. Firstly, note that the only Nash equilibrium outcome of B_2 is $(0, 0)$. Now, because the principal cannot observe p_2 directly, the only way for them to ensure that p_2 is rationally chosen by player 2 is to offer a contract to *player* 1 to make the choice of setting $p_1 = 1$ more rational than setting it to 0. However, in order to do so, such a contract κ must satisfy $\kappa_1(p_1) > \kappa_1(\bar{p}_1) + 9$, under which the new unique Nash equilibrium becomes $(1, 1)$. This illustrates a scenario in which it is lucrative for an employee to control what is observable by the employer so as to benefit from indirect incentivisation.

Before proceeding, we find it useful to define some terminology [225]. Firstly, we define the set $\text{INIT}(B) = \text{NE}(B^0)$ of *initial equilibria* to be the set of Nash equilibria in the costless game B^0. Formally, we will say that given a game B and an observable set $\mathcal{O}$, a strategy profile $\vec{v}$ is an *inducible equilibrium* if there exists a contract κ such that $\vec{v} \in \text{NE}(B^\kappa)$, and we let $\text{IND}(B, \mathcal{O})$ be the set of inducible equilibria of a game B with respect to an observable set $\mathcal{O}$. Additionally, we will say that a strategy profile $\vec{v}$ is a *hard equilibrium* (with respect to $\mathcal{O}$), written $\vec{v} \in \text{HARD}(B, \mathcal{O})$, if $\vec{v} \in \text{NE}(B)$ and there is no contract κ such that $\vec{v} \notin \text{NE}(B^\kappa)$. Complementarily, we say that a strategy profile $\vec{v}$ is a *soft equilibrium*, written $\vec{v} \in \text{SOFT}(B, \mathcal{O})$, if $\vec{v} \in \text{NE}(B) \backslash \text{HARD}(B, \mathcal{O})$. It is easy to verify that the previously defined sets are related in the following way:

Proposition 3.2. *For all Boolean games B and observable sets $\mathcal{O}$, it holds that* $\text{HARD}(B, \mathcal{O}) \subseteq \text{IND}(B, \mathcal{O})$, $\text{SOFT}(B, \mathcal{O}) \subseteq \text{IND}(B, \mathcal{O})$, *and* $\text{NE}(B^\kappa) \subseteq \text{IND}(B, \mathcal{O}) \subseteq \text{INIT}(B)$ *for all contracts κ.*

To reason about the principal's ability to design effective contracts, a method of analysing the inducibility and eliminability of sets of strategy profiles is needed. Following the approach of [226], we will introduce different types of potential deviations between strategy profiles. Suppose that $\vec{v}, \vec{v}' \in V$ are two distinct strategy profiles such that $\vec{v}' = (\vec{v}_{-i}, v_i')$ for some $i \in N$ and $v_i' \in V_i \setminus \{v_i\}$. Then we say that:

- $\vec{v}'$ is an *initial deviation* for i from $\vec{v}$, written $\vec{v} \rightharpoonup_i \vec{v}'$, if $i \in W(\vec{v}) \Rightarrow i \in W(\vec{v}')$.
- $\vec{v}'$ is an *inducible deviation* for i from $\vec{v}$, written $\vec{v} \rightarrow_i \vec{v}'$, if there is a contract κ such that $\vec{v}' \succ_i^\kappa \vec{v}$. Under such a contract, we say that κ *induces* a deviation from $\vec{v}$ to $\vec{v}'$.
- $\vec{v}'$ is a *soft deviation* for i from $\vec{v}$, written $\vec{v} \leftrightarrows_i \vec{v}'$, if both $\vec{v} \rightarrow_i \vec{v}'$ and $\vec{v}' \rightarrow_i \vec{v}$.
- $\vec{v}'$ is a *hard deviation* for i from $\vec{v}$, written $\vec{v} \rightrightarrows_i \vec{v}'$, if $\vec{v} \rightarrow_i \vec{v}'$ but not $\vec{v}' \rightarrow_i \vec{v}$.

Note that just because a strategy profile is an initial deviation from another strategy profile, this does not necessarily imply that a contract exists which could *induce* that deviation, i.e., convert it into a beneficial deviation. Instead, we need the stronger notion of *inducible deviations*, which is, in turn, used to define soft and hard deviations. The following proposition characterises the relationship between initial and inducible deviations.

Proposition 3.3. *Let B be a game, $\mathcal{O} \subseteq \Phi$ an observable set, and $\vec{v}, \vec{v}' \in V$ be two distinct strategy profiles such that $\vec{v}' = (\vec{v}_{-i}, v_i')$ for some $i \in N$ and $v_i' \in V_i \setminus \{v_i\}$. Then, $\vec{v} \rightarrow_i \vec{v}'$ if and only if $\vec{v} \rightharpoonup_i \vec{v}'$ and one of the following conditions hold: (1) $\vec{v} \neq_{\mathcal{O}} \vec{v}'$, or (2) $\vec{v} =_{\mathcal{O}} \vec{v}'$ and $\vec{v}' \succ_i \vec{v}$.*

Proof. For the forward direction, suppose that it is not the case that $\vec{v} \rightharpoonup_i \vec{v}'$. By definition, this means that $i \in W(\vec{v}) \cap L(\vec{v}')$. Since any winner's utility is guaranteed to be positive and any loser's utility is guaranteed to be negative regardless of the contract chosen by the principal, there is no contract κ such that $\vec{v}' \succ_i^\kappa \vec{v}$. Now suppose that $\vec{v} \rightharpoonup_i \vec{v}'$, but that neither of two conditions hold, i.e., $\vec{v} =_{\mathcal{O}} \vec{v}'$ and $\vec{v} \succeq_i \vec{v}'$. For any contract κ, the difference in utility $\Delta_i = u_i^\kappa(\vec{v}) - u_i^\kappa(\vec{v}_{-i}, v_i')$ of agent i between $\vec{v}$ and $\vec{v}'$ is

$$\Delta_i = \begin{cases} c_i(\vec{v}') - c_i(\vec{v}) & i \in W(\vec{v}_{-i}, v_i') \cap W(\vec{v}) \\ c_i(\vec{v}') - (1 + c_i^* + \\ \qquad \kappa_i^* + c_i(\vec{v})) & i \in W(\vec{v}_{-i}, v_i') \cap L(\vec{v}) \\ c_i(\vec{v}') - c_i(\vec{v}) & i \in L(\vec{v}_{-i}, v_i') \cap L(\vec{v}). \end{cases}$$

Two observations are in order. Firstly, because the two strategy profiles are indistinguishable, any contract will pay every agent the same amount under both strategy profiles, so the difference in utility between the two is only dependent on the initial cost structure of the game and potentially the highest possible contract

payoff κ_i^*. Secondly, we can infer from the assumption that $\vec{v} \succeq_i \vec{v}'$ that this difference in utility is non-negative in all three cases. Thus, we have that $\vec{v} \succeq_i^\kappa \vec{v}'$ for all contracts κ.

For the reverse direction suppose that $\vec{v} \rightharpoonup_i \vec{v}'$ and that the first condition holds. Then, it can be easily verified that any contract κ such that $\kappa_i(\vec{v}|_{\mathcal{O}}) = 0$ and $\kappa_i(\vec{v}'|_{\mathcal{O}}) = c_i^* + 1$ will satisfy $\vec{v}' \succ_i^\kappa \vec{v}$. Now assume instead that $\vec{v} =_{\mathcal{O}} \vec{v}'$ and $\vec{v}' \succ_i \vec{v}$. Clearly, a null contract κ^0 that simply offers all agents no payment for all observations satisfies $\vec{v}' \succ_i^\kappa \vec{v}$. □

This result highlights the fact that the ability to design contracts to induce deviations between two strategy profiles is wholly determined by the initial cost and goal structure of the game in the case where the strategy profiles are indistinguishable. Next, we present a characterisation of the set of inducible equilibria of a game, which tells a principal when it is possible to design a contract κ such that a particular outcome becomes a Nash equilibrium under κ:

Proposition 3.4. *Let B be a game, $\mathcal{O} \subseteq \Phi$ an observable set, and $\vec{v}$ a strategy profile. Then, $\vec{v} \in \text{IND}(B, \mathcal{O})$ if and only if the following conditions hold: (1) $\vec{v} \in \text{INIT}(B)$; and (2) for all agents $i \in N$, and all choices $v_i' \in V_i$ such that $(\vec{v}_{-i}, v_i') =_{\mathcal{O}} \vec{v}$ and $i \in W(\vec{v}) \Leftrightarrow i \in W(\vec{v}_{-i}, v_i')$, we have that $c_i(\vec{v}_{-i}, v_i') \geq c_i(\vec{v})$.*

Proof. For the forward direction, suppose that $\vec{v} \notin \text{INIT}(B)$. Then, there is some agent $i \in L(\vec{v})$ and an alternative choice $v_i' \in V_i$ such that $i \in W(\vec{v}_{-i}, v_i')$. Thus, under any contract κ, we will have $u_i^\kappa(\vec{v}) \leq 0$ and $u_i^\kappa(\vec{v}_{-i}, v_i') \geq 1$, so $\vec{v} \notin \text{NE}(B^\kappa)$. Since this holds for all contracts, we can conclude that $\vec{v}$ is not an inducible equilibrium of B.

Now assume that the first condition is satisfied, but for some agent $i \in N$ and some $v_i' \in V_i$ such that $(\vec{v}_{-i}, v_i') =_{\mathcal{O}} \vec{v}$ and $i \in W(\vec{v}) \Leftrightarrow i \in W(\vec{v}_{-i}, v_i')$, it holds that $c_i(\vec{v}_{-i}, v_i') < c_i(\vec{v})$. First observe that under the first condition, it must be the case that $i \in W(\vec{v}_{-i}, v_i') \Rightarrow i \in W(\vec{v})$, otherwise i would have a beneficial deviation from $\vec{v}$ to $(\vec{v}_{-i}, v_i')$. Now, because $(\vec{v}_{-i}, v_i') =_{\mathcal{O}} \vec{v}$, any contract κ will pay the same amount to i under $\vec{v}$ and $(\vec{v}_{-i}, v_i')$. Thus, taking the difference in utility of agent i under any contract κ and using the assumption that $i \in W(\vec{v}) \Leftrightarrow i \in W(\vec{v}_{-i}, v_i')$, we have $u_i^\kappa(\vec{v}) - u_i^\kappa(\vec{v}_{-i}, v_i') = c_i(\vec{v}_{-i}, v_i') - c_i(\vec{v})$. Under the other assumption that $c_i(\vec{v}_{-i}, v_i') < c_i(\vec{v})$, it follows that $u_i^\kappa(\vec{v}) < u_i^\kappa(\vec{v}_{-i}, v_i')$ and hence, i has a beneficial deviation from $\vec{v}$ to $(\vec{v}_{-i}, v_i')$. Since this holds for any contract, we can conclude that $\vec{v}$ is not an inducible equilibrium of B.

For the reverse direction, assume that the two conditions hold. Then, consider the contract κ such that for all agents $i \in N$ and observations $\vec{o} \in \mathbb{B}^{\mathcal{O}}$, we have:

$$\kappa_i(\vec{o}) = \begin{cases} c_i^* + 1 & \text{if } \vec{v}|_{\mathcal{O}} = \vec{o} \\ 0 & \text{otherwise.} \end{cases}$$

What remains is to show that $\vec{v} \in \text{NE}(B^\kappa)$. Recall that because $\vec{v} \in \text{INIT}(B)$ by assumption, we have $i \in W(\vec{v}_{-i}, v_i') \Rightarrow i \in W(\vec{v})$, for all agents $i \in N$ and all strategies $v_i' \in V_i$. Thus, the only reason that some agent could have a beneficial deviation from v_i to some v_i' is if they achieve a higher net payoff, being the difference between their reward under the contract κ_i and the cost c_i that they incur. We can consider two classes of possible cases for deviation $v_i' \in V_i \setminus \{v_i\}$ by any given agent $i \in N$: cases where $(\vec{v}_{-i}, v_i') \neq_{\mathcal{O}} \vec{v}$ and cases where $(\vec{v}_{-i}, v_i') =_{\mathcal{O}} \vec{v}$. In the first case, we know that $\kappa_i(\vec{v}|_{\mathcal{O}}) = c_i^* + 1$ and $\kappa_i(\vec{v}'|_{\mathcal{O}}) = 0$, so the difference in utility $\Delta_i = u_i^\kappa(\vec{v}) - u_i^\kappa(\vec{v}_{-i}, v_i')$ for agent i is

$$\Delta_i = \begin{cases} 1 + c_i^* - c_i(\vec{v}) + c_i(\vec{v}_{-i}, v_i') & i \in W(\vec{v}_{-i}, v_i') \cap W(\vec{v}) \\ 3 + 3c_i^* - c_i(\vec{v}) + c_i(\vec{v}_{-i}, v_i') & i \in L(\vec{v}_{-i}, v_i') \cap W(\vec{v}) \\ 1 + c_i^* - c_i(\vec{v}) + c_i(\vec{v}_{-i}, v_i') & i \in L(\vec{v}_{-i}, v_i') \cap L(\vec{v}). \end{cases}$$

Note that the case where $i \in W(\vec{v}_{-i}, v_i') \cap L(\vec{v})$ is not possible because $\vec{v} \in \text{INIT}(B)$ by assumption. Furthermore, in all of these possibilities, the difference in utility is non-negative and hence, no agent $i \in N$ has a beneficial deviation from $\vec{v}$ to a distinguishable strategy profile $(\vec{v}_{-i}, v_i')$. Now, for the case where $(\vec{v}_{-i}, v_i') =_{\mathcal{O}} \vec{v}$, note that the contract payoff for both strategy profiles is the same, giving a difference in utility of

$$\Delta_i = \begin{cases} c_i(\vec{v}_{-i}, v_i') - c_i(\vec{v}) & i \in W(\vec{v}_{-i}, v_i') \cap W(\vec{v}) \\ 2 + 2c_i^* - c_i(\vec{v}) + c_i(\vec{v}_{-i}, v_i') & i \in L(\vec{v}_{-i}, v_i') \cap W(\vec{v}) \\ c_i(\vec{v}_{-i}, v_i') - c_i(\vec{v}) & i \in L(\vec{v}_{-i}, v_i') \cap L(\vec{v}). \end{cases}$$

Under the second condition, all three terms are non-negative and hence, there is no beneficial deviation for any agent $i \in N$ from $\vec{v}$ under κ. Thus, $\vec{v} \in \text{IND}(B, \mathcal{O})$. □

Now, we focus on characterising the sets of hard and soft equilibria, which allows a principal to check whether an undesirable initial equilibrium can be eliminated from a game.

Proposition 3.5. *Let B be a game, $\mathcal{O} \subseteq \Phi$ an observable set, and $\vec{v}$ a strategy profile. Then, $\vec{v} \in \text{SOFT}(B, \mathcal{O})$ if and only if the following conditions hold: (1) $\vec{v} \in \text{INIT}(B)$; and (2) there is an agent $i \in N$ and a strategy $v_i' \in V_i \setminus \{v_i\}$ such that $(\vec{v}_{-i}, v_i') \neq_{\mathcal{O}} \vec{v}$ and $i \in W(\vec{v}_{-i}, v_i') \Leftrightarrow i \in W(\vec{v})$.*

Proof. For the forward direction, it is immediate that if $\vec{v} \notin \text{INIT}(B)$, then $\vec{v} \notin \text{SOFT}(B, \mathcal{O}) \subseteq \text{INIT}(B)$. Hence, assume that $\vec{v} \in \text{INIT}(B)$ and suppose that for all agents $i \in N$ and all strategies $v_i' \in V_i \setminus \{v_i\}$ such that $(\vec{v}_{-i}, v_i') \neq_{\mathcal{O}} \vec{v}$, it is not the case that $i \in W(\vec{v}_{-i}, v_i') \Leftrightarrow i \in W(\vec{v})$. Now, for a given agent $i \in N$, it is not possible that $i \in L(\vec{v}) \cap W(\vec{v}_{-i}, v_i')$ because otherwise, $\vec{v}$ would not be an initial equilibrium. Thus, it must be that $i \in W(\vec{v}) \cap L(\vec{v}_{-i}, v_i')$, so no contract κ could convince any agent i to deviate to some distinguishable strategy profile $(\vec{v}_{-i}, v_i') \neq_{\mathcal{O}} \vec{v}$. Also, because $\vec{v} \in \text{INIT}(B)$, no agent has a beneficial deviation to any indistinguishable strategy profile. Moreover, because no contract can induce different costs for indistinguishable strategy profiles, there is no contract that can induce a beneficial deviation from $\vec{v}$ for any agent. Thus, we have that $\vec{v} \notin \text{SOFT}(B, \mathcal{O})$. For the reverse direction, assume that both conditions hold. Then, it is easy to check that any contract κ such that $\kappa_i(\vec{v}_{-i}, v_i') = c_i^* + 1$ and $\kappa_i(\vec{v}) = 0$ induces a beneficial deviation for i from $\vec{v}$ to $(\vec{v}_{-i}, v_i')$. Thus, it follows that $\vec{v} \in \text{SOFT}(B, \mathcal{O})$. □

Because $\text{HARD}(B) = \text{NE}(B) \setminus \text{SOFT}(B)$ by definition, a characterisation of the set of hard equilibria in a game follows immediately from Proposition 3.5:

Corollary 3.6. *Let B be a game, $\mathcal{O} \subseteq \Phi$ a non-empty observable set, and $\vec{v}$ a strategy profile. Then, $\vec{v} \in \text{HARD}(B, \mathcal{O})$ if and only if the following conditions hold: (1) $\vec{v} \in \text{NE}(B)$; and (2) for all agents $i \in N$ and all strategies $v_i' \in V_i \setminus \{v_i\}$ such that $(\vec{v}_{-i}, v_i') \neq_{\mathcal{O}} \vec{v}$, we have $(\vec{v}_{-i}, v_i') \rightrightarrows_i \vec{v}$.*

So far, our analysis has focused on whether or not a single strategy profile is an inducible, hard, or soft equilibrium. However, it is not sufficient in general to consider individual strategy profiles in this way, as there may be several strategy profiles that the principal would like to induce or eliminate. The more general problem faced by the principal is thus how it can determine whether a contract can be designed to jointly eliminate several equilibria in a game.

Therefore, we will extend the concept of eliminability to sets of strategy profiles. Given a set $X \subseteq V$ of strategy profiles, we say that X is *eliminable* if there is a contract κ such that $X \cap \text{NE}(B^\kappa) = \emptyset$. Any such contract κ is then said to *eliminate* X. To characterise the ability of the principal to eliminate sets of equilibria, we will utilise a graph-based approach to represent inducible deviations between strategy profiles. Given a Boolean game B and an observable set $\mathcal{O}$, we define the *potential deviation graph* of B with respect to $\mathcal{O}$ to be the directed graph $\mathcal{G}(B, \mathcal{O}) = (\mathcal{V}, E)$, where

- $\mathcal{V}$ is the set of vertices, where each vertex corresponds to a strategy profile $\vec{v} \in V$;
- $E = \{(\vec{v}, \vec{v}') \in \mathcal{V} \times \mathcal{V} \mid \vec{v} \rightarrow_i \vec{v}' \text{ for some } i \in N\}$ is the set of directed edges, where there is an edge from a vertex $\vec{v} \in \mathcal{V}$ to another vertex $\vec{v}' \in \mathcal{V}$ if and only if it is the case that $\vec{v} \rightarrow_i \vec{v}'$ for some $i \in N$.

The potential deviation graph of a game B with respect to an observable set $\mathcal{O}$ represents all possible deviations that could be induced by a contract κ. We say that a subgraph $D = (\mathcal{V}_D, E_D)$ of $\mathcal{G}(B, \mathcal{O})$ is the *deviation graph* induced by a contract κ iff it is the case that for every pair $\vec{v}, \vec{v}' \in \mathcal{V}_D$, we have $(\vec{v}, \vec{v}') \in E_D \Leftrightarrow \vec{v}' \succ_i^{\kappa} \vec{v}$. Intuitively, a contract κ induces a deviation graph D when only those inducible deviations that are actually made beneficial for some $i \in N$ under κ are included as edges in D, along with both of their endpoints. For the purposes of eliminating undesirable equilibria, we will also need to determine, given a deviation graph D of $\mathcal{G}(B, \mathcal{O})$, whether there exists a contract κ such that D is the deviation graph induced by κ. If such a contract exists, we will say that D is an *inducible deviation graph* of B wrt. $\mathcal{O}$.

We say that a tuple $P = (\vec{v}^1, \vec{v}^2, \ldots, \vec{v}^k)$ is a *deviation path* if it holds that $\vec{v}^1 \rightarrow_{i^1} \vec{v}^2 \rightarrow_{i^2} \ldots \rightarrow_{i^{k-1}} \vec{v}^k$, for some tuple of agents $(i^1, i^2, \ldots, i^{k-1})$. Equivalently, a deviation path may be defined simply as any directed path within the potential deviation graph $\mathcal{G}(B, \mathcal{O})$. A deviation path $(\vec{v}^1, \vec{v}^2, \ldots, \vec{v}^k)$ is a *deviation cycle* if $k \geq 2$ and $\vec{v}^1 = \vec{v}^k$. Additionally, we will say that a deviation path $(\vec{v}^1, \vec{v}^2, \ldots, \vec{v}^k)$ *involves agent* i if and only if we have $\vec{v}^j \rightarrow_i \vec{v}^{j+1}$ for some $j \in \{1, \ldots, k-1\}$ and $i \in N$.

Although deviation paths are defined with respect to sequences of strategy profiles that are pairwise related by inducible deviations, it will be remembered that the influence of the principal is restricted to providing incentives only based on what they can observe, rather than the strategy profile that is actually chosen. In particular, given two strategy profiles $\vec{v}, \vec{v}' \in V$ such that $\vec{v} \neq_{\mathcal{O}} \vec{v}'$ and $\vec{v} \rightarrow_i \vec{v}'$ for some $i \in N$, the principal cannot design a contract that rewards i for choosing $\vec{v}'$ over $\vec{v}$ directly, but this must instead be done indirectly, by rewarding i sufficiently more under the *observation* $\vec{v}'|_{\mathcal{O}}$ than the observation $\vec{v}|_{\mathcal{O}}$. From the point of view of the principal, then, it will sometimes be more useful to reason about *observed deviation paths*, which we define as a tuple $P_{\mathcal{O}} = (\vec{o}^1, \vec{o}^2, \ldots, \vec{o}^k)$ such that there exists a set $\{\vec{v}^1, \vec{v}^2, \ldots, \vec{v}^k\} \subseteq \mathcal{V}$ where for all $j \in \{1, \ldots, k\}$, we have (1) $\vec{v}^j|_{\mathcal{O}} = \vec{o}^j$ and (2) if $j < k$, then $(\vec{v}^j, \vec{v}^{j+1'}) \in E_D$ for some $\vec{v}^{j+1'} =_{\mathcal{O}} \vec{v}^{j+1}$. In other words, an observed deviation path need not be an actual deviation path

Valuation $\vec{v}$	Observation $\vec{o}$	$\vec{v} \in \text{NE}(B)$?	$\vec{v} \models \varphi$?
(0,0,0)	0	✗	✗
(0,0,1)	0	✗	✗
(0,1,0)	0	✓	✗
(0,1,1)	0	✗	✗
(1,0,0)	1	✓	✗
(1,0,1)	1	✗	✗
(1,1,0)	1	✓	✗
(1,1,1)	1	✓	✓

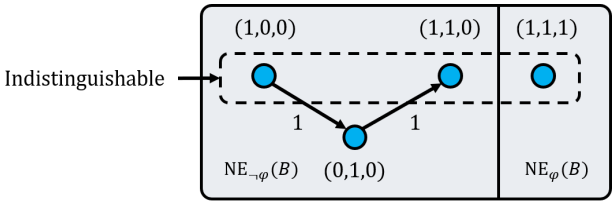

Figure 3.6: Example of a two-player costless game in which $\Phi = \{p_1, p_2, p_3\}$, $\Phi_1 = \{p_1, p_2\}$, $\Phi_2 = \{p_3\}$, $\gamma_1 = p_1 \vee p_2$, $\gamma_2 = p_3 \Rightarrow (p_1 \wedge p_2)$, $\mathcal{O} = \{p_1\}$, $o_0 = 0$, $o_1 = (1)$ and $\varphi = p_1 \wedge p_2 \wedge p_3$. The table on the top illustrates different properties of each strategy profile, and the diagram on the bottom depicts a deviation path which contains an observed deviation cycle.

in $\mathcal{G}(B, \mathcal{O})$; it simply needs to look like one from the principal's perspective. Given this, a deviation path $P = (\vec{v}^1, \vec{v}^2, \ldots, \vec{v}^k)$ is said to contain an *observed deviation cycle* if $\vec{v}^1 =_{\mathcal{O}} \vec{v}^j$, for some $j \in \{2, \ldots, k\}$. Terminology regarding the involvement of agents in observed deviation paths/cycles is extended in the natural way – an observed deviation path $P_{\mathcal{O}} = (\vec{o}^1, \vec{o}^2, \ldots, \vec{o}^k)$ *involves agent* i iff $\vec{v}^j \rightarrow_i \vec{v}^{j+1}$ for some $j \in \{1, \ldots, k-1\}$ such that $\vec{v}^j|_{\mathcal{O}} = \vec{o}^j$ and $\vec{v}^{j+1}|_{\mathcal{O}} = \vec{o}^{j+1}$.

Example 3.3. To illustrate some of the recently introduced concepts, consider the two-player game in Figure 3.6. The diagram at the bottom depicts a deviation path $P = ((1,0,0), (0,1,0), (1,1,0))$ consisting of three Nash equilibria of B. Note that P involves only one agent and is not a deviation cycle. Nevertheless, P contains an *observed* deviation cycle that involves only player 1, as the corresponding observed deviation path is $P_{\mathcal{O}} = (o_1, o_0, o_1)$. We will see shortly that this property is crucial in determining whether a set of strategy profiles can be eliminated or not.

Now, we present a characterisation of the eliminability of sets of initial equilibria in a game:

Theorem 3.7. *Let B be a Boolean game, $\mathcal{O}$ an observable set, and $X \subseteq \text{INIT}(B)$ a non-empty set of initial equilibria. Then, X is eliminable if and only if there exists a deviation graph $D = (\mathcal{V}_D, E_D)$ of B with respect to $\mathcal{O}$ that satisfies the following properties:*

- *For every $\vec{v} \in X$, there is some $\vec{v}' \in \mathcal{V}_D$ such that $(\vec{v}, \vec{v}') \in E_D$;*
- *Every observed deviation cycle in D involves at least two agents.*

Proof. For the forward direction, suppose that X is eliminable, and let κ be a contract that eliminates X. We construct a deviation graph D satisfying both properties. Since κ eliminates X, for every $\vec{v} \in X$, the profile $\vec{v}$ is not a Nash equilibrium of B^κ, so there is some agent $i_{\vec{v}} \in N$ and strategy $v'_{i_{\vec{v}}} \in V_{i_{\vec{v}}}$ such that $(\vec{v}_{-i_{\vec{v}}}, v'_{i_{\vec{v}}}) \succ^\kappa_{i_{\vec{v}}} \vec{v}$. Since this deviation is beneficial under κ, it is an inducible deviation and hence an edge of $\mathcal{G}(B, \mathcal{O})$. Define $D = (\mathcal{V}_D, E_D)$ by selecting, for each $\vec{v} \in X$, exactly one such edge, and let $\mathcal{V}_D$ consist of all endpoints of the selected edges. Then D is a subgraph of $\mathcal{G}(B, \mathcal{O})$ satisfying property (1) by construction.

It remains to show that D satisfies property (2). Observe that in D, the only vertices with outgoing edges are those in X, so in any path $(\vec{v}^1, \ldots, \vec{v}^k)$ within D, all vertices $\vec{v}^1, \ldots, \vec{v}^{k-1}$ must lie in X. Suppose for contradiction that there is an observed deviation cycle $P = (\vec{v}^1, \vec{v}^2, \ldots, \vec{v}^k)$ in D involving only one agent $i \in N$. Since each edge is a unilateral deviation by i, all profiles agree on the strategies of every other agent: $\vec{v}^j_{-i} = \vec{v}^1_{-i}$ for all $j \in \{1, \ldots, k\}$. In particular, $\vec{v}^k = (\vec{v}^1_{-i}, v^k_i)$ for some $v^k_i \in V_i$, so $\vec{v}^k$ is a unilateral deviation by i from $\vec{v}^1$. Now, $\vec{v}^1 \in X \subseteq \textsc{init}(B)$, so $\vec{v}^1$ is a Nash equilibrium of the costless game and hence $u_i(\vec{v}^1) \geq u_i(\vec{v}^k)$. Moreover, $\vec{v}^1 \in \textsc{init}(B)$ ensures that $i \in W(\vec{v}^k) \Rightarrow i \in W(\vec{v}^1)$. By definition of an observed deviation cycle, $\vec{v}^1 =_\mathcal{O} \vec{v}^k$, so $\kappa_i(\vec{v}^1|_\mathcal{O}) = \kappa_i(\vec{v}^k|_\mathcal{O})$. It follows that $u^\kappa_i(\vec{v}^k) \leq u^\kappa_i(\vec{v}^1)$: if i has the same winning status at both profiles, the contract payments cancel and the inequality reduces to the base-game condition; if i transitions from winner at $\vec{v}^1$ to loser at $\vec{v}^k$, the inequality is only strengthened. However, the chain of strict preferences along P yields $\vec{v}^k \succ^\kappa_i \vec{v}^{k-1} \succ^\kappa_i \cdots \succ^\kappa_i \vec{v}^1$, and by transitivity, $u^\kappa_i(\vec{v}^k) > u^\kappa_i(\vec{v}^1)$, a contradiction. Hence, D satisfies property (2).

For the backward direction, suppose that there exists a deviation graph D which satisfies both of the given properties. Then, let $\textsc{in}(\vec{o}, D)$ denote the set of agents $i \in N$ for which there is some $\vec{v} \in \textsc{cons}(B, \vec{o})$ and $\vec{v}' \in \mathcal{V}_D$ such that $(\vec{v}', \vec{v}) \in E_D$. For each $i \in N$, let ℓ_i be the length of the longest observed deviation path in D that involves only i. Additionally, for all $\vec{o} \in \mathbb{B}^\mathcal{O}$, let $d_i(\vec{o})$ denote the length of the longest observed deviation path in D *starting from* $\vec{o}$ that involves only i. Note that because no observed deviation cycle involves only one agent by assumption, ℓ_i is finite for all $i \in N$ and $\vec{o} \in \mathbb{B}^\mathcal{O}$. Thus, we can construct a contract κ as follows:

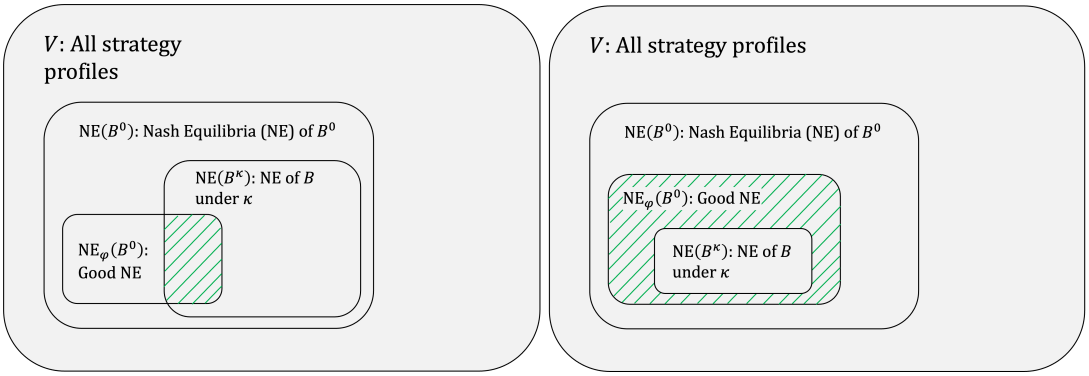

Figure 3.7: Venn Diagrams representing the nesting of sets of strategy profiles relevant to our analysis for the Nash Contractibility problems, which asks whether a contract exists that renders the depicted intersections non-empty. Left: Diagram depicting the relevant intersection for the E-Nash Contractibility problem. Right: Diagram depicting the relevant intersection for the A-Nash Contractibility problem.

$$\kappa_i(\vec{o}) = \begin{cases} (\ell_i - d_i(\vec{o}))(c_i^* + 1) & \text{if } i \in \text{IN}(\vec{o}, D) \\ 0 & \text{otherwise.} \end{cases}$$

Intuitively, κ is designed so that for every edge $(\vec{v}, \vec{v}') \in E_D$, the agent who has an inducible deviation from $\vec{v}$ to $\vec{v}'$ is offered a reward when $\vec{v}'|_{\mathcal{O}}$ is observed that is significantly greater than when $\vec{v}|_{\mathcal{O}}$ is observed. More precisely, observe first that for all $(\vec{v}, \vec{v}') \in E_D$, we have $d_i(\vec{v}|_{\mathcal{O}}) > d_i(\vec{v}'|_{\mathcal{O}})$ for the agent i such that $\vec{v} \rightarrow_i \vec{v}'$, because D is assumed not to contain any observed deviation cycles involving only one agent. Thus, κ offers i at least $c_i^* + 1$ more under $\vec{v}'|_{\mathcal{O}}$ than under $\vec{v}|_{\mathcal{O}}$. By the definition of inducible deviations, we can then conclude that for any $(\vec{v}, \vec{v}') \in E_D$, we have $\vec{v}' \succ_i^\kappa \vec{v}$. Then, by the first property, it holds that for every $\vec{v} \in X$, there is some $\vec{v}' \in V$ such that $\vec{v}' \succ_i^\kappa \vec{v}$ for some $i \in N$, so κ eliminates X. □

Contract Design for Logical Objectives

Returning to the problem of designing contracts for the guaranteed satisfaction of logical objectives, the natural question to ask is, given a Boolean game B, does there exist a contract κ such that the principal can ensure that their goal φ has been satisfied on some or every Nash equilibrium of the game B^κ? The E-Nash version of this problem is defined as follows:

E-NASH CONTRACTIBILITY:
Given: Game B, observable set $\mathcal{O}$, formula φ.
Question: Does there exist a contract κ such that for some $\vec{o} \in \mathbb{B}^{\mathcal{O}}$, we have $\text{ENV}(B^{\kappa}, \vec{o}, \varphi) \neq \emptyset$?

Note that for the problem of contract design, the principal is not given an observation to begin with, but must consider *all possible observations* that are consistent with at least one Nash equilibrium of a given Boolean game – since the agents' utilities are affected by the imposed contract, their strategies will likewise be influenced by the incentives introduced by the contract and hence, this component must be specified *before* agents in the game select their strategies. We have found that this problem is no harder than the special case where the principal has full observability within the game [224]:

Theorem 3.8. E-NASH CONTRACTIBILITY *is* Σ_2^p*-complete.*

Proof. For membership in Σ_2^p, first observe that the answer to E-Nash Contractibility is "yes" if and only if there exists some $\vec{v} \in \text{IND}(B, \mathcal{O})$ such that $\vec{v} \models \varphi$, that is, if there exists an inducible equilibrium that satisfies φ. Making use of the characterisation of inducible equilibria in Proposition 3.4, we can guess a strategy profile $\vec{v}$ and then check whether (1) $\vec{v} \in \text{INIT}(B)$ and $\vec{v} \models \varphi$; and (2) for all agents $i \in N$, and all choices $v_i' \in V_i$ such that $(\vec{v}_{-i}, v_i') =_{\mathcal{O}} \vec{v}$ and $i \in W(\vec{v}) \Leftrightarrow i \in W(\vec{v}_{-i}, v_i')$, we have that $c_i(\vec{v}_{-i}, v_i') \geq c_i(\vec{v})$. Checking the first condition amounts to verifying that $\vec{v}$ is a Nash equilibrium of the cost-free game B^0 and that $\vec{v} \models \varphi$, which is a coNP problem [224]. Moreover, checking the second condition is also in coNP – simply guess an alternative choice $v_i' \in V_i$ for each agent $i \in N$ and then checking whether $(\vec{v}_{-i}, v_i') =_{\mathcal{O}} \vec{v}$, $i \in W(\vec{v}) \Leftrightarrow i \in W(\vec{v}_{-i}, v_i')$, and $c_i(\vec{v}_{-i}, v_i') < c_i(\vec{v})$ can be done in polynomial time. If the answer is "yes", then condition 2 is not satisfied. Thus, the overall procedure is in the Σ_2^p complexity class.

Hardness follows again from the fact that the problem of checking whether or not a Nash equilibrium exists in a cost-free Boolean game is Σ_2^p-complete [224]. Given an instance of such a problem, we can check for E-Nash contractibility of the formula $\varphi = \top$ in the same game, where the principal is able to observe all of the variables (i.e., $\mathcal{O} = \Phi$). The answer to the E-Nash contractibility problem will be "yes" if and only if a Nash equilibrium exists in the corresponding cost-free Boolean game, by Proposition 3.2. □

Finally, we introduce and settle the complexity of the universal counterpart to the E-Nash Contractibility problem:

> A-NASH CONTRACTIBILITY:
> *Given*: Game B, observable set $\mathcal{O}$, formula φ.
> *Question*: Is there a contract κ s.t. $\text{NE}(B^\kappa) \neq \emptyset$ and for all observations $\vec{o} \in \mathbb{B}^{\mathcal{O}}$, if $\text{CONS}(B^\kappa, \vec{o}) \neq \emptyset$, then $\text{CONS}(B^\kappa, \vec{o}) \subseteq \text{NE}_\varphi(B^\kappa)$?

With this definition in place, we can then show the following result, which contrasts with Proposition 14 of [224], where the analogous problem was claimed to be Σ_2^p-complete.[3] The discrepancy arises because the earlier analysis did not account for the additional quantifier alternation introduced by the universal condition on observations: whereas the existential contractibility problem requires only that *some* observation yields a satisfying equilibrium, the universal version demands that *every* consistent observation does so, which adds a further layer of alternation and lifts the problem into Σ_3^p. This correction is not merely a technical adjustment; it establishes that universal contract design is genuinely harder than existential contract design, a separation that has direct consequences for the tractability of mechanism design in partially-observable Boolean games.

Theorem 3.9. A-NASH CONTRACTIBILITY *is* Σ_3^p*-complete.*

Proof. For membership, note that the A-Nash Contractibility problem is equivalent to deciding the following statement:

$$\exists \kappa \in \mathcal{K}, \exists \vec{v} \in V, \forall \vec{v}\,' \in V, (\vec{v} \in \text{NE}(B^\kappa)) \wedge (\vec{v}\,' \in \text{NE}(B^\kappa) \Rightarrow \vec{v}\,' \models \varphi),$$

which is a Σ_3^p predicate. This follows from two main observations: firstly, the contract κ constructed in the backward direction in the proof of Theorem 3.7 is guaranteed to eliminate $\text{NE}_{\neg\varphi}(B^\kappa)$ if φ is A-Nash contractible. The maximum payment awarded by such a contract is $\ell_i(c_i^* + 1)$, where we recall that ℓ_i is the length of the longest observed deviation path in D that involves only i. The length of such a path is bounded from above by the longest non-cyclic path through the set of all strategy profiles, which is $2^{|\Phi|}$. Thus, if φ is A-Nash contractible, then a contract can be designed to achieve this using payments of at most $2^{|\Phi|} \cdot \max_{i \in N}(c_i^* + 1)$, which can be represented using a polynomial number of bits. Thus, the size of each contract is polynomial in the size of the game, since the size of the original cost functions c_i are already exponential in $|\Phi|$. There are at

[3] Private communication with the authors has confirmed that there was an error in the original publication [224].

most an exponential number of contracts to check with this maximum value and hence, guessing a contract can be done in NP. Secondly, the problem of checking whether $\vec{v} \in \text{NE}(B^\kappa)$ for some $\vec{v} \in V$ and $\kappa \in \mathcal{K}$ is in fact coNP-complete. These combined give membership in Σ_3^p.

For hardness, we reduce from QSAT$_3$, which is known to be Σ_3^p-complete [263]. Suppose that we have an instance of QSAT$_3$, which is given by a Boolean formula φ over a set X of Boolean variables and a partition of X into three sets X_1, X_2, X_3. The question is to decide whether $\exists X_1 \forall X_2 \exists X_3\ \varphi$. We transform this into an instance of A-NASH CONTRACTIBILITY by defining a three-agent game $B = (\{1,2,3\}, \Phi, (\Phi_i)_{i\in N}, (\gamma_i)_{i\in N}, (c_i)_{i\in N})$, where: $\Phi = X \cup \{p, q\}$, $\Phi_1 = X_1, \Phi_2 = X_2 \cup \{p\}$, and $\Phi_3 = X_3 \cup \{q\}$; $\gamma_1 = \top$; $\gamma_2 = \neg\varphi \vee (p \leftrightarrow q)$; $\gamma_3 = \varphi \vee \neg(p \leftrightarrow q)$; for all $\vec{v} \in V$ and $i \in \{1,2\}$, we have $c_i(\vec{v}) = 0$; for all $\vec{v} \in V$, we have $c_3(\vec{v}) = 1$ if $\vec{v} \not\models \varphi$ and is 0 otherwise. The objective that the principal wishes to implement is simply φ, and the observable set is given by $\mathcal{O} = X_1$. We now show that the answer to an instance of QSAT$_3$ is "yes" if and only if the answer to A-NASH CONTRACTIBILITY in the above construction is also "yes."

Suppose that the answer to the instance of QSAT$_3$ is "yes" and let $v_1^* = X_1$ for some assignment X_1 that satisfies the QSAT$_3$ instance. Then, consider the contract κ such that for all $i \in N$ and $\vec{o} \in \mathcal{O}$:

$$\kappa_i(\vec{o}) = \begin{cases} c_1^* + 1 & \vec{o} = v_1^* \text{ and } i = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Intuitively, κ offers agent 1 a very high reward if they do choose v_1^* and offers no reward to agents 2 or 3. Let $\vec{v}$ be a strategy profile such that the following three conditions hold:

1. $v_1 = v_1^*$;
2. $\vec{v} \models \varphi$;
3. $p \leftrightarrow q$.

Such a strategy profile is guaranteed to exist under the assumption that we have a "yes" instance of QSAT$_3$. Then, it is easy to check that all agents achieve their goals and minimise their costs under κ, so there is no incentive to unilaterally deviate. Thus, $\vec{v} \in \text{NE}_\varphi(B^\kappa)$.

Now, suppose that $\vec{v}$ is a strategy profile that does not satisfy one of the three conditions. If $v_1 \neq v_1^*$, then agent 1 has a beneficial deviation to v_1^* to reduce their

cost. Thus, assume that $v_1 = v_1^*$ and suppose that $\vec{v} \not\models \varphi$. Then, because $v_1 = v_1^*$, no matter what strategy agent 2 chooses, agent 3 has a deviation that satisfies φ and will benefit from doing so because their cost is strictly lower when φ is satisfied than when it is not. Thus, agent 3 has a beneficial deviation to achieve their goal. Finally, if $\neg(p \leftrightarrow q)$ holds, then agent 2 has a beneficial deviation by setting p to the same value as q. Thus, any strategy profile that does not satisfy the three conditions is not a Nash equilibrium of B^κ. This means all Nash equilibria under B^κ satisfy φ, and hence, the answer to A-NASH CONTRACTIBILITY is "yes."

Suppose then, that the answer to the instance of QSAT_3 is "no." Then, it holds that $\forall X_1 \exists X_2 \forall X_3 \neg\varphi$. Let κ be any contract and let $\mathsf{resp}(v_1) = \{v_2 \in V_2 \mid \forall v_3 \in V_3\ (v_1, v_2, v_3) \not\models \varphi\}$. Then, there exists a strategy profile $\vec{v}$ satisfying the following three conditions:

1. $\kappa_1(\vec{v}\,|_{\mathcal{O}}) - c_1(\vec{v}\,) \geq \kappa_1((\vec{v}_{-1}, v_1')|_{\mathcal{O}}) - c_1(\vec{v}_{-1}, v_1')$, for all $v_1' \in V_1$;
2. $v_2 \in \mathsf{resp}(v_1)$;
3. $\vec{v} \models \neg(\varphi \vee (p \leftrightarrow q))$.

To see why $\vec{v}$ necessarily exists, firstly note that because the answer to this instance of QSAT_3 is "no", then it follows that $\mathsf{resp}(v_1) \neq \emptyset$ for all $v_1 \in V_1$ so condition 2 can always be satisfied for any strategy that agent 1 chooses. Next, observe that a strategy profile which satisfies condition 1 always exists, because V_1 is finite. Finally, it is clear that because $v_2 \in \mathsf{resp}(v_1)$, we have $\vec{v} \models \neg\varphi$ and no matter what value agent 3 assigns to q, agent 2 has an assignment to p so that $\vec{v} \models \neg(p \leftrightarrow q)$.

Now, under $\vec{v}$ and κ, agent 1 achieves the lowest cost attainable by the property 1 of $\vec{v}$ and trivially achieves their goal, so there is no incentive for them to deviate. Moreover, both agents 2 and 3 achieve their goals by property 3, so there is no incentive for them to deviate in order to achieve their goals. Thus, the only possible deviation that could occur from $\vec{v}$ is one by either agents 2 or 3 in order to reduce their costs, whilst retaining goal satisfaction. However, for any fixed v_1 and contract κ, all choices made by agents 2 and 3 will result in the same contract payoff to them, because none of their actions are observable by the principal. Additionally, agent 3 cannot deviate in order to satisfy φ and reduce their cost because $v_2 \in \mathsf{resp}(v_1)$. Hence, there is no incentive for either agent to deviate in pursuit of lower costs. Since no beneficial deviation exists from $\vec{v}$, we

have $\vec{v} \in \text{NE}_{\neg\varphi}(B^\kappa)$. Thus, for all κ, we have $\text{NE}_{\neg\varphi}(B^\kappa) \neq \emptyset$, so the answer to A-NASH CONTRACTIBILITY is "no." □

This result formally establishes the intuitive notion that the task of designing incentives to eliminate *all* undesirable behaviours in a multi-agent system is, in general, significantly more challenging than creating space for a desirable outcome to be chosen among many possible outcomes in the game.

3.5 Discussion

Boolean games provide a very natural model in which to draw a connection between two previously separate yet closely related areas of study: moral hazard problems (from economics) and rational verification (from AI verification). By extending the moral hazard problem to a mixed quantitative and qualitative setting through the use of Boolean variables and propositional logic goals, this framework provides a method for expressing relationships between discrete tasks which may require a threshold of resources (costs) to complete, rather than being tied to a continuous level of effort. This chapter develops this connection with a number of contributions: the formulation of a model of multi-agent moral hazard problems with combined qualitative and quantitative preferences, a characterisation of when equilibria can be induced or eliminated, and results settling the computational complexity of the verifiability and contractibility problems associated with this model of games. Ultimately, these results demonstrate that the contractibility problem under partial observability is no harder than the corresponding problems in the fully-observable setting and that the additional complexity of the universal contractibility problem arises from the difficulty of designing the contract as opposed to the problem of verifying the contract. Due to the combinatorial nature of these problems, it is natural that the problems studied in this chapter are generally computationally hard to solve, and additional assumptions such as those made in the combinatorial contracts model [190, 259] are required to make them tractable. However, carrying out this analysis nonetheless allows a relative comparison of the complexity of different components of the contract design problem – namely verification and contract design – which allows us to better understand where the source of the difficulty lies.

As this chapter and previous work on moral hazard problems demonstrate, the presence of hidden actions limits a principal's ability to design contracts

that successfully align agents' local incentives with the principal's higher-level objectives. This study presents a model which opens up many possible extensions and questions. Further work may refine the model in several ways by introducing an explicit *quantitative utility function for the principal* which is decreasing in contract payments, adding *randomness* to observations, allowing agents to *report* some of their actions, requiring that contracts be *individually rational*, and letting the principal *decide which observations* to make under the assumption that observations are costly. Such extensions would provide further insights into the computational aspects of incentive design in moral hazard settings, which are important concerns in AI research.

4

Incentive Engineering for Concurrent Games

In this chapter, we consider the problem of incentivising desirable behaviours in multi-agent systems by way of *taxation schemes* in a temporally extended setting. In the previous chapter, we studied how limitations in observability can constrain the decision problem of a contracting principal in the context of one-shot Boolean games. However, many real-world scenarios involve ongoing interactions where players make decisions over time that influence the state of the game. For example, consider a scenario where autonomous vehicles are navigating a traffic intersection. The autonomous vehicles must decide whether to proceed through the intersection or wait for other vehicles to clear the intersection. This decision is influenced by the state of the intersection, such as the presence of other vehicles, the presence of pedestrians, and the state of the traffic lights. Suppose that the goals of the autonomous agent's vehicle crossing the intersection are to: 1. Always avoid collisions, 2. Eventually reach its destination, and 3. Do so in the fastest amount of time. The incentive designer, on the other hand, might wish to coordinate the scheduling of the traffic lights to ensure that vehicles are never signalled to proceed through the intersection when it is unsafe to do so and to ensure that no vehicle has to wait for a certain number of cycles before proceeding through the intersection. The scheduling of the traffic lights could be set according to a fixed schedule, which might be simple to design but may fail to achieve the goal of ensuring effective flow of traffic through the

intersection. On the other hand, the scheduling could depend on the observed traffic conditions at the intersection, implementing a dynamic incentive scheme that, while potentially more complex to design, may enable the achievement of the desired goal where the static scheme might fail.

The dynamic nature of interactions between agents and the environment ultimately calls for the use of a richer model to describe these processes. For this reason, we extend the framework of [224] to the concurrent games model: in this model, each agent is primarily motivated to seek the satisfaction of a goal, expressed as a Linear Temporal Logic (LTL) formula; secondarily, agents seek to minimise costs, where costs are imposed based on the actions taken by agents in different states of the game. In this setting, we consider an external principal who can influence agents' preferences by imposing taxes (additional costs) on the actions chosen by agents in different states. The principal imposes taxation schemes to motivate agents to choose a course of action that will lead to the satisfaction of their goal, also expressed as an LTL formula. However, taxation schemes are limited in their ability to influence agents' preferences: an agent will always prefer to satisfy its goal rather than otherwise, no matter what the costs. The fundamental question that we study is whether the principal can impose a taxation scheme such that, in the resulting game, the principal's goal is satisfied in at least one or all runs of the game that could arise by agents choosing to follow game-theoretic equilibrium strategies. We consider two different types of taxation schemes: in a *static* scheme, the same tax is imposed on a state-action profile pair in all circumstances, while in a *dynamic* scheme, the principal can choose to vary taxes depending on the circumstances. We investigate the main game-theoretic properties of this model as well as the computational complexity of the relevant decision problems.

4.1 Related Work

We are particularly interested here in the question of how desirable properties can be incentivised, and undesirable behaviours eliminated. We take as our starting point the work of [224], who considered the possibility of influencing one-shot Boolean games by introducing taxation schemes, which impose additional costs onto a game at the level of individual actions. In the model of preferences considered in [224], agents are primarily motivated to achieve a goal expressed as a (propositional) logical formula, and only secondarily motivated to minimise costs. This limits the principal's ability to influence agent preferences: an agent

can never be motivated by a taxation scheme away from achieving its goal. Wooldridge et al. defined the following implementation problem [224]: given a game G and an objective φ, expressed as a propositional logic formula, does there exist a taxation scheme τ that could be imposed upon G such that, in the resulting game G^τ, the objective φ will be satisfied in at least one Nash equilibrium?

Here, we develop these ideas in the context of *concurrent games* [38]. Preferences in such games are defined by associating with each agent i a Linear Temporal Logic (LTL) goal γ_i, which agent i desires to see satisfied. In this work, we also assume that actions incur costs, and that agents seek to minimise their limit average costs. Since, in contrast to the model of [224], play in our games continues for an infinite number of rounds, we find there are two natural variations of taxation schemes for concurrent games. In a *static* taxation scheme, we impose a fixed cost on state-action profiles, so that the same state-action profile will always incur the same tax, no matter when it is performed. In a *dynamic* taxation scheme, the same state-action profile may incur different taxes in different circumstances. For both static and dynamic taxation schemes, we characterise the conditions under which an LTL objective φ can be implemented in a game.

4.2 Concurrent Games

Concurrent Game Arenas We work with concurrent game structures, which in this work we will refer to as *arenas* (to distinguish them from the game structures that we introduce later in this section) [38]. Formally, a concurrent game arena is given by a structure

$$\mathcal{A} = (S, N, A_1, \ldots, A_n, \mathcal{T}, \mathcal{C}, \mathcal{L}, s^0),$$

where:

- S is a finite and non-empty set of *arena states*;
- $N = \{1, \ldots, n\}$ is the set of *agents* – for any $i \in N$, we let $-i = N \setminus \{i\}$ denote the set of *all agents excluding* i;
- for each $i \in N$, A_i is the finite and non-empty set of unique actions available to agent i – we let $A = \bigcup_{i \in N} A_i$ denote the set of all *actions* and $\vec{A} = A_1 \times \cdots \times A_n$ denote the set of all *action profiles*;

- $\mathcal{T} : S \times A_1 \times \cdots \times A_n \to S$ is the *state transformer function* which prescribes how the state of the arena is updated for each possible action profile – we refer to a pair $(s, \vec{a})$, consisting of a state $s \in S$ and an action profile $\vec{a} \in \vec{A}$ as a *state-action profile*;
- $\mathcal{C} : S \times A_1 \times \cdots \times A_n \to \mathbb{R}^n_+$ is the *cost function* – given a state-action profile $(s, \vec{a})$ and an agent $i \in N$, we write $\mathcal{C}_i(s, \vec{a})$ for the i-th component of $\mathcal{C}(s, \vec{a})$, which corresponds to the cost that agent i incurs;
- $\mathcal{L} : S \to \mathbf{2}^{\Phi}$ is the *propositional labelling function* that specifies which propositional variables are true in each state $s \in S$; and
- $s^0 \in S$ is the *initial state* of the arena.

In what follows, we find it useful to define for every agent $i \in N$ the value c_i^* to be the maximum cost that agent i could incur through the performance of an individual state-action profile: $c_i^* = \max\{\mathcal{C}_i(s, \vec{a}) \mid s \in S, \vec{a} \in \vec{A}\}$.

Runs Games are played in an arena as follows. The arena begins in its initial state s^0, and each agent $i \in N$ selects an action $a_i \in A_i$ to perform; [1] the actions so selected define an *action profile*, $\vec{a} \in A_1 \times \cdots \times A_n$. The arena then transitions to a new state $s^1 = \mathcal{T}(s^0, a_1, \ldots, a_n)$. Each agent then selects another action $a_i' \in A_i$, and the arena again transitions to a new state $s^2 = \mathcal{T}(s^1, \vec{a}\,')$. In this way, we trace out an infinite interleaved sequence of states and action profiles, which we refer to as a *run*, ρ:

$$\rho : s^0 \xrightarrow{\vec{a}^0} s^1 \xrightarrow{\vec{a}^1} s^2 \xrightarrow{\vec{a}^2} \cdots$$

Where ρ is a run and $k \in \mathbb{N}$, we write $s(\rho, k)$ to denote the state indexed by k in ρ, so $s(\rho, 0)$ is the first state in ρ, $s(\rho, 1)$ is the second, and so on. In the same way, we denote the k-th action profile played in a run ρ by $\vec{a}\,(\rho, k-1)$ and to single out an individual agent i's k-th action, we write $a_i(\rho, k-1)$. Given this, we let the set of all possible runs in an arena $\mathcal{A}$ be $\mathbf{runs}(\mathcal{A}) = \{(s^0, \vec{a}^{\,0})(s^1, \vec{a}^{\,1})\ldots \mid \forall i \in \mathbb{N},\ s^{i+1} = \mathcal{T}(s^i, \vec{a}^{\,i})\}$. The sequence of states in a given run ρ induces a sequence of labels called a *trace*, written $\mathrm{tr}(\rho) = \mathcal{L}(s(\rho, 0))\mathcal{L}(s(\rho, 1)) \ldots \in (2^{\Phi})^{\omega}$.

[1] In this study, we assume that there is no agent $i \in N$ whose actions are inconsequential, in the sense that the output of the state transformer function $\mathcal{T}$ is always the same for any state $s \in S$, partial action profile $\vec{a}_{-i} \in \prod_{j \in N \setminus \{i\}} A_j$, and action $a_i \in A_i$.

It will also be convenient to refer to finite prefixes or infinite suffixes of a particular run. Given a run ρ, we write $\rho^k = (s^k, \vec{a}^{\,k})$ to denote its k-th state-action profile. Then, let $\rho^{j:k} = \rho^j \dots \rho^k$ denote the sequence of state-action profiles in the run ρ from the j-th round to the k-th round inclusive, where $j \leq k$. One distinguished case of this notation is the infinite suffix of a run ρ from a given position $k \in \mathbb{N}$, which we write as $\rho^{k:} = \rho^k \rho^{k+1} \dots$. Likewise, we write $s(\rho, j : k) = s(\rho, j) \dots s(\rho, k)$ and $s(\rho, k :) = s(\rho, k)s(\rho, k+1) \dots \in S^\omega$. Similarly, let $\vec{a}\,(\rho, j : k)$ and $\vec{a}\,(\rho, k :)$ be defined in the obvious manner.

Above, we defined the cost function $\mathcal{C}$ with respect to individual state-action profiles. In what follows, we find it useful to lift the cost function from individual state-action profiles to sequences of state-action profiles and runs. Since runs are infinite, simply summing costs is not appropriate: instead, we consider the cost of a run to be the *average* cost incurred by an agent i over the run; more precisely, we define the average cost incurred by agent i over the first t steps of the run ρ as $\mathcal{C}_i(\rho, 0 : t) = \frac{1}{t+1} \sum_{j=0}^{t} \mathcal{C}_i(\rho, j)$ for $t \geq 1$, where by $\mathcal{C}_i(\rho, j)$ we mean $\mathcal{C}_i(s(\rho, j), \vec{a}\,(\rho, j))$. Similarly, we define $\mathcal{C}_i(\rho, s : t) = \frac{1}{t-s+1} \sum_{j=s}^{t} \mathcal{C}_i(\rho, j)$ for $1 \leq s < t$. Finally, we define the cost incurred by an agent i over the run ρ, denoted $\mathcal{C}_i(\rho)$, as the *inferior limit of means*:

$$\mathcal{C}_i(\rho) = \liminf_{t \to \infty} \mathcal{C}_i(\rho, 0 : t) = \liminf_{t \to \infty} \frac{1}{t+1} \sum_{j=0}^{t} \mathcal{C}_i(\rho, j). \tag{4.1}$$

Alert readers will ask whether the value $\mathcal{C}_i(\rho)$ is guaranteed to be well defined. In our setting, not only is the limit inferior finite, but the limit of means in fact exists, as we now show:

Proposition 4.1. *For all $i \in N$, the sequence of averages $\mathcal{C}_i(\rho, 0 : t)$ converges as $t \to \infty$, and hence $\mathcal{C}_i(\rho)$ is equal to this limit.*

Proof. Since strategies and arenas are both modelled as finite-state machines, the composite state space of the arena and agent machines is finite. Consequently, every run ρ is ultimately periodic: it begins with a finite non-repeating prefix, followed by a sequence of state-action profiles that repeats infinitely often. Let ℓ denote the index at which the periodic part begins and let p denote the length of the repeating cycle, so that $\rho^j = \rho^{j+p}$ for all $j \geq \ell$. For t large enough that $t > \ell + p$, we may decompose the average cost as

$$\mathcal{C}_i(\rho, 0 : t) = \frac{1}{t+1} \sum_{j=0}^{\ell-1} \mathcal{C}_i(\rho, j) + \frac{1}{t+1} \sum_{j=\ell}^{t} \mathcal{C}_i(\rho, j).$$

The first sum is a fixed constant divided by $t+1$, and hence converges to 0. For the second sum, the periodicity of costs after index ℓ implies that as $t \to \infty$, the average converges to the mean cost over one period, namely $\frac{1}{p}\sum_{j=\ell}^{\ell+p-1} \mathcal{C}_i(\rho, j)$. Therefore $\mathcal{C}_i(\rho, 0:t)$ converges, and $\mathcal{C}_i(\rho)$ equals this limit. □

Strategies We model strategies for agents as finite state machines with output. Formally, a strategy σ_i for agent $i \in N$ is given by a structure

$$\sigma_i = (Q_i, \mathsf{next}_i, \mathsf{do}_i, q_i^0),$$

where:

- Q_i is a finite set of machine states;
- $\mathsf{next}_i : Q_i \times A_1 \times \cdots \times A_n \to Q_i$ is the machine's state transformer function;
- $\mathsf{do}_i : Q_i \to A_i$ is the machine's action selection function; and
- $q_i^0 \in Q_i$ is the machine's initial state.

A collection of strategies, one for each agent $i \in N$, is a *strategy profile*: $\vec{\sigma} = (\sigma_1, \ldots, \sigma_n)$. A strategy profile $\vec{\sigma}$ enacted in an arena $\mathcal{A}$ will generate a unique run, which we denote by $\rho(\vec{\sigma}, \mathcal{A})$; the formal definition is standard, and we will omit it here [121]. Where $\mathcal{A}$ is clear from context we will simply write $\rho(\vec{\sigma})$. For each agent $i \in N$, we write Σ_i for the set of all possible strategies for the agent and $\Sigma = \Sigma_1 \times \cdots \times \Sigma_n$ for the set of all possible strategy profiles. For a set of distinct agents $A \subseteq N$, we write $\Sigma_A = \prod_{i \in A} \Sigma_i$ for the set of partial strategy profiles available to the group A and $\Sigma_{-A} = \prod_{i \in N \setminus A} \Sigma_i$ for the set of partial strategy profiles available to the set of all agents excluding those in A.

Where $\vec{\sigma} = (\sigma_1, \ldots, \sigma_i, \ldots, \sigma_n)$ is a strategy profile and σ_i' is a strategy for agent i, we denote the strategy profile obtained by replacing the i-th component of $\vec{\sigma}$ with σ_i' by $(\vec{\sigma}_{-i}, \sigma_i')$. Similarly, given a strategy profile $\vec{\sigma}$ and a set of agents $A \subseteq N$, we write $\vec{\sigma}_A = (\sigma_i)_{i \in A}$ to denote a partial strategy profile for the agents in A and if $\vec{\sigma}'_A \in \Sigma_A$ is another partial strategy profile for A, we write $(\vec{\sigma}_{-A}, \vec{\sigma}'_A)$ for the strategy profile obtained by replacing $\vec{\sigma}_A$ in $\vec{\sigma}$ with $\vec{\sigma}'_A$.

Games, Utilities, and Preferences We obtain a *concurrent game* from an arena $\mathcal{A}$ by associating with each agent i a goal γ_i, represented as an LTL formula. Formally, a concurrent game $\mathcal{G}$ is given by a structure

$$\mathcal{G} = (S, N, (A_i)_{i \in N}, \mathcal{T}, \mathcal{C}, \mathcal{L}, s^0, (\gamma_i)_{i \in N}),$$

where $(S, N, (A_i)_{i \in N}, \mathcal{T}, \mathcal{C}, \mathcal{L}, s^0)$ is a concurrent game arena, and γ_i is the LTL goal of agent i, for each $i \in N$. Runs in a concurrent game $\mathcal{G}$ are defined over the game's arena $\mathcal{A}$ and hence, we use the notations $\rho(\vec{\sigma}, \mathcal{G})$ and $\rho(\vec{\sigma}, \mathcal{A})$ interchangeably. When the game or arena is clear from the context, we omit the $\mathcal{G}$ and simply write $\rho(\vec{\sigma})$. Given a strategy profile $\vec{\sigma}$, the generated run $\rho(\vec{\sigma})$ will satisfy the goals of some agents and not satisfy the goals of others, that is, there will be a set $W(\vec{\sigma}) = \{i \in N : \rho(\vec{\sigma}) \models \gamma_i\}$ of *winners* and a set $L(\vec{\sigma}) = N \setminus W(\vec{\sigma})$ of *losers*.

We are now ready to define preferences for agents. Our basic idea is that, as in the previous chapter and in [224], agents' preferences are structured: they first desire to accomplish their goal, and secondarily desire to minimise their costs. To capture this idea, it is convenient to define preferences via utility functions u_i over runs, where i's utility for a run ρ is

$$u_i(\rho) = \begin{cases} 1 + c_i^* - \mathcal{C}_i(\rho) & \text{if } \rho \models \gamma_i \\ -\mathcal{C}_i(\rho) & \text{otherwise.} \end{cases}$$

Defined in this way, if an agent i gets her goal achieved, her utility will lie in the range $[1, c_i^* + 1]$[2] (depending on the cost she incurs), whereas if her goal is not achieved, then her utility will lie in the range $[-c_i^*, 0]$. Thus, if an agent gets her goal achieved, then she will always receive utility strictly greater than 0, while if she fails to get her goal achieved, then she will receive utility at most 0.

Preference relations $\succeq_i$ over runs are then defined in the obvious way: $\rho_1 \succeq_i \rho_2$ if and only if $u_i(\rho_1) \geq u_i(\rho_2)$, with indifference relations $\sim_i$ and strict preference relations $\succ_i$ defined in the usual way.

Nash equilibrium We work with the solution concept of pure strategy Nash equilibria in this chapter. Given a game $\mathcal{G}$, let $\text{NE}(\mathcal{G})$ denote its set of Nash

[2]When we introduce taxation schemes later, c_i^* will be taken to be the largest incurrable cost in the game under the given taxation scheme.

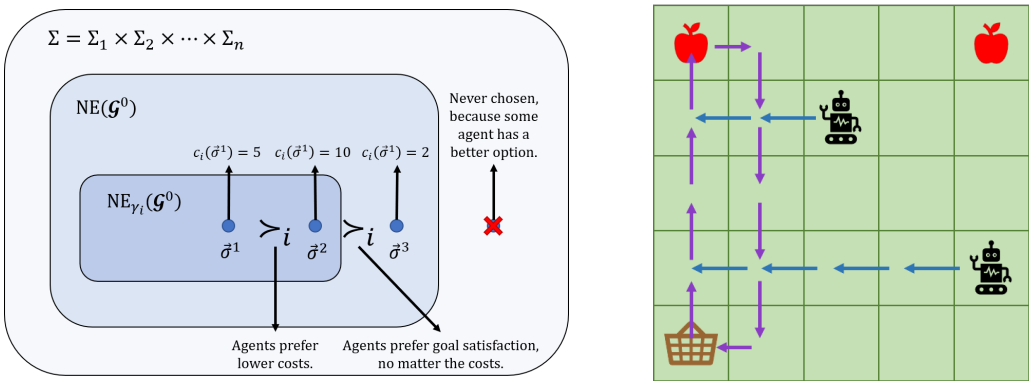

Figure 4.1: (a) Illustration of the lexicographic quantitative and qualitative preferences of agents. (b) A concurrent game where two robots are situated in a grid world and are programmed to (1) never crash into another robot and (2) to secondarily minimise their limit-average costs. The arrows indicate how a run may be decomposed into a non-repeating and an infinitely-repeating component.

equilibria.[3] Where φ is an LTL formula, we find it useful to define $\mathrm{NE}_\varphi(\mathcal{G})$ to be the set of Nash equilibrium strategy profiles that result in φ being satisfied:

$$\mathrm{NE}_\varphi(\mathcal{G}) = \{\vec{\sigma} \in \mathrm{NE}(\mathcal{G}) \mid \rho(\vec{\sigma}) \models \varphi\}.$$

It is sometimes useful to consider a concurrent game that is modified so that no costs are incurred in it. We call such a game a *cost-free game*. Where $\mathcal{G}$ is a game, let $\mathcal{G}^{\mathbf{0}}$ denote the game that is the same as $\mathcal{G}$ except that the cost function $\mathcal{C}^{\mathbf{0}}$ of $\mathcal{G}^{\mathbf{0}}$ is such that $\mathcal{C}^{\mathbf{0}}_i(s, \vec{a}) = 0$ for all $i \in N$, $s \in S$, and $\vec{a} \in \vec{A}$. The following result is readily established by direct application of the results in [121]:

Theorem 4.2. *Given a game $\mathcal{G}$, the problem of checking whether $\mathrm{NE}(\mathcal{G}^{\mathbf{0}}) \neq \emptyset$ is* 2ExpTime*-complete.*

Deviations and Distinguishability: The notion of Nash equilibrium is closely related to the concept of beneficial deviations. Given the way that preferences are defined in this study, it will be useful to categorise the different kinds of deviation that may be available to an agent [226]. In particular, in the same manner as the previous chapter, we will distinguish between initial, soft, and hard deviations. Firstly, given a game $\mathcal{G}$, we say that a strategy profile $\vec{\sigma}' \in \Sigma$ is *distinguishable*

[3] In general, Nash equilibria in this model of concurrent games may require agents to play infinite memory strategies [126]. These do not affect the results in the present study, although they can be used to the principal's advantage in closely related problems not studied here.

from another strategy profile $\vec{\sigma}^{\,2} \in \Sigma$ if $\rho(\vec{\sigma}^{\,1}, \mathcal{G}) \neq \rho(\vec{\sigma}^{\,2}, \mathcal{G})$. Then, for an agent i, a strategy profile $\vec{\sigma}$, and an alternative strategy $\sigma_i' \neq \sigma_i$, we say that:

1. σ_i' is an *initial deviation* for i from $\vec{\sigma}$, written $\vec{\sigma} \rightarrow_i (\vec{\sigma}_{-i}, \sigma_i')$, if $i \in W(\vec{\sigma}) \Rightarrow i \in W(\vec{\sigma}_{-i}, \sigma_i')$ and $\vec{\sigma}$ is distinguishable from $(\vec{\sigma}_{-i}, \sigma_i')$.
2. σ_i' is a *soft deviation* for i from $\vec{\sigma}$, written $\vec{\sigma} \leftrightarrows_i (\vec{\sigma}_{-i}, \sigma_i')$, if $i \in W(\vec{\sigma}) \iff i \in W(\vec{\sigma}_{-i}, \sigma_i')$ and $\vec{\sigma}$ is distinguishable from $(\vec{\sigma}_{-i}, \sigma_i')$.
3. σ_i' is a *hard deviation* for i from $\vec{\sigma}$, written $\vec{\sigma} \rightrightarrows_i (\vec{\sigma}_{-i}, \sigma_i')$, if $\vec{\sigma} \rightarrow_i (\vec{\sigma}_{-i}, \sigma_i')$ but not $(\vec{\sigma}_{-i}, \sigma_i') \rightarrow_i \vec{\sigma}$.

Note that the requirement that the pairs of strategy profiles be distinguishable does not appear in the definition given in [226], as in their case it is assumed that if a player makes a unilateral deviation to a different strategy, then the two must necessarily be distinguishable. In our case however, two strategy profiles may be different, yet induce the same run in a game. In such a situation, there is no way for a taxation scheme to distinguish between the two and impose taxes to make one more preferable than the other. Indeed, if a pair of strategy profiles $\vec{\sigma}, (\vec{\sigma}_{-i}, \sigma_i') \in \Sigma$ are *not* distinguishable, then there is no taxation scheme that can induce agent i to deviate from one to the other.

Also worth mentioning is the fact that just because a strategy σ_i' is an *initial* deviation for i from a strategy profile $\vec{\sigma}$, it is not necessarily the case that σ_i' is a *beneficial* deviation for i from $\vec{\sigma}$, because i's preference also depends on the cost that she will incur under both strategy profiles. Thus, the ability to manipulate these cost structures affords the possibility of converting initial deviations into beneficial deviations for a given agent i. The behaviours which can be incentivised or disincentivised under such manipulations is the focus of this chapter.

4.3 Static Taxation Schemes

We now introduce a model of incentives for concurrent games. For incentives to work, they clearly must appeal to an agent's preferences $\succeq_i$. As we saw above, incentives for our games are defined with respect to both goals and costs: an agent's primary desire is to see their goal achieved – the desire to minimise costs is strictly secondary to this. We will assume that we have no ability to change

the goals of agents: they are assumed to be fixed and immutable. It follows that any incentives we offer an agent to alter their behaviour must appeal to the costs incurred by that agent. Our basic model of incentives assumes that we can alter the cost structure of a game by imposing *taxes* which depend on the collective actions that agents choose in different states: taxes may increase an agent's costs, and so influence their preferences and rational choices. This model of costs and taxation naturally captures the context-dependence as well as the inherent limitations of real-world taxation schemes in influencing agents' behaviours. Formally, we model static taxation schemes as functions

$$\tau : S \times \vec{A} \to \mathbb{R}^n_+.$$

A static taxation scheme τ imposed on a game $\mathcal{G}$ will result in a new game, denoted by $\mathcal{G}^\tau = (S, N, (A_i)_{i\in N}, \mathcal{T}, \mathcal{C}^\tau, \mathcal{L}, s^0, (\gamma_i)_{i\in N})$, which is the same as $\mathcal{G}$ except that the cost function $\mathcal{C}^\tau$ of $\mathcal{G}^\tau$ is defined as

$$\mathcal{C}^\tau(s, \vec{a}\,) = \mathcal{C}(s, \vec{a}\,) + \tau(s, \vec{a}\,).$$

Similarly, we write $\mathcal{A}^\tau$ to denote the arena with modified cost function $\mathcal{C}^\tau$ associated with the game $\mathcal{G}^\tau$ and $u_i^\tau(\rho)$ to denote the utility function of agent i over run ρ with the modified cost function $\mathcal{C}^\tau$. Given a concurrent game $\mathcal{G}$ and a taxation scheme τ, we write $\rho_1 \succeq_i^\tau \rho_2$ iff $u_i^\tau(\rho_1) \geq u_i^\tau(\rho_2)$. The indifference relations $\sim_i^\tau$ and strict preference relations $\succ_i^\tau$ are defined analogously.

Example 4.1. Two robots are situated in a grid world (Figure 4.1b), where atomic propositions represent events where a robot picks up an apple (label a_{ij} represents agent i picking up apple j), has delivered an apple to the basket (label b_i represents agent i delivering an apple to the basket), or where the robots have crashed into each other (label c). Additionally, suppose that both robots are programmed with LTL goals $\gamma_1 = \gamma_2 = \mathbf{G}\neg c$. In this way, the robots are not pre-programmed to perform specific tasks, and it is therefore the duty of the *principal* to design taxes that motivate the robots to perform a desired function, e.g., pick apples and deliver them to the basket quickly. Because the game is initially costless, there is an infinite number of Nash equilibria that could arise from this scenario and it is by no means obvious that the robots will choose one in which they perform the desired function. Hence, the principal may attempt to design a taxation scheme to eliminate those that do not achieve their objective, thus motivating the robots to collect apples and deliver them to the basket. Clearly, using dynamic taxation schemes affords the principal more control over how the robots should accomplish this than static taxation schemes.

To begin our study of static taxation schemes, we observe that one important feature of taxation schemes in this setting is that *they are only effective on infinitely repeating action profile sequences of any run* ρ. Since strategies and arenas are both modelled as finite state machines, ρ will begin with a finite non-repeating sequence of states and actions, followed by a sequence of arena states and actions that repeats infinitely often. The vector $\mathcal{C}(\rho)$ will then contain the average cost of state-action profiles over the finite repeating sequence for each player.

We denote the sequence of state-action profiles in the run ρ which is repeated infinitely often by $\mathrm{Inf}(\rho) = \rho^{\ell_\rho : m_\rho - 1}$, where ℓ_ρ is the smallest integer such that $(\mathrm{Inf}(\rho))^\omega = \rho^{\ell_\rho :}$ and m_ρ is the first integer after ℓ_ρ such that $s(\rho, \ell_\rho) = s(\rho, m_\rho)$ and $q_i^{\ell_\rho} = q_i^{m_\rho}$ for all $i \in N$. Additionally, we write $(s, \vec{a}\,) \in \mathrm{Inf}(\rho)$ if there exists a $j \in \{\ell_\rho, \ldots, m_\rho - 1\}$ such that $\rho^j = (s, \vec{a}\,)$. Similarly, if $\ell_\rho > 0$, let $\mathrm{Fin}(\rho) = \rho^{0:\ell_\rho - 1}$ denote the finite sequence of state-action profiles which precedes the ultimately repeating component $\mathrm{Inf}(\rho)$ on the run ρ, or the empty word if $\ell_\rho = 0$. With this, we show that a taxation scheme only affects the utilities of agents in a game under a given strategy profile if it taxes state-action profiles which are played infinitely often:

Proposition 4.3. *Let $\mathcal{G}$ be a game, $\vec{\sigma}$ be a strategy profile, and τ, τ' be two taxation schemes such that $\tau \neq \tau'$ only on state-action profiles in $(Inf(\rho(\vec{\sigma}\,)))$. For any agent $i \in N$, we have that $u_i^\tau(\rho(\vec{\sigma}\,, \mathcal{G}^\tau)) \geq u_i^{\tau'}(\rho(\vec{\sigma}\,, \mathcal{G}^{\tau'}))$ if and only if it holds that*

$$\sum_{j=\ell_\rho}^{m_\rho - 1} \mathcal{C}_i^\tau(\rho, j) \leq \sum_{j=\ell_\rho}^{m_\rho} \mathcal{C}_i^{\tau'}(\rho, j).$$

Proof. Because utilities are computed with respect to the same run ρ, the choice of taxation scheme does not have any bearing on goal satisfaction for the agents, so it suffices to compare mean-payoffs under the two taxation schemes. In other words, it holds that $\rho(\vec{\sigma}\,, \mathcal{G}^\tau) \succeq_i \rho(\vec{\sigma}\,, \mathcal{G}^{\tau'})$ if and only if $\mathcal{C}_i^\tau(\rho) \leq \mathcal{C}_i^{\tau'}(\rho)$. We now show that this condition can be verified by computing the two sums given in the proposition.

Observe that for a large enough $t \in \mathbb{N}$, we can write the average cost of ρ over the first t steps under the taxation scheme τ as $\mathcal{C}_i^\tau(\rho, 0 : t) = \frac{1}{t} \sum_{j=0}^{\ell_\rho - 1} \mathcal{C}_i^\tau(\rho, j) + \frac{1}{t} \sum_{j=\ell_\rho}^{t} \mathcal{C}_i^\tau(\rho, j)$. The first sum represents the cost to i of the maximal finite non-repeating sequence of state-action profile pairs taken by the agents under τ and converges to 0 as $t \to \infty$. For the second sum, observe that the average cost to i over k loops of the infinitely repeating sequence $\mathrm{Inf}(\rho)$ under τ is $\frac{1}{m_\rho - \ell_\rho} \sum_{j=\ell_\rho}^{m_\rho} \mathcal{C}_i^\tau(\rho, j)$ which does not depend on k. Therefore, the second sum

in our decomposition of $\mathcal{C}_i^\tau(\rho, 0 : t)$ converges to this quantity as $t \to \infty$, so putting the two together, we get $\mathcal{C}_i^\tau(\rho) = \frac{1}{m_\rho - \ell_\rho} \sum_{j=\ell_\rho}^{m_\rho} \mathcal{C}_i^\tau(\rho, j)$. By a basic property of the limit inferior, it follows that

$$\mathcal{C}_i^\tau(\rho) \leq \mathcal{C}_i^{\tau'}(\rho) \iff \sum_{j=\ell_\rho}^{m_\rho} \mathcal{C}_i^\tau(\rho, j) \leq \sum_{j=\ell_\rho}^{m_\rho} \mathcal{C}_i^{\tau'}(\rho, j).$$

□

The implication of this is that if one is to devise taxation schemes to elicit certain behaviours from the agents, it suffices to incentivise only recurring behaviour. Given an arena $\mathcal{A}$ and strategy profile $\vec{\sigma}$, we can compute $\mathcal{C}(\rho(\vec{\sigma}, \mathcal{A}))$ in exponential time with respect to the number of agents n, since the number of arena/agent configurations is $O(|S \times \mathcal{A} \times Q|)$.

Now, a natural question to consider is whether given two different strategy profiles $\vec{\sigma}, \vec{\sigma}'$, there exists a taxation scheme which renders one of the strategies more preferable than the other. We will see shortly that the answer to this question depends on both the goal-satisfaction properties of the two strategies and the ultimately repeating action sequences they induce. Before proceeding, it will be useful to introduce some notation which will recur often in our analysis of static taxation schemes.

For a game $\mathcal{G}$ and any agent i, let $\kappa((s, \vec{a}), \vec{\sigma})$ be the number of times that the state-action profile $(s, \vec{a}) \in S \times \vec{A}$ appears in the infinite sequence of state-action profiles $\text{Inf}(\rho(\vec{\sigma}))$ and let $r((s, \vec{a}), \vec{\sigma}) = \kappa((s, \vec{a}), \vec{\sigma}) / (\sum_{\beta \in \text{Inf}(\rho(\vec{\sigma}))} \kappa(\beta, \vec{\sigma}))$ be the proportion of times that $(s, \vec{a})$ is played in $\text{Inf}(\rho(\vec{\sigma}))$. Then, given two strategy profiles $\vec{\sigma}_1$ and $\vec{\sigma}_2$, let $\Delta((s, \vec{a}), \vec{\sigma}_1, \vec{\sigma}_2) = r((s, \vec{a}), \vec{\sigma}_1) - r((s, \vec{a}), \vec{\sigma}_2)$ be the difference in the ratios $r((s, \vec{a}), \vec{\sigma}_1)$ and $r((s, \vec{a}), \vec{\sigma}_2)$ for a state-action profile $(s, \vec{a})$. Given this, we say that a strategy profile $\vec{\sigma}$ is *minimal* for i if for all other strategies $\sigma_i' \in \Sigma_i \setminus \{\sigma_i\}$ such that $i \in W(\vec{\sigma}) \iff i \in W(\vec{\sigma}_{-i}, \sigma_i')$, and for all $(s, \vec{a}) \in \text{Inf}(\rho(\vec{\sigma}))$, we have $\Delta((s, \vec{a}), \vec{\sigma}, (\vec{\sigma}_{-i}, \sigma_i')) = 0$.

Intuitively, a strategy profile is minimal for an agent i if each state-action profile that is played in the infinitely repeating sequence $\text{Inf}(\rho(\vec{\sigma}))$ is played in the same ratio among all possible deviations $(\vec{\sigma}_{-i}, \sigma_i')$ for that agent which lead to the same goal satisfaction outcome for them as $\vec{\sigma}$ does. The lemma below provides some intuition on why the notion of minimality is important in static taxation schemes:

Lemma 4.4. *For all games $\mathcal{G}$, any agent $i \in N$, and any two strategy profiles $\vec{\sigma}_1$ and $\vec{\sigma}_2$ such that $\rho(\vec{\sigma}_1) \succeq_i \rho(\vec{\sigma}_2)$, there exists a static taxation scheme $\tau : S \times \vec{A} \to \mathbb{R}^n_+$ such that $\rho(\vec{\sigma}_2) \succ_i^\tau \rho(\vec{\sigma}_1)$ if and only if the following conditions hold:*

1. $i \in W(\vec{\sigma}_1) \iff i \in W(\vec{\sigma}_2)$;
2. *There exists at least one state-action profile $(s, \vec{a}) \in Inf(\rho(\vec{\sigma}_1))$ such that*

$$\Delta((s, \vec{a}), \vec{\sigma}_1, \vec{\sigma}_2) \neq 0.$$

Proof. For the forward implication, assume that a static taxation scheme τ exists such that $\rho(\vec{\sigma}_2) \succ_i^\tau \rho(\vec{\sigma}_1)$ and that either $i \in W(\vec{\sigma}_1) \cap L(\vec{\sigma}_2)$ or $i \in W(\vec{\sigma}_1) \iff i \in W(\vec{\sigma}_2)$ and that for all state-action profile pairs $(s, \vec{a}) \in \text{Inf}(\rho(\vec{\sigma}_1))$, it holds that $\Delta((s, \vec{a}), \vec{\sigma}_1, \vec{\sigma}_2) \neq 0$.[4] Under the first possibility that $i \in W(\vec{\sigma}_1) \cap L(\vec{\sigma}_2)$, we immediately see a contradiction with the assumption because the utility of agent i under strategy profile $\vec{\sigma}_1$ is strictly positive whereas their utility under $\vec{\sigma}_2$ is strictly negative under any taxation scheme.

Under the second possibility, we first observe that because γ_i is satisfied under both strategies, the difference in utilities must arise from mean-payoff considerations. Hence, the only way to raise the utility (to agent i) of playing $\vec{\sigma}_2$ above $\vec{\sigma}_1$ is by raising the cost of the latter strategy profile above the former's. However, this is impossible in this case because any state-action profile pair played on the ultimately repeating state-action profile sequence component $\text{Inf}(\rho(\vec{\sigma}_1))$ is played in exactly the same ratio in $\text{Inf}(\rho(\vec{\sigma}_2))$ and thus, any static taxation scheme which raises i's cost under $\vec{\sigma}_1$ also raises her cost under $\vec{\sigma}_2$ by at least the same amount. This is again a contradiction with the assumption that such a taxation scheme exists.

For the reverse implication, suppose that $i \in W(\vec{\sigma}_1) \iff i \in W(\vec{\sigma}_2)$, and there exists at least one $(s, \vec{a}) \in \text{Inf}(\rho(\vec{\sigma}_1))$ such that $\Delta((s, \vec{a}), \vec{\sigma}_1, \vec{\sigma}_2) \neq 0$. As noted earlier, the difference in utilities in this situation arises solely due to mean-payoff considerations, so we will construct a taxation scheme to satisfy the required inequality. There are, in general, many ways to do this. We present one that distributes the taxes evenly amongst the state-action profile pairs $(s, \vec{a}) \in \text{Inf}(\rho(\vec{\sigma}_1))$ for which $\Delta((s, \vec{a}), \vec{\sigma}_1, \vec{\sigma}_2) \neq 0$. To do this, we compute the difference in cost between the two runs and then raise the cost to i of $\rho(\vec{\sigma}_1)$ by a sufficient amount so that its cost is higher than $\rho(\vec{\sigma}_2)$. To this end, we will utilise the quantity $\Delta((s, \vec{a}), \vec{\sigma}_1, \vec{\sigma}_2)$. This quantity will tell us how much weighting we

[4] The case where $i \in L(\vec{\sigma}_1) \cap W(\vec{\sigma}_2)$ is not possible under the assumption that $\rho(\vec{\sigma}_1) \succeq_i \rho(\vec{\sigma}_2)$.

should place on a particular state-action profile when taxing it so that the taxation scheme is sure to recover the difference in cost (and hence utility) between the two runs. With this, we then let the taxation scheme τ be such that for all $(s, \vec{a}) \in S \times \vec{A}$ where $\Delta((s, \vec{a}), \vec{\sigma}_1, \vec{\sigma}_2) < 0$, we have

$$\tau_i(s, \vec{a}) = (c_i^* + 1) \left(\sum_{j=1}^{2} \sum_{\beta \in \text{Inf}(\rho(\vec{\sigma}_j))} \kappa(\beta, \vec{\sigma}_j) \right).$$

For all $(s, \vec{a})$ where $\Delta((s, \vec{a}), \vec{\sigma}_1, \vec{\sigma}_2) \geq 0$, we simply let $\tau_i(s, \vec{a}) = 0$. Finally, for all other agents $j \neq i$, let τ_j be 0 for all state-action profile pairs. Note that under condition 2, there is always some $(s, \vec{a}) \in \text{Inf}(\rho(\vec{\sigma}_1))$ for which $\Delta((s, \vec{a}), \vec{\sigma}_1, \vec{\sigma}_2) < 0$, otherwise $\sum_{\beta \in \text{Inf}(\rho(\vec{\sigma}_1))} r(\beta, \vec{\sigma}_1) \neq 1$. Thus, the taxation scheme τ is well-defined and positive under the assumptions and condition 2. Moreover, under this taxation scheme, the difference in utility between the two strategies $\vec{\sigma_1}$ and $\vec{\sigma}_2$ is guaranteed to be exceeded and thus, $\rho(\vec{\sigma}_2) \succ_i^\tau \rho(\vec{\sigma}_1)$. □

This result not only tells us when an effective taxation scheme exists that would make one strategy profile more appealing over another, but also gives us insight into how we can construct alternative strategy profiles and taxation schemes to "eliminate" a strategy profile $\vec{\sigma}$ as a rational choice for a given agent.

As we saw above in Lemma 4.4, the ability to design effective taxation schemes to make a strategy $\vec{\sigma}_2$ weakly preferable to another strategy $\vec{\sigma}_1$ for an agent i relies on the existence of a state-action profile which can be taxed so as to increase the costs incurred under $\vec{\sigma}_1$ by a strictly greater amount than under $\vec{\sigma}_2$. Moreover, the existence of such a pair is determined by whether or not some $(s, \vec{a})$ is played more times in the ultimately repeating component $\text{Inf}(\rho(\vec{\sigma}_1))$ than in $\text{Inf}(\rho(\vec{\sigma}_2))$. The following corollary applies Lemma 4.4 to minimal strategies:

Corollary 4.5. *Let $\mathcal{G}$ be a game and $\vec{\sigma}$ be a strategy profile which is not minimal for some agent $i \in N$. There exists a static taxation scheme and an alternative strategy $\sigma_i' \in \Sigma_i \setminus \{\sigma_i\}$ such that $\rho(\vec{\sigma}_{-i}, \sigma_i') \succ_i^\tau \rho(\vec{\sigma})$.*

Hard and Soft Equilibria

Due to the quasi-dichotomous nature of preferences, we may classify Nash equilibria based on whether there exist taxation schemes which can eliminate them. This may be of use if a particular undesirable Nash equilibrium is identified and the principal would like to determine whether this equilibrium can be

$$\begin{array}{rcl} \textsc{Init}(G) & = & \textsc{ne}(\mathcal{G}^{\mathbf{0}}) \\ \textsc{Hard}(\mathcal{G}) & = & \bigcap_{\tau : S \times \vec{A} \to \mathbb{R}^n_+} \textsc{ne}(\mathcal{G}^\tau) \\ \textsc{Soft}(\mathcal{G}) & = & \textsc{Init}(\mathcal{G}) \setminus \textsc{Hard}(\mathcal{G}) \end{array}$$

Figure 4.2: Varieties of equilibria.

eliminated through the choice of an appropriate taxation scheme. In this section, following [225], we will address this problem by partitioning the set of initial equilibria and characterising these partitions in terms of properties of the game and the agents' goals.

Let us define the set $\textsc{Init}(\mathcal{G})$ of *initial equilibria* of a game $\mathcal{G}$ to consist of those equilibria that are present in the game $\mathcal{G}^{\mathbf{0}}$, i.e., the set of Nash equilibria that arise in the absence of any cost considerations. Then, the set $\textsc{Hard}(\mathcal{G}) = \bigcap_{\tau : S \times \vec{A} \to \mathbb{R}^n_+} \textsc{ne}(\mathcal{G}^\tau)$ of *hard equilibria* of the game is the set of initial equilibria which are contained in $\textsc{ne}(\mathcal{G}^\tau)$, for all static taxation schemes τ. The set $\textsc{Soft}(\mathcal{G})$ of *soft equilibria* is defined as the set of all initial equilibria which are not hard equilibria, i.e., $\textsc{Init}(\mathcal{G}) \setminus \textsc{Hard}(\mathcal{G})$.

The reason for singling out this set and giving it its name is illustrated by the following observation, whose proof in [224] directly carries over to our setting:

Proposition 4.6 (cf. [224])**.**

1. *For all games $\mathcal{G}$ and static taxation schemes τ, we have* $\textsc{Hard}(\mathcal{G}) \subseteq \textsc{ne}(\mathcal{G}^\tau) \subseteq \textsc{Init}(\mathcal{G})$.
2. *Let $\mathcal{G}$ be a game and let $\vec{\sigma}$ be a strategy profile for $\mathcal{G}$. If $\vec{\sigma} \in \textsc{Soft}(\mathcal{G})$, then $\exists \tau : S \times \vec{A} \to \mathbb{R}^n_+$ such that $\vec{\sigma} \in \textsc{ne}(\mathcal{G}^\tau)$.*

Proof. For part 1, the first inclusion is immediate by the definition of $\textsc{Hard}(\mathcal{G})$. To see the second inclusion, observe that for any static taxation scheme τ and any Nash equilibrium strategy profile $\vec{\sigma} \in \textsc{ne}(\mathcal{G}^\tau)$, any agent whose goal is not satisfied does not have a beneficial deviation to satisfy their goal and thus, $\vec{\sigma} \in \textsc{Init}(\mathcal{G})$.

For part 2, define τ so that in the game $\mathcal{G}^\tau$, all state-action profiles have the same cost. Clearly $\textsc{ne}(\mathcal{G}^\tau) = \textsc{ne}(\mathcal{G}^{\mathbf{0}}) = \textsc{Init}(\mathcal{G})$, and since $\vec{\sigma} \in \textsc{Soft}(\mathcal{G}) \subseteq \textsc{Init}(\mathcal{G})$, then $\vec{\sigma} \in \textsc{ne}(\mathcal{G}^\tau)$. □

We now give necessary and sufficient conditions for strategies to be in the set of hard and soft Nash equilibria of a game, as done for one-shot games in [225]. An interesting property of such conditions is that they do not depend on the cost structure of the given game, but are rather characterised purely in terms of qualitative properties related to the satisfaction of agents' goals and the repeating sequences of actions that they induce. Note that our setting leads to a different characterisation to that presented in [225].

Proposition 4.7. *Let $\mathcal{G}$ be a game and $\vec{\sigma}$ be a strategy profile. Then, $\vec{\sigma} \in \text{HARD}(\mathcal{G})$ if and only if all of the following conditions are satisfied:*

1. *$\vec{\sigma}$ is minimal for all agents $i \in N$.*
2. *For all agents $i \in L(\vec{\sigma})$ and all alternative strategies $\sigma_i' \in \Sigma_i \setminus \{\sigma_i\}$, it holds that $i \in L(\vec{\sigma}_{-i}, \sigma_i')$.*

Proof. For the forward direction, first suppose that $\vec{\sigma}$ is not minimal for some agent $i \in N$. This means that there is some other strategy $\sigma_i' \in \Sigma_i \setminus \{\sigma_i\}$ such that $i \in W(\vec{\sigma}) \iff i \in W(\vec{\sigma}_{-i}, \sigma_i')$ and for some $(s, \vec{a}) \in \text{Inf}(\rho(\vec{\sigma}))$, it is the case that $\Delta((s, \vec{a}), \vec{\sigma}, (\vec{\sigma}_{-i}, \sigma_i')) \neq 0$. By an application of Lemma 4.4, it follows that there exists a taxation scheme τ under which $\rho(\vec{\sigma}_{-i}, \sigma_i') \succ_i^\tau \rho(\vec{\sigma})$, which implies that $\vec{\sigma}$ is not a hard equilibrium.

Now suppose that for some agent $i \in L(\vec{\sigma})$, there is an alternative strategy $\sigma_i' \in \Sigma_i \setminus \{\sigma_i\}$ such that $i \in W(\vec{\sigma}_{-i}, \sigma_i')$. Then, i clearly has a beneficial deviation from σ_i under which their utility is negative to σ_i' under which their utility is positive. Again, this implies that $\vec{\sigma}$ is not a hard equilibrium.

For the backward direction, assume that both conditions 1 and 2 are satisfied. By condition 1, it is not hard to check that no agent can deviate to reduce their average costs under any taxation scheme, whilst retaining the logical satisfaction value of their goal. By condition 2, any agent $i \in L(\vec{\sigma})$ has no deviation in which they do achieve their goal, so $\vec{\sigma} \in \text{INIT}(\mathcal{G})$. Moreover, there is no taxation scheme τ, agent $i \in N$, and strategy $\sigma_i' \in \Sigma_i$ such that $\rho(\vec{\sigma}_{-i}, \sigma_i') \succ_i^\tau \rho(\vec{\sigma})$. □

Note that our setting leads to a different characterisation to that presented in [225], in that hard equilibria need not satisfy all of the agents' goals. This is because even if some agent's goal is not satisfied in a Nash equilibrium, there may be no way of incentivising such an agent to deviate if the Nash equilibrium in question is minimal.

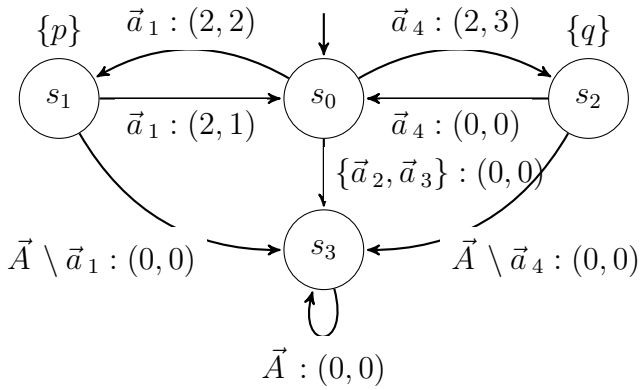

Figure 4.3: A two-agent concurrent game $\mathcal{G}_1$ with actions $A_1 = \{a, b\}, A_2 = \{c, d\}$ and goals $\gamma_1 = \mathbf{GF}p \wedge \mathbf{G}((p \rightarrow \mathbf{XX}q) \wedge (q \rightarrow \mathbf{XX}p))$ and $\gamma_2 = \gamma_1$, where we let $\vec{a}^1 = (a, c), \vec{a}^2 = (a, d), \vec{a}^3 = (b, c), \vec{a}^4 = (b, d)$. Cost vectors associated with sets denote that all action profiles within the set are assigned that vector.

It is worth noting that these conditions are not easily met in general, which is positive from the viewpoint of the principal as it implies that the only scenarios in which a particular initial equilibrium can not be eliminated are those in which every player's goal is satisfied.

Given a characterisation of the set of hard equilibria of a game $\mathcal{G}$, it is easy to check whether a given strategy profile $\vec{\sigma}$ is a soft equilibrium of $\mathcal{G}$, because the set $\text{SOFT}(\mathcal{G})$ of a game is defined as the set of its initial equilibria which are not hard equilibria. Thus, to decide whether a given $\vec{\sigma}$ is a soft equilibrium, one can first check whether it is an initial equilibrium of $\mathcal{G}$, which is the NE-CHECKING problem in [149]. If the answer is "no", then we can conclude that $\vec{\sigma}$ is not a soft equilibrium of $\mathcal{G}$. If the answer is "yes" on the other hand, then we can simply check whether $\vec{\sigma}$ is a hard equilibrium using the characterisation in Proposition 4.7, and then return the opposite answer. As an example, Figure 4.3 depicts a game with two hard equilibria, consisting of strategy profiles which output $(\vec{a}^1\vec{a}^1\vec{a}^4\vec{a}^4)^\omega$ and $(\vec{a}^4\vec{a}^4\vec{a}^1\vec{a}^1)^\omega$. Though neither equilibrium can be encouraged by the principal as the preferred strategy profile for the agents to play using static taxation schemes, we will soon explore a dynamic model of taxation schemes which does allow them to do so.

Alternatively, we can also directly characterise the set of soft equilibria of a game by providing necessary and sufficient conditions for a strategy profile to be a soft equilibrium:

Proposition 4.8. *Let $\mathcal{G}$ be a game and let $\vec{\sigma}$ be a strategy profile. Then $\vec{\sigma} \in \text{SOFT}(\mathcal{G})$ if and only if both of the following conditions are satisfied:*

1. $\vec{\sigma} \in \textsc{Init}(\mathcal{G})$.

2. $\vec{\sigma}$ *is not minimal for some agent* $i \in N$.

Proof. For the forward implication, observe that if $\vec{\sigma} \in \textsc{Soft}(\mathcal{G})$, then by definition, $\vec{\sigma} \in \textsc{Init}(\mathcal{G})$. Furthermore, since $\vec{\sigma}$ is a soft equilibrium, there must exist a taxation scheme τ such that $\vec{\sigma} \notin \textsc{ne}(\mathcal{G}^\tau)$. This means that under τ, there is some agent i and alternative strategy $\sigma_i' \neq \sigma_i$ such that $\rho(\vec{\sigma}_{-i}, \sigma_i') \succ_i^\tau \rho(\vec{\sigma})$. Clearly, this implies that $\vec{\sigma}$ is not minimal for i, otherwise such a taxation scheme could not exist. Now suppose that $i \in W(\vec{\sigma})$. Then, for σ_i' to be a beneficial deviation for agent i, it must be the case that $i \in W(\vec{\sigma}_{-i}, \sigma_i')$. Now suppose that $i \in L(\vec{\sigma})$. Then, if $i \in W(\vec{\sigma}_{-i}, \sigma_i')$, we could conclude that $\vec{\sigma} \notin \textsc{Init}(\mathcal{G})$ which is a contradiction. Thus, if $i \in L(\vec{\sigma})$, then it must also be the case that $i \in L(\vec{\sigma}_{-i}, \sigma_i')$ and so condition 2 holds.

For the backward implication, we observe that if conditions 1 and 2 hold, then Corollary 4.5 tells us that there exists a taxation scheme τ such that $\rho(\vec{\sigma}) \succ_i^\tau \rho(\vec{\sigma}_{-i}, \sigma_i')$ for some alternative strategy $\sigma_i' \in \Sigma_i \setminus \{\sigma_i\}$, which means that $\vec{\sigma}$ is a soft equilibrium. □

4.3.1 Decision Problems

As we have seen in the previous section, the notions of hard and soft equilibria are useful when a principal desires to *eliminate* individual strategies. In addition to this, we also consider the scenario in which a principal, who is external to the game, has a particular goal that they wish to see satisfied within the game; in a general economic setting, the goal might be intended to capture some principle of social welfare, for example. In our setting, the goal is specified as an LTL formula φ, and will typically represent a desirable system behaviour. The principal has the power to influence the game by choosing a taxation scheme and imposing it upon the game. Then, given a game $\mathcal{G}$ and a goal φ, our basic question is whether it is possible to design a taxation scheme τ such that, assuming the agents, individually and independently, act rationally (by choosing strategies $\vec{\sigma}$ that collectively form a Nash equilibrium in the modified game), the goal φ will be satisfied in the run $\rho(\vec{\sigma})$ that results from executing the strategies $\vec{\sigma}$. In the remainder of this section, we explore one way of interpreting this problem.

First, we need an auxiliary definition. We will say a taxation scheme τ *stabilises* a game $\mathcal{G}$ if the game $\mathcal{G}^\tau$ has at least one Nash equilibrium, i.e., if $\text{NE}(\mathcal{G}^\tau) \neq \emptyset$.

E-Nash Implementation: We say that a static taxation scheme $\tau : S \times \vec{A} \to \mathbb{R}^n_+$ *E-Nash implements* φ *in* $\mathcal{G}$ if there is a Nash equilibrium strategy profile $\vec{\sigma}$ of the game $\mathcal{G}^\tau$ such that $\rho(\vec{\sigma}) \models \varphi$. The notion of E-Nash implementation is thus analogous to the E-Nash concept in rational verification [129, 149]. Observe that, if the answer to this question is "yes", then this implies that the game $\mathcal{G}^\tau$ has at least one Nash equilibrium. Let us define the set $\text{ENI}(\mathcal{G}, \varphi)$ to be the set of taxation schemes τ that E-Nash implement φ in the game $\mathcal{G}$:

$$\text{ENI}(\mathcal{G}, \varphi) \coloneqq \{\tau : S \times \vec{A} \to \mathbb{R}^n_+ \mid \text{NE}_\varphi(\mathcal{G}^\tau) \neq \emptyset\}.$$

The obvious decision problem is then as follows:

> E-NASH IMPLEMENTATION:
> *Given*: Game $\mathcal{G}$, LTL goal φ.
> *Question*: Is it the case that $\text{ENI}(\mathcal{G}, \varphi) \neq \emptyset$?

A-Nash Implementation: Naturally enough, the counterpart of E-Nash implementation is *A-Nash Implementation*. To capture this idea, we might try to adapt the function $\text{ENI}(\cdots)$, and define $\text{ANI}(\mathcal{G}, \varphi)$ to denote the set of taxation schemes that A-Nash implement φ in $\mathcal{G}$ as follows:

$$\text{ANI}(\mathcal{G}, \varphi) \coloneqq \{\tau : S \times \vec{A} \to \mathbb{R}^n_+ \mid \text{NE}_\varphi(\mathcal{G}^\tau) = \text{NE}(\mathcal{G}^\tau)\}.$$

However, there is a problem with this definition: if there is no taxation scheme that stabilises $\mathcal{G}$, then for all taxation schemes τ we will have $\text{NE}_\varphi(\mathcal{G}^\tau) = \text{NE}(\mathcal{G}^\tau) = \emptyset$, in which case $\text{ANI}(\mathcal{G}, \varphi)$ will contain all possible taxation schemes, even though none of them actually implement φ. We therefore need a slightly more refined definition, as follows: we say that τ *A-Nash implements* φ *in* $\mathcal{G}$ if both

1. τ E-Nash implements φ in $\mathcal{G}$; and
2. $\text{NE}(\mathcal{G}^\tau) = \text{NE}_\varphi(\mathcal{G}^\tau)$.

We thus define $\text{ANI}(\mathcal{G}, \varphi)$ as follows:

$$\text{ANI}(\mathcal{G}, \varphi) \coloneqq \{\tau : S \times \vec{A} \to \mathbb{R}^n_+ \mid \text{NE}(\mathcal{G}^\tau) = \text{NE}_\varphi(G^\tau) \neq \emptyset\}.$$

The decision problem is then:

A-NASH IMPLEMENTATION:
Given: Game $\mathcal{G}$, LTL goal φ.
Question: Is it the case that $\text{ANI}(\mathcal{G}, \varphi) \neq \emptyset$?

It is arguably the case that A-Nash implementation is a more important problem to consider than E-Nash implementation because the principal could be assured that as long as the agents play according to a Nash equilibrium, the goal will be satisfied. In contrast, a positive answer to E-Nash implementation makes no such guarantee.

In our discussion going forward, we will sometimes refer to the set $\text{NE}_{\varphi}(\mathcal{G})$ as the set of *good equilibria* and $\text{NE}_{\neg\varphi}(\mathcal{G})$ as the set of *bad equilibria* when the game $\mathcal{G}$ and goal φ are clear from the context.

We now turn to the complexity of the decision problems we introduced above, beginning with the E-Nash implementation problem. This decision problem proves to be closely related to the aforementioned E-NASH problem, and the following result establishes its complexity:

Theorem 4.9. E-NASH IMPLEMENTATION *is* 2EXPTIME*-complete.*

Proof. For membership, suppose we are given an instance $(\mathcal{G}, \varphi)$ of the E-Nash implementation problem. Then check whether φ is satisfied on any Nash equilibrium of the cost-free concurrent game obtained from $\mathcal{G}$ by effectively removing its cost function using a taxation scheme which makes the cost of all state-action profiles the same for all agents. This then becomes the E-NASH problem, which is known to be 2EXPTIME-complete. The answer will be "yes" iff φ is satisfied on some Nash equilibrium of $\mathcal{G}^{\mathbf{0}}$; and if the answer is "yes", then from Proposition 4.6, the given LTL goal φ can be E-Nash implemented in $\mathcal{G}$.

For hardness, we can reduce the problem of checking whether a cost-free concurrent game has a Nash equilibrium (Theorem 4.2). Simply ask whether $\varphi = \top$ can be E-Nash implemented in $\mathcal{G}^{\mathbf{0}}$. The answer to this problem will clearly be "yes" iff the cost-free concurrent game $\mathcal{G}$ has a Nash equilibrium. □

The A-NASH IMPLEMENTATION problem asks whether there is a single taxation scheme under which the incentivised game admits at least one Nash equilibrium and that all such Nash equilibria satisfy the goal φ. Note that this is not equivalent to the problem of deciding whether all 'bad' Nash equilibria, i.e., ones under which φ is not satisfied, can be eliminated. This is because in the process of constructing a single taxation scheme which eliminates all bad

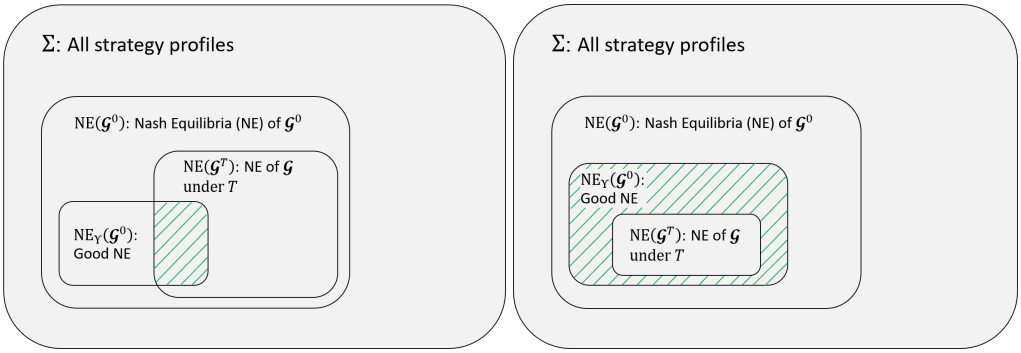

Figure 4.4: Venn diagrams depicting the sets of strategy profiles related to the Nash implementation problems. Left: E-Nash implementation problem. Right: A-Nash implementation problem.

equilibria, one may also necessarily eliminate all good equilibria along with them. Thus, a stronger condition is required to ensure A-NASH implementability, which is that the set of all bad equilibria can be eliminated by a single static taxation scheme whilst leaving at least one 'good' equilibrium remaining. To express this more formally, given a game $\mathcal{G}$, we will say that a taxation scheme τ *induces* a set of k strategies $\{\vec{\sigma}_j \mid j = 1 \ldots k\}$ if $\{\vec{\sigma}_j \mid j = 1 \ldots k\} \subseteq \text{NE}(\mathcal{G}^\tau)$. In contrast, we say that a taxation scheme τ *eliminates* a set of strategies $\{\vec{\sigma}_j \mid j = 1 \ldots k\}$ if $\{\vec{\sigma}_j \mid j = 1 \ldots k\} \cap \text{NE}(\mathcal{G}^\tau) = \emptyset$.

4.4 Dynamic Taxation Schemes

The model of taxation that we have been working with so far has been static, in the sense that it applies taxes consistently over time: performing the state-action profile $(s, \vec{a})$ now incurs the same cost for all agents as performing $(s, \vec{a})$ in one hundred time steps. This model has the advantage of simplicity, but it is naturally limited in the range of behaviours it can incentivise—particularly with respect to behaviours φ expressed as LTL formulae. In this section, we therefore introduce a *dynamic* model of taxation schemes. This model essentially allows a designer to impose taxation schemes that can choose to tax the same action different amounts, depending on the history of the run to date. A very natural model for dynamic taxation schemes is to describe them using a finite state machine with output—the same approach that we used to model strategies for individual agents. Formally, a dynamic taxation scheme T is defined by a tuple

$$T = (Q_T, \mathsf{next}_T, \mathsf{do}_T, q_T^0),$$

where:

- Q_T is the set of incentive machine states;
- $\mathsf{next}_T : Q_T \times A_1 \times \cdots \times A_n \to Q_T$ is the transition function of the machine;
- $q_T^0 \in Q_T$ is the initial state of the machine; and
- $\mathsf{do}_T : Q_T \to (S \times \vec{A} \to \mathbb{R}_+^n)$ is the output function of the machine.

With this, let $\mathscr{T}$ be the set of all dynamic taxation schemes for a game $\mathcal{G}$. We think of the incentive machine being executed alongside the agent strategies as a run unfolds, and at each time step, the machine outputs a static taxation scheme, which is applied at that time step only, with $\mathsf{do}_T(q_T^0)$ being the initial taxation scheme imposed.

When we impose taxation schemes, we no longer have a simple transformation $\mathcal{G}^\tau$ on games as we did with static taxation schemes τ. Instead we define the effect of a taxation scheme with respect to a run ρ. Formally, given a run ρ of a game $\mathcal{G}$, a dynamic taxation scheme T induces an infinite sequence of static taxation schemes, which we denote by $t(\rho, T)$. We can think of this sequence as a function

$$t(\rho, T) : \mathbb{N} \to (S \times \vec{A} \to \mathbb{R}_+^n).$$

We denote the cost of the run ρ in the presence of dynamic taxation scheme T by $\mathcal{C}^T(\rho)$:

$$\mathcal{C}^T(\rho) = \liminf_{u \to \infty} \frac{1}{u} \sum_{v=0}^{u} \mathcal{C}(\rho, v) + \underbrace{t(\rho, T)(v)(s(\rho, v), \vec{a}\,(\rho, v))}_{(*)}.$$

The expression $(*)$ denotes the vector of taxes incurred by the agents as a consequence of performing the action profile which they chose at time step v on the run ρ. The cost $\mathcal{C}_i^T(\rho)$ to agent i of the run ρ under T is then given by the i-th component of $\mathcal{C}^T(\rho)$.

4.4.1 Dynamic Hard and Soft Equilibria

Dynamic taxation schemes give a greater degree of control over how actions are incentivised than is the case with static schemes. Because of this, the conditions under which a given strategy profile is a hard or soft equilibrium will be different with dynamic taxation schemes than with static taxation schemes. To begin

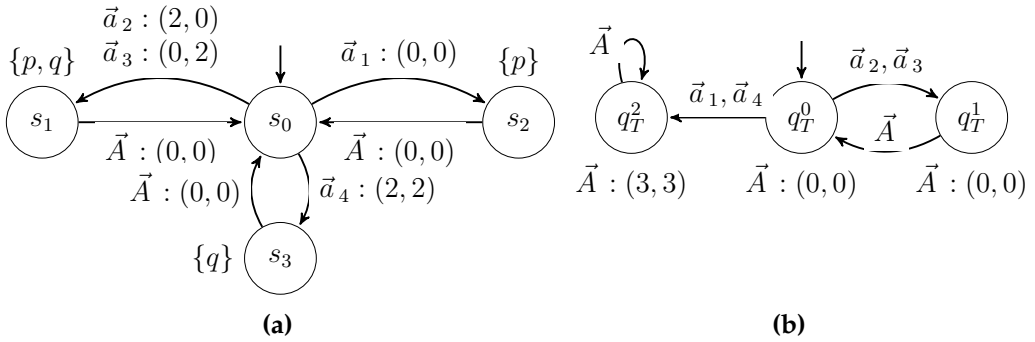

Figure 4.5: (a): A two-agent concurrent game $\mathcal{G}$ with action sets $A_1 = \{a, b\}$ and $A_2 = \{c, d\}$ and goals $\gamma_1 = \gamma_2 = \mathbf{GF}p$, where we let $\vec{a}^{\,1} = (a, c), \vec{a}^{\,2} = (a, d), \vec{a}^{\,3} = (b, c), \vec{a}^{\,4} = (b, d)$. Cost vectors associated with sets denote that all action profiles within the set are assigned those costs. (b): A dynamic taxation scheme that could be imposed on the agents in the game from (a). Labels below the states represent a static taxation scheme that applies a uniform tax for all agents and all action profiles.

with, we partition the set $\text{INIT}(\mathcal{G})$ of a game in an analogous manner to Section 4.3. Firstly, let $\text{D-HARD}(\mathcal{G}) = \cap_{T \in \mathscr{T}} \text{NE}(\mathcal{G}^T)$ be the set of *dynamic hard equilibria* and $\text{D-SOFT}(\mathcal{G}) = \text{INIT}(\mathcal{G}) \setminus \text{D-HARD}(\mathcal{G})$ be the set of *dynamic soft equilibria* of a given game $\mathcal{G}$.

We then have the following characterisation, below, of the set of dynamic hard equilibria, which is analogous to the conditions found in [225]:

Proposition 4.10. *Let $\vec{\sigma}$ be a strategy profile of a game $\mathcal{G}$. Then $\vec{\sigma} \in \text{D-HARD}(\mathcal{G})$ if and only if both of the following conditions are satisfied:*

1. *$W(\vec{\sigma}) = N$;*
2. *For all agents $i \in N$ and all strategies $\sigma_i' \in \Sigma_i$ such that $\rho(\vec{\sigma}) \neq \rho((\vec{\sigma}_{-i}, \sigma_i'))$, it holds that $i \in L(\vec{\sigma}_{-i}, \sigma_i')$.*

Proof. For the forward implication, suppose that condition 1 does not hold. Then, there exists some agent $i \in N$ such that $i \in L(\vec{\sigma})$. Such an agent will want to choose a strategy which minimises her costs. Now, if there exists some strategy σ_i' such that $i \in W(\vec{\sigma}_{-i}, \sigma_i')$, then $\vec{\sigma} \notin \text{D-HARD}(\mathcal{G})$ because there is a beneficial deviation from $\vec{\sigma}$ for i. Therefore, for all other strategies $\sigma_i' \neq \sigma_i$, it must be the case that $i \in L(\vec{\sigma}_{-i}, \sigma_i')$. It is not difficult to see that we can design a taxation scheme to make σ_i more costly than one of these other strategies $\sigma_i' \neq \sigma_i$ by constructing a dynamic taxation scheme T_1 which taxes i for playing the strategy σ_i by a sufficient amount. More precisely, we let $T_1 = (Q_{T_1}, \mathsf{next}_{T_1}, \mathsf{do}_{T_1}, q^0_{T_1})$, where

- $Q_{T_1} = \{q^0_{T_1}, \ldots, q^{|Q_i|}_{T_1}, q^{|Q_i|+1}_{T_1}\}$,
- For all $\vec{a} \in \vec{A}$ where $a_i = \mathsf{do}_i(q^j_i)$, if $j \leq |Q_i|$, then $\mathsf{next}_{T_1}(q^j_{T_1}, \vec{a}) = q^k_{T_1}$, where $\mathsf{next}_i(q^j_i, \vec{a}) = q^k_i$ and q^k_i is the k-th machine state of σ_i. Otherwise, let $\mathsf{next}_{T_1}(q^j_{T_1}, \vec{a}) = q^{|Q_i|+1}_{T_1}$;
- $\mathsf{do}_{T_1}(q^k_{T_1}) = \tau^{T_1}(\cdot)$, where for all $(s, \vec{a}) \in S \times \vec{A}$ and $j \in N$, we have that $\tau^{T_1}_j(s, \vec{a}) = \begin{cases} c^*_i + 1 & k \leq |Q_i| \text{ and } j = i; \\ 0 & \text{otherwise.} \end{cases}$

Intuitively, this dynamic taxation scheme mimics the strategy σ_i whilst ignoring the actions of all other agents, but has an additional absorbing state $q^{|Q_i|+1}_{T_1}$ which is transitioned into whenever i selects an action which does not agree with what they would have selected under σ_i. As long as i is playing according to the strategy σ_i, T_1 repeatedly outputs a taxation scheme which heavily taxes her, and immediately switches to the 0-cost taxation scheme forever as soon as a deviation has occurred, at which point T_1 will have entered the aforementioned absorbing state. The effect of this taxation scheme is that any other strategy σ'_i such that $i \in W(\vec{\sigma}) \iff i \in W(\vec{\sigma}_{-i}, \sigma'_i)$ will be strictly less costly and hence more preferable to i than σ_i. Thus, the strategy σ'_i represents a beneficial deviation for agent i under T_1 so if condition 1 does not hold, then $\vec{\sigma} \notin \text{D-HARD}(\mathcal{G})$. Now assume that condition 2 does not hold. Then, for some agent $i \in N$, there exists an alternative strategy σ'_i such that $i \in W(\vec{\sigma}_{-i}, \sigma'_i)$. Again, we can design a dynamic taxation scheme T_2 to make σ_i more costly than σ'_i, which will in turn make σ'_i a beneficial deviation for agent i because her goal is satisfied under both $\vec{\sigma}$ and $(\vec{\sigma}_{-i}, \sigma'_i)$ but the latter is less costly under T_2. Thus if condition 2 does not hold, then $\vec{\sigma} \notin \text{D-HARD}(\mathcal{G})$.

For the backward implication, assume for contradiction that both conditions 1 and 2 hold but that $\vec{\sigma} \notin \text{D-HARD}(\mathcal{G})$. Then, there must be some dynamic taxation scheme T_3 such that $\vec{\sigma} \notin \text{NE}(\mathcal{G}^{T_3})$. Because $\vec{\sigma}$ is not a Nash equilibrium under this taxation scheme, some agent i must have a beneficial deviation σ'_i which allows her to achieve a higher utility. However, by condition 1, all agents get their goal achieved under $\vec{\sigma}$ and by condition 2, any deviation from $\vec{\sigma}$ by an agent will result in their goal not being achieved and hence receiving negative utility. It follows from this that σ'_i can not be a beneficial deviation and hence, we obtain a contradiction. □

Proposition 4.11. *Let $\vec{\sigma}$ be a strategy profile of a game $\mathcal{G}$. Then $\vec{\sigma} \in$ D-SOFT($\mathcal{G}$) if and only if both of the following conditions are satisfied:*

1. $\vec{\sigma} \in$ INIT($\mathcal{G}$).
2. *There exists an agent i and an alternative strategy $\sigma_i' \neq \sigma_i$ such that $i \in W(\vec{\sigma}) \iff i \in W(\vec{\sigma}_{-i}, \sigma_i')$.*

Proof. For the forward implication, assume first that $\vec{\sigma} \in$ D-SOFT($\mathcal{G}$). By definition, condition 1 trivially follows. Furthermore, there exists a dynamic taxation scheme T_4 such that $\vec{\sigma} \notin$ NE($\mathcal{G}^{T_4}$). This means that under T_4, there is an agent i who has a beneficial deviation σ_i' from $\vec{\sigma}$. Moreover, we claim that this strategy satisfies condition 2, namely that $i \in W(\vec{\sigma}) \iff i \in W(\vec{\sigma}_{-i}, \sigma_i')$. Suppose for contradiction that this is not the case. Then we would have that either $i \in W(\vec{\sigma}) \cap L(\vec{\sigma}_{-i}, \sigma_i')$ or $i \in L(\vec{\sigma}) \cap W(\vec{\sigma}_{-i}, \sigma_i')$. In the first case, σ_i' could not possibly be a beneficial deviation because $u_i^{T_4}(\rho(\vec{\sigma}_{-i}, \sigma_i')) < 0 < u_i^{T_4}(\rho(\vec{\sigma}))$. In the second case, $\rho(\vec{\sigma}_{-i}, \sigma_i')$ would represent a beneficial deviation for i in the game $\mathcal{G}^{\mathbf{0}}$, meaning that $\vec{\sigma} \notin$ INIT($\mathcal{G}$). In both cases, we have a contradiction and hence, condition 2 must be satisfied if $\vec{\sigma} \in$ D-SOFT($\mathcal{G}$).

For the backward implication, suppose that conditions 1 and 2 hold. We can design a dynamic taxation scheme T_5 such that $\vec{\sigma} \notin$ NE($\mathcal{G}^{T_5}$) by using the same construction as the dynamic taxation scheme T_1 in the proof of Proposition 4.10. Under T_5, the alternative strategy σ_i' from condition 2 is a beneficial deviation for i and hence $\vec{\sigma}$ satisfies the definition of a soft equilibrium. □

An important defining feature of dynamic hard and soft equilibria is the fact that the actual cost structure of the concurrent game does not affect whether a strategy is a hard or soft equilibrium. The only factors of importance are whether agents achieve their goals under the strategy and under alternative strategies.

Dynamic Decision Problems

One may wonder whether there are any benefits to be gained from using a more expressive incentive model, but in order to answer this question, we need to specify *in which respect* such benefits arise. As the central focus of this study is on the question of designing taxation schemes to elicit behaviours that satisfy a principal's goal in some or all Nash equilibria of the game, we will study this problem in the context of dynamic taxation schemes and show that this model

is strictly more powerful than static taxation schemes. For this, we define the decision problems in the same way as in section 4.3.1.

Dynamic E-Nash Implementation: Given a game $\mathcal{G}$, we will say that a dynamic taxation scheme $T \in \mathscr{T}$ *D-E-Nash implements* φ *in* $\mathcal{G}$ if the set $\text{NE}_\varphi(\mathcal{G}^T)$ is nonempty. We then define the set $\text{D-ENI}(\mathcal{G}, \varphi)$ to be the set of dynamic taxation schemes T that D-E-Nash implement φ in the game $\mathcal{G}$:

$$\text{D-ENI}(\mathcal{G}, \varphi) = \{T \in \mathscr{T} \mid \text{NE}_\varphi(\mathcal{G}^T) \neq \emptyset\}.$$

The decision problem is then as follows:

> D-E-NASH IMPLEMENTATION:
> *Given*: Game $\mathcal{G}$, LTL goal φ.
> *Question*: Is it the case that $\text{D-ENI}(\mathcal{G}, \varphi) \neq \emptyset$?

Dynamic A-Nash Implementation: In the same manner as before, we define $\text{D-ANI}(\mathcal{G}, \varphi)$ as follows:

$$\text{D-ANI}(\mathcal{G}, \varphi) = \{T \in \mathscr{T} \mid \text{NE}(\mathcal{G}^T) = \text{NE}_\varphi(\mathcal{G}^T) \neq \emptyset\}.$$

The decision problem is then:

> D-A-NASH IMPLEMENTATION:
> *Given*: Game $\mathcal{G}$, LTL goal φ.
> *Question*: Is it the case that $\text{D-ANI}(\mathcal{G}, \varphi) \neq \emptyset$?

This problem asks whether there exists a dynamic taxation scheme T such that: (1) the set of Nash equilibria of $\mathcal{G}^T$ is non-empty and (2) all Nash equilibria of $\mathcal{G}^T$ satisfy φ. In essence, this means that the incentive designer would like to guarantee that no matter which equilibrium is chosen by the agents, their objective φ will be satisfied, and that one such equilibrium exists.

Given this, we observe the following relationship between the static and dynamic versions of the existential and universal Nash implementation problem:

Proposition 4.12. *For all games* $\mathcal{G}$ *and goals* φ *the following statements hold:*

1. $\text{ANI}(\mathcal{G}, \varphi) \neq \emptyset \rightarrow \text{D-ANI}(\mathcal{G}, \varphi) \neq \emptyset$;
2. $\text{ENI}(\mathcal{G}, \varphi) \neq \emptyset \iff \text{D-ENI}(\mathcal{G}, \varphi) \neq \emptyset$.

Proof. We begin with statement 1. If the answer to A-NASH implementation is "yes", then let τ be any static taxation scheme in $\text{ANI}(\mathcal{G}, \varphi)$. Construct a dynamic taxation scheme $T = (q_0, q_0 \times \vec{A} \rightarrow q_0, q_0 \rightarrow \tau, q_0)$ which has a single state q_0 and outputs τ in this state. Clearly, this dynamic taxation scheme D-A-NASH implements φ so statement 1 holds.

The same line of reasoning can be applied for the forward implication of statement 2. We prove the backward implication by contradiction. Suppose that $\text{D-ENI}(\mathcal{G}, \varphi) \neq \emptyset$ and $\text{ENI}(\mathcal{G}, \varphi) = \emptyset$. Then from Theorem 4.9, it must be the case that φ is not satisfied on any strategy $\vec{\sigma} \in \text{INIT}(\mathcal{G}) = \text{NE}(\mathcal{G}^{\mathbf{0}})$. On the other hand, there exists a dynamic taxation scheme T and a strategy profile $\vec{\sigma} \in \text{NE}_{\varphi}(\mathcal{G}^T)$. However, by the definition of Nash equilibria, any agent whose goal is not satisfied under $\vec{\sigma}$ has no deviation in which their goal is satisfied and hence, $\vec{\sigma} \in \text{INIT}(\mathcal{G})$ which is a contradiction. □

This result reveals that the degree to which dynamic taxation schemes are more useful than static taxation schemes is restricted to the A-Nash implementation problem. The following is an immediate corollary of Theorem 4.9 and statement 2 of Proposition 4.12:

Theorem 4.13. D-E-NASH IMPLEMENTATION *is* 2EXPTIME*-complete.*

In light of Proposition 4.12, it is rather natural to ask whether the reverse of statement 1 holds, that is, whether it is the case that $\text{D-ANI}(\mathcal{G}, \varphi) \neq \emptyset \rightarrow \text{ANI}(\mathcal{G}, \varphi) \neq \emptyset$. As the following proposition shows, this is not the case in general. What this highlights is that although dynamic taxation schemes provide no added benefit when it comes to the static and dynamic variants of the E-Nash implementation problem, they do in fact afford the principal a greater capability to eliminate undesirable strategy profiles from a game.

Proposition 4.14. *There exists a game $\mathcal{G}$ and an LTL goal φ such that* $\text{D-ANI}(\mathcal{G}, \varphi) \neq \emptyset$*, but* $\text{ANI}(\mathcal{G}, \varphi) = \emptyset$*.*

Proof. Consider the concurrent game $\mathcal{G}$ in Figure 4.5a. Intuitively, both agents desire to always eventually visit either s^1 or s^2. Suppose that the principal's objective is $\varphi = \mathbf{G}(p \leftrightarrow q)$, i.e., they would like the agents to never visit s^2 or s^3. Firstly, observe that there is no *static* taxation scheme which can A-Nash implement φ, as any modification to the costs of the game will not eliminate any Nash equilibria where the agents visit s^2 or s^3 a finite number of times. This is due to the prefix-independence of costs in infinite games with limiting-average

payoffs [133]. However, the dynamic taxation scheme depicted in Figure 4.5b A-Nash implements φ. To see this, observe that for any strategy profile that visits s^2 or s^3 a finite number of times, there exists a deviation for some agent to ensure that s^2 and s^3 are never visited. Such a deviation will result in all agents $i \in \{1, 2\}$ satisfying their goals γ_i and strictly reducing their average costs from at least $c_i^* + 1$ to some value strictly below this. This constitutes a beneficial deviation and hence, there is no Nash equilibrium under T that does not satisfy φ. Moreover, any strategy profile $\vec{\sigma}$ that leads to the sequence of states $s(\rho(\vec{\sigma}), 0:) = (s^0 s^1)^\omega$ is a Nash equilibrium of $\mathcal{G}^T$ and hence goal φ is A-Nash implemented by T in this game. □

The key mechanism at work in the above proof is the *prefix-independence* of limiting-average costs. For any run ρ, modifying a finite prefix of ρ does not alter $\mathcal{C}_i(\rho)$, since the contribution of any finite segment vanishes in the limit. A static taxation scheme assigns the same cost function at every step, so the limiting-average cost it induces depends only on the long-run frequency of state–action pairs, not on the order in which they occur. Consequently, a static scheme cannot distinguish a run that visits an undesirable state finitely many times from one that never visits it at all, as both yield the same limiting-average cost. Dynamic taxation schemes circumvent this barrier by conditioning on the observed history. Once an undesirable state is visited, the scheme can transition to a new internal state and permanently escalate the taxes it levies.

Before proceeding with the D-A-NASH IMPLEMENTATION problem, we will need to reintroduce some additional terminology and concepts, beginning first with deviation paths and cycles as we saw in the previous chapter. We say that a finite sequence $\vec{\sigma}^{\,1}, \vec{\sigma}^{\,2}, \ldots, \vec{\sigma}^{\,m}$ of strategy profiles is a *deviation path* if for some sequence of indices $i_1, i_2, \ldots, i_{m-1}$ where each $i_j \in N$, it holds that $\vec{\sigma}^{\,m} \to_{i_{m-1}} \vec{\sigma}^{\,m-1} \to_{i_{m-2}} \ldots \to_{i_1} \vec{\sigma}^{\,1}$, and for all $j \in 1 \ldots m-1$, we have $\vec{\sigma}^{\,j}_{-i_j} = \vec{\sigma}^{\,j+1}_{-i_j}$. In other words, a deviation path is a sequence of initial deviations from a strategy profile $\vec{\sigma}^{\,1}$ by individual agents, where each deviation results in a strategy profile that is distinguishable from its predecessor. A *deviation cycle* is a deviation path $\vec{\sigma}^{\,1}, \ldots, \vec{\sigma}^{\,m}$ where $\vec{\sigma}^{\,1} = \vec{\sigma}^{\,m}$. Given a game $\mathcal{G}$, a *Nash deviation path* is a deviation path where all strategy profiles in the path are initial equilibria of $\mathcal{G}$ and a *Nash deviation cycle* is defined in the obvious manner. A particularly important property of Nash deviation cycles is that since all strategy profiles are initial equilibria, all initial deviations between consecutive strategy profiles on such cycles are in fact soft deviations.

Proposition 4.15. *Let $\mathcal{G}$ be a game and $X \subset \Sigma$ be a finite set of strategy profiles in $\mathcal{G}$. Then, X is eliminable if and only if there exists a finite deviation graph $\Gamma = (\mathcal{V}, E)$ that satisfies the following properties: (1) For every $\vec{\sigma} \in X$, there is some $\vec{\sigma}\,' \in \mathcal{V}$ such that $(\vec{\sigma}, \vec{\sigma}\,') \in E$; and (2) every deviation cycle in Γ involves at least two agents.*

Proof. For the forward direction, suppose that there is no deviation graph Γ satisfying both properties (1) and (2) in the statement. Then, for all deviation graphs Γ, either for some $\vec{\sigma} \in X$, there is no $\vec{\sigma}\,' \in \mathcal{V}$ such that $(\vec{\sigma}, \vec{\sigma}\,') \in E$, or there is some deviation cycle in Γ involving only one agent. Now consider any deviation graph $\Gamma = (\mathcal{V}, E)$, where $\mathcal{V} = X \cup \{\vec{\sigma}\,' \mid \vec{\sigma} \to_i \vec{\sigma}\,' \text{ for some } \vec{\sigma} \in X \text{ and } i \in N\}$. In the first case, it is clear that any taxation scheme that induces Γ does not eliminate $\{\vec{\sigma}\}$ and hence X. In the second case, no taxation scheme can induce the deviation graph Γ. To see why, suppose for contradiction that some taxation scheme T induces Γ and let i be the agent for which there is a deviation cycle $C = \{\vec{\sigma}^{\,1}, \ldots, \vec{\sigma}^{\,m}\}$ in Γ involving only agent i. Then, we have $\vec{\sigma}^{\,1} \succ_i^T \vec{\sigma}^{\,2} \succ_i^T \ldots \succ_i^T \vec{\sigma}^{\,m}$ and by transitivity of the preference relation $\succ_i^T$, we can conclude that $\vec{\sigma}^{\,1} \succ_i^T \vec{\sigma}^{\,m}$. However, by definition of a deviation cycle, $\vec{\sigma}^{\,1}$ and $\vec{\sigma}^{\,m}$ are indistinguishable, so agent i will always receive the same utility under both $\vec{\sigma}^{\,1}$ and $\vec{\sigma}^{\,m}$, no matter what taxation scheme is imposed on them and hence, we have a contradiction. From this, we can conclude that every deviation graph that can be induced by a taxation scheme does not eliminate X and hence, X is not eliminable, proving this part of the statement.

For the backward direction, assume that there is a deviation graph Γ that satisfies both properties. Under this assumption, we will construct a dynamic taxation scheme T that eliminates X. To assign the appropriate costs to different strategy profiles, we will make use of the lengths of deviation paths within Γ. For every $i \in N$, let ℓ_i denote the length of the longest deviation path in Γ that involves only agent i. Additionally, for all $\vec{\sigma} \in \mathcal{V}$, let $d_i(\rho(\vec{\sigma}))$ denote the length of the longest deviation path in Γ that starts from $\rho(\vec{\sigma})$ and involves only i. The difference between these two quantities will serve as the basis for how much taxation an agent i will incur for any given strategy profile in $\mathcal{V}$. Observe that because it is assumed that no deviation cycle involves only one agent, both quantities are well-defined and finite for all agents and strategy profiles. Then, for a deviation graph Γ and a run ρ, let $\textsc{in}(\rho, \Gamma)$ be the set of agents $i \in N$ for which there is some pair of strategy profiles $\vec{\sigma}, \vec{\sigma}\,' \in \mathcal{V}$ such that we have both $(\vec{\sigma}, \vec{\sigma}\,') \in E$ and $\rho = \rho(\vec{\sigma}\,')$. In other words, $\textsc{in}(\rho, \Gamma)$ represents the set of agents who have an initial deviation from some other strategy profile in $\mathcal{V}$ to one that

generates the run ρ. With this, we would like to construct a dynamic taxation scheme such that for any strategy profile $\vec{\sigma}$, the following criteria are satisfied:

- $\mathcal{C}_i^T(\rho(\vec{\sigma})) \geq (\ell_i - d_i(\rho(\vec{\sigma}))) \cdot (c_i^* + 1)$ if $i \in \text{IN}(\rho, \Gamma)$;
- $\mathcal{C}_i^T(\rho(\vec{\sigma})) = C_i(\rho(\vec{\sigma}))$ otherwise.

Intuitively, the idea is to ensure that for every edge $(\vec{\sigma}, \vec{\sigma}') \in E$, the agent $i \in N$ for whom $\vec{\sigma} \rightarrow_i \vec{\sigma}'$ gets taxed by a significantly higher amount for choosing $\vec{\sigma}$ compared to when they choose $\vec{\sigma}'$. To see why it is possible to construct such a taxation scheme, first observe that if $\rho \neq \rho'$ for any two runs ρ, ρ', then there is some dynamic taxation scheme that can distinguish between the two by simply tracing out the two runs up to the first point in which they differ and then branching accordingly. From this point onwards, the dynamic taxation scheme can then output static taxation schemes, which assign different limiting average costs to the agents according to the above criteria. Extending this approach to a taxation scheme that distinguishes between all unique runs generated by elements of $\mathcal{V}$, it follows that there is a dynamic taxation scheme T that satisfies the two criteria. Consequently, for all $(\vec{\sigma}, \vec{\sigma}') \in E$, it follows that $\vec{\sigma}' \succ_i^T \vec{\sigma}$ because $\rho(\vec{\sigma}) \neq \rho(\vec{\sigma}')$ by definition of the initial deviation relation $\rightarrow_i$. Moreover, because it is assumed that no deviation cycle involves only one agent, T gives rise to a strict total ordering $\succ_i^T$ on the elements of $\mathcal{V}$ for each $i \in N$. Finally, by property (1), it holds that for every $\vec{\sigma} \in X$, some agent has a beneficial deviation from $\vec{\sigma}$ to another $\vec{\sigma}' \in \mathcal{V}$ under T and hence, T eliminates X. □

We now show that if a set of strategy profiles is eliminable, then it can in fact be eliminated by a *local* taxation scheme – one that applies non-zero taxes only to runs generated by the targeted profiles. This will be needed for the characterisation theorem that follows.

Lemma 4.16. *Let $\mathcal{G}$ be a game and $X \subset \Sigma$ be a finite set of strategy profiles. If X is eliminable, then there exists a dynamic taxation scheme T that eliminates X and is X-local, i.e., T applies non-zero taxes only to runs generated by strategy profiles in X.*

Proof. Since X is eliminable, there exists a dynamic taxation scheme T_0 that eliminates X. For each $\vec{\sigma} \in X$, fix a beneficial deviation from $\vec{\sigma}$ under T_0: an agent $j(\vec{\sigma})$ and a strategy $\sigma'_{j(\vec{\sigma})}$ such that $(\vec{\sigma}_{-j(\vec{\sigma})}, \sigma'_{j(\vec{\sigma})}) \succ_{j(\vec{\sigma})}^{T_0} \vec{\sigma}$. Let $\vec{\sigma}'(\vec{\sigma})$ denote the resulting strategy profile after the deviation. Build a deviation graph $\Gamma = (\mathcal{V}, E)$ whose edges are precisely these chosen beneficial deviations $\{(\vec{\sigma}, \vec{\sigma}'(\vec{\sigma})) \mid \vec{\sigma} \in$

$X\}$. This graph satisfies property (1) by construction and property (2) because any single-agent cycle $\vec{\sigma}^{\,1} \to_j \vec{\sigma}^{\,2} \to_j \cdots \to_j \vec{\sigma}^{\,1}$ would yield $\vec{\sigma}^{\,1} \succ_j^{T_0} \vec{\sigma}^{\,1}$ by transitivity, contradicting the irreflexivity of strict preference. Moreover, since each edge corresponds to a beneficial deviation, the deviating agent's goal satisfaction is either preserved or improved along each edge.

We now construct an X-local dynamic taxation scheme T from Γ. For runs not generated by any profile in X, let T output zero taxes. For each $\vec{\sigma} \in X$ and each agent $i \in N$, we assign taxes using the acyclic structure of Γ. For each agent i, let $\delta_i(\vec{\sigma})$ denote the length of the longest chain of i-deviations in Γ ending at $\vec{\sigma}$, and let $\delta_{\max} = \max_{i \in N, \vec{\sigma} \in X} \delta_i(\vec{\sigma})$. We set the tax on agent i at run $\rho(\vec{\sigma})$ to $(\delta_{\max} + 1 - \delta_i(\vec{\sigma})) \cdot (c_i^* + 1)$. This is well defined because the scheme can distinguish between all runs of profiles in X (which are all distinct) by tracing them to the first point at which they diverge.

To see that T eliminates X, consider any edge $(\vec{\sigma}, \vec{\sigma}\,'(\vec{\sigma})) \in E$ with deviating agent $j = j(\vec{\sigma})$. If $\vec{\sigma}\,'(\vec{\sigma}) \notin X$, then agent j's cost at $\rho(\vec{\sigma}\,'(\vec{\sigma}))$ is $\mathcal{C}_j(\rho(\vec{\sigma}\,'(\vec{\sigma}))) \leq c_j^*$, whereas the tax ensures $\mathcal{C}_j^T(\rho(\vec{\sigma})) \geq c_j^* + 1$, so j's cost strictly decreases. If $\vec{\sigma}\,'(\vec{\sigma}) \in X$, then $\delta_j(\vec{\sigma}\,'(\vec{\sigma})) \geq \delta_j(\vec{\sigma}) + 1$, so the tax on agent j at $\vec{\sigma}\,'(\vec{\sigma})$ is lower by at least $(c_j^* + 1)$, which exceeds the maximum possible difference in original costs (c_j^*), so j's total cost again strictly decreases. In both cases, since j's goal satisfaction is preserved or improved and j's cost strictly decreases, the deviation is beneficial under T. By property (1), every $\vec{\sigma} \in X$ admits such a beneficial deviation, so T eliminates X. □

Given this, we are ready to prove our main result for the dynamic A-Nash implementation problem:

Theorem 4.17. *Let $\mathcal{G}$ be a game and φ be an LTL formula. Then* D-ANI$(\mathcal{G}, \varphi) \neq \emptyset$ *if and only if the following conditions hold:*

1. D-ENI$(\mathcal{G}, \varphi) \neq \emptyset$*;*
2. NE$_{\neg\varphi}(\mathcal{G}^{\mathbf{0}})$ *is eliminable.*

Proof. For the forward direction, it is obvious by definition of the problem that if D-ENI$(\mathcal{G}, \varphi) = \emptyset$, then D-ANI$(\mathcal{G}, \varphi) = \emptyset$. Moreover, it is also clear that if NE$_{\neg\varphi}(\mathcal{G}^{\mathbf{0}})$ is not eliminable, then it is impossible to design a taxation scheme such that only good equilibria remain and hence, D-ANI$(\mathcal{G}, \varphi) = \emptyset$.

For the reverse direction, suppose that the two conditions hold. By Lemma 4.16, since $\text{NE}_{\neg\varphi}(\mathcal{G}^{\mathbf{0}})$ is eliminable, there exists a dynamic taxation scheme T that eliminates $\text{NE}_{\neg\varphi}(\mathcal{G}^{\mathbf{0}})$ and is $\text{NE}_{\neg\varphi}(\mathcal{G}^{\mathbf{0}})$-local (i.e., T only applies non-zero taxes to runs generated by strategy profiles in $\text{NE}_{\neg\varphi}(\mathcal{G}^{\mathbf{0}})$). Now consider the result of augmenting T by adding a static taxation scheme τ such that $\tau(s, \vec{a}) = \hat{c}$ for all $(s, \vec{a}) \in S \times \vec{A}$ and some $\hat{c} \geq \max_{i \in N} c_i^*$. Combining this with T gives us a taxation scheme T^* such that for each state $q \in Q_{T^*} = Q_T$ and $(s, \vec{a}) \in S \times \vec{A}$, we have $\mathsf{do}_{T^*}(q)(s, \vec{a}) = \mathsf{do}_T(q)(s, \vec{a}) + \tau(s, \vec{a})$. Now, because T eliminates $\text{NE}_{\neg\varphi}(\mathcal{G}^{\mathbf{0}})$, and $\text{NE}(\mathcal{G}^{\tau}) = \text{NE}(\mathcal{G}^{\mathbf{0}})$, it follows that the combined taxation scheme T^* eliminates $\text{NE}_{\neg\varphi}(\mathcal{G}^{\mathbf{0}})$.

Because the satisfaction of an LTL formula on a given run is solely dependent on the run's trace, it follows that all good equilibria, i.e., strategy profiles in $\text{NE}_{\varphi}(\mathcal{G}^{\mathbf{0}})$, must be distinguishable from all bad equilibria. Combining this observation with the assumption that T is $\text{NE}_{\neg\varphi}(\mathcal{G}^{\mathbf{0}})$-local, it follows that no run induced by a strategy profile in $\text{NE}_{\varphi}(\mathcal{G}^{\mathbf{0}})$ incurs a limiting-average cost for any agent that is different from $\hat{c}$. Moreover, for any $\vec{\sigma} \in \text{NE}_{\varphi}(\mathcal{G}^{\mathbf{0}})$, no agent will want to deviate to some $\vec{\sigma}' \in \text{NE}_{\neg\varphi}(\mathcal{G}^{\mathbf{0}})$ because the cost that they incur can only increase by deviating to such a strategy profile and by virtue of membership in $\text{NE}_{\varphi}(\mathcal{G}^{\mathbf{0}})$, it is not possible for any agent $i \in L(\vec{\sigma})$ to unilaterally deviate and become a winning agent. From these two points, it follows that all strategy profiles in $\text{NE}_{\varphi}(\mathcal{G}^{\mathbf{0}})$ are Nash equilibria under T^* and no strategy profiles in $\text{NE}_{\neg\varphi}(\mathcal{G}^{\mathbf{0}})$ are Nash equilibria under T^*, so we have that $\text{NE}(\mathcal{G}^{T^*}) = \text{NE}_{\varphi}(\mathcal{G}^{\mathbf{0}})$. Finally, because $\text{D-ENI}(\mathcal{G}, \varphi) \neq \emptyset$, we have that $\text{NE}_{\varphi}(\mathcal{G}^{\mathbf{0}}) \neq \emptyset$ and hence, $\text{D-ANI}(\mathcal{G}, \varphi) \neq \emptyset$.

□

4.5 Discussion

The work in this chapter was motivated by [224], and based on that work, presents three main contributions: the introduction of *dynamic* taxation schemes for an extension of their model to concurrent games with a demonstration of their increased capability to incentivise behaviours, formal characterisations of the sets of *soft and hard equilibria* in this setting as well as the D-A-NASH implementation problem, and a study of the computational *complexity* of the existential Nash implementation decision problems.

For each of the three main contributions, we highlight some previous work in these areas. *Dynamic taxation schemes* are a new concept and technical construction

in the setting of concurrent games, that has its origins in the addition of "static" taxation schemes in [224] to (one-shot) Boolean games, and was subsequently tailored to reason about iterated games of infinite duration such as those that arise in the context of concurrent games, where mean-payoff utility functions are used to capture the long-term performance of a multi-agent system. The second main contribution is necessary and sufficient conditions are given to characterise the sets of *soft and hard equilibria*, which were originally introduced in [225] where, again, only (one-shot) Boolean games were considered. In addition, although the complexity of the universal Nash implementation problems remain open, we make a significant step towards solving them by providing a characterisation of the dynamic version of the problem. The third main contribution, related to new complexity results, builds on previous work in the area of rational verification [121, 129, 149], particularly for the case of concurrent games played on graphs with combined qualitative and quantitative goals.

An obvious starting point for future work is to explore scenarios and further restrictions which would make the problem of incentive design more tractable. For example, one could consider fragments of LTL, such as safety properties or GR1 properties, which, as highlighted in Chapter 2, make the problems of rational verification easier. Additionally, the goal of the principal has been assumed here to be given by an LTL formula. However, many reward scenarios involve not just the binary satisfaction of some property, but the optimisation of a particular cost or reward function. Extending the model to allow for such scenarios would be a natural extension of the work in this chapter. Finally, while dynamic taxation schemes are a natural and more powerful tool for incentive design, they can also be more complex to reason about and design. Moreover, the imposition of dynamic taxation schemes in practical scenarios can make decision-making more complex from the agents' perspectives and can be more costly to implement on the part of the principal. While these issues are not the focus of this chapter, they are important practical considerations that should be taken into account when designing and implementing incentive schemes. Future work to understand and address these practical considerations would be crucial in applying the ideas and techniques presented in this chapter to large-scale practical scenarios.

5

A General Logic-Based Approach to Automatic Incentive Design

In this chapter, we introduce a new class of problems for equilibrium verification and synthesis in multi-agent systems. In our model, agents select strategies to maximise their utility functions, which are expressed as formulae in $\mathsf{LTL}[\mathcal{F}]$, a quantitative extension of Linear Temporal Logic with functions computable in polynomial time. Quantitative temporal specifications are a natural extension of the qualitative ones that we saw in the previous chapter, as they allow us to reason about scenarios with quantitative objectives, such as the maximisation of a particular cost or reward function. Although in Chapter 4, we considered agents with mixed qualitative and quantitative utilities, these were treated separately and combined using a lexicographic ordering. In $\mathsf{LTL}[\mathcal{F}]$, we can express the quantitative objectives directly as a logical formula over real-valued (rather than binary-valued) variables in the system.

As usual in game theory, we consider agents who seek to maximise their personal utility function. In our setting, such utilities are expressed as formulae of $\mathsf{LTL}[\mathcal{F}]$ [47], a quantitative extension of Linear Temporal Logic [31] with functions that are computable in polynomial time. This enables the representation of agents' utilities as *quantitative temporal specifications*. The rationality of an outcome is defined with respect to game theoretic solution concepts, including *dominant strategy equilibrium*, *Nash equilibrium*, and *resilient equilibria*. For each of these solution concepts, we study the problem of finding and verifying an appropriate

incentive mechanism to achieve a societal objective, be it the social welfare or any other aggregating measure of collective wellbeing. To address these challenges, we introduce two problems: incentive verification and incentive synthesis. In incentive *verification*, the aim is to verify whether the application of a given incentive scheme guarantees some level of satisfaction for a given objective. In incentive *synthesis*, we are interested in finding an incentive scheme, if it exists, which guarantees that the satisfaction value of some objective attains at least a certain level.

We focus on two major classes of incentive mechanisms: *static* and *dynamic*. The former assigns new satisfaction values to some of the variables in the system, regardless of the execution history. The latter dynamically reassigns satisfaction values based on the history of the game thus far. We demonstrate (as in Chapter 4) that dynamic incentives are more powerful than their static counterparts. For all solution concepts that we consider, we show that incentive verification is 2EXPTIME-complete for both static and dynamic incentives, and is hence no harder than the corresponding qualitative rational verification problems. Finally, we solve the incentive synthesis problem and prove that it is 2EXPTIME-complete in the static case and likewise 2EXPTIME-complete for dynamic existential synthesis. Only dynamic universal synthesis incurs a higher cost, falling in 3EXPTIME.

5.1 Related work

The model and approach presented in this chapter is inspired by the rational verification and synthesis literature [142, 264–268].

As highlighted in [11], the fields of economics, control theory, and machine learning have studied the problem of designing incentives. The economics approach typically focuses on the treatment of information asymmetries and the design of incentive-compatible mechanisms [170, 269]. The control theory approach introduces variables to encode information about the evolution of the environment, as it depends on agents' choices [270, 271]. Finally, learning approaches to dynamic incentives in multi-agent settings focus solely on adaptively modifying the rewards for quantitative reward-maximising agents [233, 236, 253]. Our approach to incentive design is rooted in formal methods, which guarantees the correctness of incentive schemes with respect to the input specification. Instead of specifying a particular global objective (such as utilitarian social welfare maximisation) and then designing a solution to achieve it in a wide

variety of environments, our approach is more general in that we allow arbitrary global objectives specified in $\mathsf{LTL}[\mathcal{F}]$ to be provided as input to our procedures.

As with the previous two chapters, the core ideas that underpin our approach are inspired by the work of [224], who first considered taxation schemes for one-shot Boolean games. This approach was later extended to concurrent games where players have lexicographic LTL and mean-payoff objectives [272]. We built upon this in Chapter 3 by considering the case where the principal can only observe some subset of the system variables and then extended it to the domain of concurrent games with dynamic incentives in Chapter 4. The model of incentives in this chapter is more general, as we allow for changes to any of the atomic features of the game states. Here, we consider a setting in which both the agents *and* the incentive designer have goals expressed in a quantitative temporal logic.

5.2 Preliminaries

For the remainder of the chapter, we fix a set of atomic propositions Φ, a set of agents N, and a set of strategy variables Var. We also let $\mathcal{F} \subseteq \{f \colon [-1,1]^m \to [-1,1] \mid m \in \mathbb{N}\}$ be a set of functions computable in polynomial time over $[-1,1]$ of possibly different arities, that will parameterise the logics we consider. With slight abuse of notation, we denote by $f \in \mathcal{F}$ both the function and the corresponding function symbol. It will be clear from the context what the symbol corresponds to.

In this chapter, we consider a variant of the classic model of concurrent game structures (CGS) [38], where atomic propositions are associated with a weight in different states:

Definition 5.1. A *weighted concurrent game structure* ($\mathcal{A}$) is a tuple

$$\mathcal{A} = (\Phi, N, (\vec{A}_i)_{i \in N}, S, s^0, \mathcal{T}, \ell),$$

where

1. Φ is a finite set of *atomic propositions*;
2. $N = \{1, \ldots, n\}$ is a finite set of n *agents*;
3. $\vec{A}_i$ is a finite set of *actions* for agent i and $\vec{A} = \bigtimes_{i \in N} \vec{A}_i$ is the set of *joint actions*;

4. S is a finite set of *positions*;
5. $s^0 \in S$ is an *initial state*;
6. $\mathcal{T} : S \times \vec{A} \to S$ is a *transition function*;
7. $\ell : S \times \Phi \to [-1, 1]$ is a *weight function*.

In a state $s \in S$, each player i chooses an action $a_i \in \vec{A}_i$, and the state proceeds to the state $\mathcal{T}(s, \vec{a})$ where $\vec{a}$ is an *action profile* $(a_i)_{i \in N}$.

A *run* $\rho = s^0 s^1 \ldots$ is an infinite sequence of states such that for every $t \geq 0$ there exists an action profile $\vec{a}$ such that $\mathcal{T}(s^t, \vec{a}) = s^{t+1}$. We write $\rho^t = s^t$ for the state at index i in run ρ and $\rho^{\geq t} = s^t s^{t+1} \ldots$ for the suffix of ρ starting from the state at index t. A *history* h is a finite prefix of a run, last(h) is the last state of history h, $|h|$ is the length of h and $\mathbb{H}$ is the set of histories.

We will find it useful to generalise the weight function so that it may vary over the course of a run. Given a set Φ of atomic propositions, an infinite sequence of weight functions $\vec{\ell} = \ell^0 \ell^1 \ldots$, and a run ρ, the $(\Phi, \vec{\ell})$-*labelling* of ρ, written $L(\rho; \Phi, \vec{\ell})$, is given by the infinite sequence of weight assignments

$$L(\rho; \Phi, \vec{\ell}) := (\ell^0(s^0, p))_{p \in \Phi} (\ell^1(s^1, p))_{p \in \Phi} \ldots$$

that is obtained by evaluating each weight function ℓ^t in the sequence $\vec{\ell}$ at the corresponding state s^t over the set Φ. In a standard wCGS, the sequence $\vec{\ell}$ is given by the repeating sequence $\vec{\ell} = \ell\ell\ell \ldots$. Later, we explore the possibility for an incentive designer to dynamically modify the weight functions, leading to a non-trivial sequence $\vec{\ell}$.

A *strategy* for an agent i is a function $\sigma_i : h \to \vec{A}_i$ that maps each history to an action. We let Σ_i be the set of strategies for i and let $\Sigma = \prod_{i \in N} \Sigma_i$ be the set of strategy profiles. An *assignment* $\mathscr{A} : N \cup \text{Var} \to \bigcup_{i \in N} \Sigma_i$ is a function from players and variables to strategies. For an assignment $\mathscr{A}$, an agent i, and a strategy σ for i, $\mathscr{A}[i \mapsto \sigma]$ is the assignment that maps i to σ and is otherwise equal to $\mathscr{A}$, and $\mathscr{A}[x \mapsto \sigma]$ is defined similarly, where x is a variable. For a strategy profile $\vec{\sigma} = (\sigma_i)_{i \in N}$, we write $\mathscr{A}[N \mapsto \vec{\sigma}]$ for the assignment where $\mathscr{A}[i \mapsto \sigma_i]$ for each $i \in N$.

For an assignment $\mathscr{A}$ and a history h, we let $\vee(\mathscr{A}, h)$ be the unique run that continues h following the strategies assigned by $\mathscr{A}$. Formally, $\vee(\mathscr{A}, h)$ is the run $hs^0 s^1 \ldots$ such that for all $t \geq 0$, $s^t = \mathcal{T}(s^{t-1}, \vec{a})$ where for all $i \in N$, $\vec{a}_i = \mathscr{A}(i)(hs^0 \ldots s^{t-1})$, and $s^{-1} = \text{last}(h)$.

Example 5.1. Consider a scenario involving two manufacturing firms who share the usage of a river and a village community who live downstream of the firms on the same river. We represent this by the set $N = \{1, 2, 3\}$, where agents 1 and 2 are the firms. At every time step, each firm has two choices: to discharge waste water directly into the river, or to first treat the waste water to clean it before discharging it into the river. We represent this by letting $\vec{A}_i = \{c_i, d_i\}, i \in \{1, 2\}$ to denote the 'clean' and 'discharge' actions for the two firms respectively. The c_i action incurs a fixed cost for firm i, whereas the d_i action incurs no immediate cost for the firm. At every time step, the village community also has two choices: to use the river water for its day-to-day activities, or to use another water source. We denote this by the actions $\vec{A}_3 = \{r, o\}$, which stand for using the 'river' or an 'other' source of water respectively. The use of the river by the village incurs a cost proportional to the level of contamination of the river, up to a threshold of k units, in which case the river becomes unusable and incurs a very high cost. The use of the other source of water incurs a high but fixed cost for the community.

In this setting, suppose that the river carries water away (whether clean or dirty) at a constant rate of 1 unit per timestep. We can define states to represent varying degrees of contamination of the river water. Let $S = \{s_0\} \cup \{s^{i,\vec{a}} : i \in \{0, \ldots, k\}, \vec{a} \in \vec{A}\}$,[1] where $k \in \mathbb{N}_+$ is a threshold such that if the river contamination level ever reaches k units, the river ecosystem is destroyed and the water becomes permanently unusable. The river starts in a clean state ($s_0 = s^0$), and the transition function between intermediate states $s^{i,\vec{a}}, i \in \{1, \ldots, k-1\}$ can be represented by $\mathcal{T}(s^{i,\vec{a}}, \vec{a}\,') = v_j^{\vec{a}\,'}$, where $j = \max(0, \min(k, i + (D - C)/2))$ and D represents the number of agents who selected the d action in $\vec{a}$ and C represents the number of agents who selected the c action. Transitions out of the initial state are defined in the same manner, and for the ruin states $s_k^{\vec{a}}$, we let $\mathcal{T}(s_k^{\vec{a}}, \vec{a}\,') = s_k^{\vec{a}\,'}$. In other words, if both firms choose the discharge action, the water quality deteriorates by 1 unit, up to the point of ruin. If only one firm chooses the discharge action, the overall quality remains as it is, and if both firms choose the clean action, the river quality improves by 1 unit, up to a maximum level.

We use an atomic proposition q to represent water quality. We can assign $\ell(s^{i,\vec{a}}, q) = 1 - 2i/k$ for all $\vec{a} \in \vec{A}$, so that the water quality is 1 when the river is perfectly clean and -1 when the river is unusable.

[1] The duplication of states for each joint action is used to record the previously taken joint action.

In order to model the effects of agent actions on their objectives, we additionally assume that the game has atomic propositions $\{u_1, u_2, u_3\}$, which represent the 'utility' that each player achieves at a particular state. For a firm $i \in \{1, 2\}$, the weight function for these variables at a state $s^{i,\vec{a}}$ is given by $\ell(s^{i,\vec{a}}, u_i) = 0.2$ if $\vec{a}_i = c_i$ and $\ell(s^{i,\vec{a}}, u_i) = 0.4$ if $\vec{a}_i = d_i$. For the community, we let $\ell(s^{i,\vec{a}}, u_3) = \ell(s^{i,\vec{a}}, q)$ if $\vec{a}_3 = r$, $\ell(s^{i,\vec{a}}, u_3) = 0$ if $\vec{a}_3 = o$. Finally, we let $\ell(s_0, q) = 1$ and $\ell(s_0, p) = 0$ for all $p \in \Phi \setminus \{q\}$.

Quantitative Linear Temporal Logic. To specify players' objectives and define the concept of a rational outcome, we make use of Linear Temporal Logic with quantitative semantics (LTL[$\mathcal{F}$]) [47]. We work with this version over LTL primarily because assuming that formula satisfaction values (over which utilities are typically defined) only take on the values of 0 or 1 limits the kinds of incentive schemes that an incentive designer can devise and implement. With quantitative logics such as LTL[$\mathcal{F}$], there is much more flexibility in capturing the more common conception of an incentive. A classic example of this is in income taxes where the utility of employees depends on their net income, which cannot be reduced to a Boolean outcome.

Definition 5.2. The syntax of LTL[$\mathcal{F}$] is defined by the BNF

$$\varphi ::= p \mid f[\varphi, \ldots, \varphi] \mid \mathbf{X}\varphi \mid \varphi \,\mathbf{U}\, \varphi,$$

where $p \in \Phi$ and $f \in \mathcal{F}$.

Here, $\mathbf{X}$ and $\mathbf{U}$ are the usual temporal operators "next" and "until." The meaning of $f[\varphi_1, \ldots, \varphi_n]$ depends on the function f. We use $\top$, $\vee$, and $\neg$ to denote, respectively, function 1, function $x, y \mapsto \max(x, y)$ and function $x \mapsto -x$.

Definition 5.3. Let $\mathcal{A} = (\Phi, S, N, (\vec{A}_i)_{i \in N}, \mathcal{T}, \ell, S^0)$ be a wCGS, and $\mathscr{A}$ an assignment. The satisfaction value $[\![\varphi]\!]^{\mathcal{A}}_{\mathscr{A}}(h) \in [-1, 1]$ of an LTL[$\mathcal{F}$] formula φ in a history h is defined as follows, where ρ denotes $\vee(\mathscr{A}, h)$:

$$\begin{aligned}
[\![p]\!]^{\mathcal{A}}_{\mathscr{A}}(h) &= \ell(\mathsf{last}(h), p) \\
[\![f[\varphi^1, \ldots, \varphi^m]]\!]^{\mathcal{A}}_{\mathscr{A}}(h) &= f([\![\varphi^1]\!]^{\mathcal{A}}_{\mathscr{A}}(h), \ldots, [\![\varphi^m]\!]^{\mathcal{A}}_{\mathscr{A}}(h)) \\
[\![\mathbf{X}\varphi]\!]^{\mathcal{A}}_{\mathscr{A}}(h) &= [\![\varphi]\!]^{\mathcal{A}}_{\mathscr{A}}(\rho^{|h|+1}) \\
[\![\varphi_1 \,\mathbf{U}\, \varphi_2]\!]^{\mathcal{A}}_{\mathscr{A}}(h) &= \sup_{t \geq 0} \min\Big([\![\varphi_2]\!]^{\mathcal{A}}_{\mathscr{A}}(\rho^{|h|+t}), \\
&\qquad\qquad \min_{0 \leq j < t} [\![\varphi_1]\!]^{\mathcal{A}}_{\mathscr{A}}(\rho^{|h|+j})\Big).
\end{aligned}$$

When the satisfaction value does not depend on the assignment, we write $[\![\varphi]\!]^{\mathcal{A}}(h)$ for $[\![\varphi]\!]^{\mathcal{A}}_{\mathscr{A}}(h)$ where $\mathscr{A}$ is any assignment. Additionally, because arena transitions are deterministic with respect to actions in this setting, every assignment gives rise to a unique run. Therefore, we also define the semantics of $\mathsf{LTL}[\mathcal{F}]$ formulae over a run ρ with respect to a valuation function ℓ over a set Φ, written $[\![\varphi]\!]^{\Phi,\ell}_{\rho}$, in the same manner as above. We define some abbreviations, corresponding to classical logic and LTL operators: $\bot := \neg\top$, $\varphi \wedge \varphi' := \neg(\neg\varphi \vee \neg\varphi')$, $\varphi \rightarrow \varphi' := \neg\varphi \vee \varphi'$, $\mathbf{F}\psi := \top\,\mathbf{U}\,\psi$, $\mathbf{G}\psi := \neg\mathbf{F}\neg\psi$. We assume that $\mathcal{F}$ contains the comparison function

$$\leq : (x, y) \mapsto \begin{cases} 1 & \text{if } x \leq y, \\ -1 & \text{otherwise.} \end{cases}$$

Similarly, the equality function is defined as

$$= : (x, y) \mapsto \begin{cases} 1 & \text{if } x = y, \\ -1 & \text{otherwise.} \end{cases}$$

We assume here that an agent i's decision making aims to maximise the satisfaction value of an $\mathsf{LTL}[\mathcal{F}]$ formula γ_i. Once goals for each agent are specified, we have a *weighted concurrent game*, defined formally as follows:

Definition 5.4. A *weighted concurrent game* (wCG) is a tuple $\mathcal{G} = (\mathcal{A}, (\gamma_i)_{i\in N})$, where $\mathcal{A}$ is a wCGS and γ_i is an $\mathsf{LTL}[\mathcal{F}]$ formula over Φ, for all $i \in N$.

Solution concepts. In this chapter, we consider agents with $\mathsf{LTL}[\mathcal{F}]$ goals, where the formula γ_i denotes the goal of agent i. The utility of agent i in a wCGS $\mathcal{A}$ given the strategy profile $\vec{\sigma} = (\sigma_i)_{i\in N}$ is denoted $\mathrm{util}^{\mathcal{A}}_i(\vec{\sigma})$ and is defined as the satisfaction value of γ_i in the initial state of $\mathcal{A}$ when agents are assigned to their strategies in $\vec{\sigma}$, that is, $\mathrm{util}^{\mathcal{A}}_i(\vec{\sigma}) = [\![\gamma_i]\!]^{\mathcal{A}}_{\mathscr{A}[N\mapsto\vec{\sigma}]}(s^0)$.

We now recall some key game theoretic solution concepts. Let $\mathcal{G} = (\mathcal{A}, (\gamma_i)_{i\in N})$ be a wCG and $\vec{\sigma} = (\sigma_i)_{i\in N}$ be a strategy profile. We define Nash equilibrium in the usual way. We say that σ is a dominant strategy equilibrium if the strategy associated with each agent weakly maximises her utility, for all possible strategies of other agents. Formally, $\vec{\sigma} = (\sigma_i)_{i\in N}$ is a *dominant strategy equilibrium* in $\mathcal{G}$ if for all agents i and for all strategy profiles $\vec{\sigma'} \in \prod_{i\in N} \Sigma_i$, $\mathrm{util}^{\mathcal{A}}_i(\vec{\sigma'}) \leq \mathrm{util}^{\mathcal{A}}_i(\vec{\sigma'}_{-i}, \sigma_i)$.

In an m-resilient equilibrium, we consider deviations by coalitions of agents rather than individuals: this solution concept tolerates deviations of up to m

agents [273]. Formally, $\vec{\sigma} = (\sigma_i)_{i \in N}$ is an *m-resilient equilibrium* in $\mathcal{G}$ if for all coalitions $\mathrm{C} \subseteq N$ with size at most m, all partial strategy profiles $\vec{\sigma'}_{\mathrm{C}} \in \prod_{i \in \mathrm{C}} \Sigma_i$, and all $i \in \mathrm{C}$, we have that $\mathrm{util}_i^{\mathcal{A}}(\vec{\sigma}_{N \setminus \mathrm{C}}, \vec{\sigma'}_{\mathrm{C}}) \leq \mathrm{util}_i^{\mathcal{A}}(\vec{\sigma})$.

We write $\mathrm{DSE}(\mathcal{G}), \mathrm{NE}(\mathcal{G}), \mathrm{RE}_m(\mathcal{G})$ to denote the sets of dominant strategy equilibria, Nash equilibria, and m-resilient equilibria of a game $\mathcal{G}$, respectively.

Example 5.2. Continuing from Example 5.1, we assume that the objective of all players is to maintain a high level of utility at all times. Suppose in addition that there is a regulator who is able to impose taxes on the firms. We can model this by introducing two new variables T_1, T_2 to represent the taxes imposed on firms 1 and 2 respectively. Taxes are initially set to 0, i.e., $\ell(s, T_i) = 0$ for all $s \in S, i \in \{1, 2\}$. Given this, we define the goals for the players on a given run π to be $\gamma_i = \mathbf{G}(u_i - T_i)$ for $i \in \{1, 2\}$ and $\gamma_3 = \mathbf{G}u_3$, where the '$-$' function is defined in the obvious manner.

In this scenario, the two firms do not internalise the effect of their actions on the quality of the river. Since discharging wastewater directly into the river is cheaper for them, there is a Nash equilibrium where both firms choose the d action every round, leading to the spoiling of the river after k time steps. An incentive designer may wish to avoid this scenario by putting in place appropriate incentives to discourage indiscriminate pollution. To realise this aim, we now define our model of incentives.

5.3 Incentives and Decision Problems

Given a wCG $\mathcal{G} = (\Phi, N, (\vec{A}_i)_{i \in N}, S, s_0, \mathcal{T}, \ell, (\gamma_i)_{i \in N})$, we interpret incentives as assignments of satisfaction values to a particular set of propositional variables.

Let $U \subseteq \Phi$ be a set of propositional variables in $\mathcal{G}$. An *incentive scheme* over U is a weight function $\theta : U \to [-1, 1]$ which induces a modified weight function ℓ_θ such that for all $s \in V$ and $p \in \Phi$, we have

$$\ell_\theta(s, p) = \begin{cases} \theta(p) & p \in U \\ \ell(s, p) & p \notin U. \end{cases}$$

Defined in this way, an incentive scheme applied at a particular point in time modifies the satisfaction values of the variables in U. Depending on the interpretation of the variables, this definition of incentives may take on different

meanings. For example, if a variable p represents the monetary cost of being in a particular state, then an incentive on p may be interpreted as a tax or subsidy on the visitation of a state. Moreover, the variables can be used to represent incentives that are agent-specific or ones that are common to multiple agents, depending on the dependencies of utility functions on these variables. Thus, the set U contains the possible variables in the players' shared environment, which the regulator can intervene upon to influence their decision making, and are referred to as *incentivised variables*.

Let Θ denote the set of incentive schemes over a set $U \subseteq \Phi$. We will assume that incentive schemes have a fixed level of granularity. More specifically, we assume that Θ is a finite set of incentive schemes, and we say that an incentive scheme θ has a granularity $g \in \mathbb{Q}^+$ if for all $v \in V, p \in U$, we have $\theta(p) - \ell(v,p) = k \cdot g$, for some $k \in \mathbb{Z}$.

A *dynamic incentive scheme* is a function $T : h \to \Theta$ that maps each history to an incentive scheme that is applied on the same round, i.e., the incentive scheme $\theta^t = T(s^1 \dots s^t)$ is applied in round t of the game. A dynamic incentive scheme thus defines a dynamic modification to the satisfaction values of a fixed subset of the propositional variables Φ defined in a wCG $\mathcal{G}$. In contrast, a *static incentive scheme* is a dynamic incentive scheme that outputs the same incentive scheme for all histories, i.e., $T(h) = T(h')$ for all $h, h' \in \mathbb{H}$. In other words, a static incentive scheme applies a single, fixed modification to the satisfaction values throughout the entire game.

It is relatively straightforward to interpret what the application of a static incentive scheme T over a set U means in a given wCG, and we write $\mathcal{G}_T := \mathcal{G}_{\theta^T} = (\Phi, N, (\vec{A}_i)_{i\in N}, S, s_0, \mathcal{T}, \ell_{\theta^T}, (\gamma_i)_{i\in N})$, where θ^T is the incentive scheme that T outputs at each point in the game. A dynamic incentive scheme, on the other hand, can be thought of as inducing a dynamic (history-dependent) weight function on the game $\mathcal{G}$. In particular, given a run ρ of a game $\mathcal{G}$, a dynamic incentive scheme T induces an infinite sequence $\vec{\ell}_T(\rho) = \ell_{\theta^1}\ell_{\theta^2}\dots$ of weight functions, such that for every history $h = s^1 \dots s^k$ of ρ, we have $\theta^k = T(h)$.

More generally, dynamic weight functions can be accommodated in our model by defining new semantics for the satisfaction value $[\![\varphi]\!]^{\mathcal{A},\vec{\ell}}_{\mathscr{A}}(h)$ of an $\mathsf{LTL}[\mathcal{F}]$ formula in a wCGS $\mathcal{A}$, given an infinite sequence of weight functions $\vec{\ell} = \ell^1\ell^2\dots$ over a finite set $\mathcal{L} = \{\ell^1, \ell^2, \dots, \ell^{|\mathcal{L}|}\}$ of possible weight functions. This is achieved by letting

$$[\![p]\!]^{\mathcal{A},\vec{\ell}}_{\mathscr{A}}(h) := \ell^{|h|}(\mathsf{last}(h), p),$$

while the semantics for f, the $\mathbf{X}$ operator, and the $\mathbf{U}$ operator remain the same. Thus, in the specific context of a dynamic incentive scheme T over a set $U \subseteq \Phi$, we then have

$$[\![p]\!]^{\mathcal{A}_T}_{\mathscr{A}}(h) := [\![p]\!]^{\mathcal{A},\vec{\ell}_T}_{\mathscr{A}}(h) = \begin{cases} T(h)(p) & p \in U \\ \ell(\mathsf{last}(h), p) & p \notin U. \end{cases}$$

We also define the semantics of $\mathsf{LTL}[\mathcal{F}]$ over $(\Phi, \vec{\ell})$-labelled runs $L(\rho; \Phi, \vec{\ell})$, written $[\![\varphi]\!]^{\Phi,\vec{\ell}}_{\rho}$, in the same way.

Due to the fixed granularity assumption, a dynamic incentive scheme's set of possible outputs is always finite. As we demonstrate later, this assumption is important as it allows one to model check $\mathsf{SL}[\mathcal{F}]$ formulae with this modified semantics, albeit at some additional computational cost. However, this does not affect the worst-case complexity results we derive here.

We argue that for most practical purposes, this is a reasonable assumption to make, as it will usually be the case that beyond a certain level of granularity, the benefits of introducing an even finer partition of the incentive space will be outweighed by the additional complexity introduced in computing the optimal incentive scheme and then implementing it in practice. Examples of granularity appearing in real-world incentive policies include taxation rates or the US Federal Reserve's interest rate on reserve balances.

Example 5.3. With incentives defined, we can now consider possible incentive implementations by a regulator in our running example. In most instances, a regulator typically has an *objective* that they would like to see accomplished as the result of the implementation of incentive schemes.

In our case, suppose that the regulator wants to impose taxes on the firms to maintain a stable and acceptable water quality level. To achieve this aim, the regulator can intervene on the T_i variables (i.e., $U = \{T_1, T_2\}$), setting them to some positive taxation rate with a granularity of $g = 0.2$. This objective and the taxation constraints may be expressed by the formula $\varphi = \mathbf{G}(q = 1 \wedge T_1 \geq 0 \wedge T_2 \geq 0)$.

With static incentive schemes, the only way to achieve this is to set the values of T_1 and T_2 so that at least one of the firms is worse off by performing the d action. If only one firm is taxed in this way, the policy may be criticised as being

unfair. If both firms are taxed in this way, this results in an unnecessary loss of profits to both firms. However, if a dynamic incentive scheme is used, which alternates between taxing the firms a sufficient amount for selecting the d action, the regulator can achieve their objective in a fair and efficient manner.

Problem statement. Using the definitions thus far, we are able to state the core problems that we study in the remainder of this text. We begin with the problems in the family of *incentive verification* problems, which ask whether a given static/dynamic incentive scheme guarantees a certain best- or worst-case equilibrium satisfaction value for a given $\mathsf{LTL}[\mathcal{F}]$ formula φ.

ζ-S-E-INCENTIVE-VERIFICATION:
Given: wCG $\mathcal{G}$, $\mathsf{LTL}[\mathcal{F}]$ formula φ,static incentive scheme T, solution concept ζ, threshold $c \in [-1, 1]$.
Question: Is there a strategy profile $\vec{\sigma} \in \zeta(\mathcal{G}_T)$ such that $[\![\varphi]\!]^{\mathcal{A}_T}_{\mathscr{A}[N \mapsto \vec{\sigma}]} \geq c$?

ζ-S-A-INCENTIVE-VERIFICATION:
Given: wCG $\mathcal{G}$, $\mathsf{LTL}[\mathcal{F}]$ formula φ, static incentive scheme T, solution concept ζ, threshold $c \in [-1, 1]$.
Question: Does it hold that for all strategy profiles $\vec{\sigma} \in \zeta(\mathcal{G}_T)$, we have $[\![\varphi]\!]^{\mathcal{A}_T}_{\mathscr{A}[N \mapsto \vec{\sigma}]} \geq c$?

Similarly, we can define the incentive verification problems for dynamic incentive schemes:

ζ-D-E-INCENTIVE-VERIFICATION:
Given: wCG $\mathcal{G}$, $\mathsf{LTL}[\mathcal{F}]$ formula φ, dynamic incentive scheme T, solution concept ζ, threshold $c \in [-1, 1]$.
Question: Is there a strategy profile $\vec{\sigma} \in \zeta(\mathcal{G}_T)$ such that $[\![\varphi]\!]^{\mathcal{A}_T}_{\mathscr{A}[N \mapsto \vec{\sigma}]} \geq c$?

ζ-D-A-INCENTIVE-VERIFICATION:
Given: wCG $\mathcal{G}$, $\mathsf{LTL}[\mathcal{F}]$ formula φ, dynamic incentive scheme T, solution concept ζ, threshold $c \in [-1, 1]$.
Question: Does it hold that for all strategy profiles $\vec{\sigma} \in \zeta(\mathcal{G}_T)$, we have $[\![\varphi]\!]^{\mathcal{A}_T}_{\mathscr{A}[N \mapsto \vec{\sigma}]} \geq c$?

While the ability to verify the effectiveness of a proposed incentive scheme is arguably a useful tool to have, policymakers are often faced with the task of *designing* effective incentive schemes for a given purpose. This motivates the study of what we refer to as the class of *incentive synthesis* problems, which are defined as follows:

ζ-S-E-INCENTIVE-SYNTHESIS:
Given: wCG $\mathcal{G}$, LTL$[\mathcal{F}]$ formula φ, solution concept ζ, incentivised variables $U \subseteq \Phi$, threshold $c \in [-1, 1]$.
Question: Is there a static incentive scheme T over U such that for some strategy profile $\vec{\sigma} \in \zeta(\mathcal{G}_T)$, we have $[\![\varphi]\!]^{\mathcal{A}_T}_{\mathscr{M}[N \mapsto \vec{\sigma}]} \geq c$?

ζ-S-A-INCENTIVE-SYNTHESIS:
Given: wCG $\mathcal{G}$, LTL$[\mathcal{F}]$ formula φ, solution concept ζ, incentivised variables $U \subseteq \Phi$, threshold $c \in [-1, 1]$.
Question: Is there a static incentive scheme T over U such that for all strategy profiles $\vec{\sigma} \in \zeta(\mathcal{G}_T)$, we have $[\![\varphi]\!]^{\mathcal{A}_T}_{\mathscr{M}[N \mapsto \vec{\sigma}]} \geq c$?

ζ-D-E-INCENTIVE-SYNTHESIS:
Given: wCG $\mathcal{G}$, LTL$[\mathcal{F}]$ formula φ, solution concept ζ, incentivised variables $U \subseteq \Phi$, threshold $c \in [-1, 1]$.
Question: Is there a dynamic incentive scheme T over U such that for some strategy profile $\vec{\sigma} \in \zeta(\mathcal{G}_T)$, we have $[\![\varphi]\!]^{\mathcal{A}_T}_{\mathscr{M}[N \mapsto \vec{\sigma}]} \geq c$?

ζ-D-A-INCENTIVE-SYNTHESIS:
Given: wCG $\mathcal{G}$, LTL$[\mathcal{F}]$ formula φ, solution concept ζ, incentivised variables $U \subseteq \Phi$, threshold $c \in [-1, 1]$.
Question: Is there a dynamic incentive scheme T over U such that for all strategy profiles $\vec{\sigma} \in \zeta(\mathcal{G}_T)$, we have $[\![\varphi]\!]^{\mathcal{A}_T}_{\mathscr{M}[N \mapsto \vec{\sigma}]} \geq c$?

Broadly speaking, the incentive synthesis problems ask whether an incentive scheme exists that guarantees at least a certain satisfaction value for a given LTL$[\mathcal{F}]$ formula under some/all stable outcomes of a game, according to some notion of stability. We remark that although these problems are stated in terms of thresholds, our method of solving these problems allows one to find the optimal satisfaction values of the given formula under some/all stable outcomes of a game.

5.4 Static Incentives

In this section, we show how to solve emptiness and synthesis for the static incentive case. We do this by employing procedures for the model checking of formulae in SL$[\mathcal{F}]$, an extension of LTL$[\mathcal{F}]$ that includes strategy quantifiers that are useful to quantify over the strategic abilities of agents [274]. We first recall the syntax and semantics of SL$[\mathcal{F}]$, together with a statement about its

model checking against a wCGS. Subsequently, we show how $\mathsf{SL}[\mathcal{F}]$ can be used to express all the various solution concepts considered in this chapter. Finally, we show how to use such a representation to solve both the verification and synthesis problems for the case of static incentives.

Definition 5.5. The syntax of $\mathsf{SL}[\mathcal{F}]$ is defined by the BNF

$$\varphi ::= \psi \mid \exists x.\, \varphi \mid (i, x)\varphi.$$

where ψ is an $\mathsf{LTL}[\mathcal{F}]$ formula, $x \in \mathrm{Var}$, and $i \in N$.

Intuitively, the new operator $\exists x.\, \varphi$ selects a strategy that maximises the satisfaction value of φ; $(i, x)\varphi$ means that the strategy x is assigned to agent i in the evaluation of φ.

Definition 5.6. Let $\mathcal{A} = (\Phi, N, (\vec{A}_i)_{i \in N}, S, s^0, \mathcal{T}, \ell)$ be a wCGS, and $\mathscr{A}$ an assignment. The satisfaction value $[\![\varphi]\!]^{\mathcal{A}}_{\mathscr{A}}(h) \in [-1, 1]$ of an $\mathsf{SL}[\mathcal{F}]$ formula φ in a history h is defined as follows:

$$[\![\exists x.\, \varphi]\!]^{\mathcal{A}}_{\mathscr{A}}(h) = \max_{\sigma \in \Sigma} [\![\varphi]\!]^{\mathcal{A}}_{\mathscr{A}[x \mapsto \sigma]}(h)$$
$$[\![(i, x)\varphi]\!]^{\mathcal{A}}_{\mathscr{A}}(h) = [\![\varphi]\!]^{\mathcal{A}}_{\mathscr{A}[i \mapsto \mathscr{A}(x)]}(h).$$

The other cases are similar to the semantics of $\mathsf{LTL}[\mathcal{F}]$.

We also use the standard universal quantifiers $\forall x.\, \varphi \coloneqq \neg\exists x.\, \neg\varphi$. For a variable profile $\vec{x} = (x_1, \ldots, x_n)$ with $1 \leq n \leq |N|$, we use the abbreviations $\exists \vec{x}\ \varphi \coloneqq \exists x_1, \ldots, \exists x_n\, \varphi$ and $\forall \vec{x}\ \varphi \coloneqq \forall x_1, \ldots, \forall x_n\, \varphi$.

Definition 5.7 ($\mathsf{SL}[\mathcal{F}]$ model-checking [274])**.** Given an $\mathsf{SL}[\mathcal{F}]$ formula φ, a wCGS $\mathcal{A}$, an assignment $\mathscr{A}$, a history h, and a predicate $P \subseteq [-1, 1]$, decide whether $[\![\varphi]\!]^{\mathcal{A}}_{\mathscr{A}}(h) \in P$.

In [274], it is shown that the model-checking problem for $\mathsf{SL}[\mathcal{F}]$ is decidable and its complexity is $(k+1)$-ExpTime-complete, where k is the *block-nesting* depth of the checked formula. Informally, the block-nesting depth counts how many times an $\mathsf{SL}[\mathcal{F}]$ formula in prenex-normal form alternates between an existential and a universal quantifier.[2]

[2]By prenex normal form, we mean formulae in which all the quantifiers appear in the outermost part of the formula.

Expressing solution concepts in $\mathsf{SL}[\mathcal{F}]$. Let $\vec{x} = (x_i)_{i \in N}$ be a profile of strategy variables and γ_i be an $\mathsf{LTL}[\mathcal{F}]$ formula denoting agent i's goal. The following $\mathsf{SL}[\mathcal{F}]$ formulae express the notion that the strategy profile assigned by $\vec{\sigma}$ is a strategic equilibrium:

$$\mathrm{DSE}(\vec{x}) := \bigwedge_{i \in N} \forall \vec{y}. \Big[(N, \vec{y}) \gamma_i \leq (i, x_i)(N_{-i}, \vec{y}_{-i}) \gamma_i \Big].$$

σ is a *Nash equilibrium* (NE) if no agent can increase her utility with a unilateral change of strategy. We can capture it with the $\mathsf{SL}[\mathcal{F}]$ formula:

$$\mathrm{NE}(\vec{x}) := \bigwedge_{i \in N} \forall y. \Big[(N_{-i}, \vec{x}_{-i})(i, y) \gamma_i \leq (N, \vec{x}) \gamma_i \Big].$$

The strategy profile σ is an *m-resilient* equilibrium if it tolerates deviations of up to m agents of the MAS [273]. Finally, for m-resilient equilibria, we write:

$$\mathrm{RE}_m(\vec{x}) := \bigwedge_{\mathsf{C} \subseteq N : |\mathsf{C}| \leq m} \forall \vec{y}_{\mathsf{C}} \Big[\bigwedge_{i \in \mathsf{C}} \Big((N_{-\mathsf{C}}, \vec{x}_{-\mathsf{C}})(\mathsf{C}, \vec{y}_{\mathsf{C}}) \gamma_i \leq (N, \vec{x}) \gamma_i \Big) \Big].$$

The following proposition relates the satisfaction value of the above formulae to the solution concepts.

Proposition 5.1. *For a wCG $\mathcal{G} = (\mathcal{A}, (\gamma_i)_{i \in N})$, a strategy profile $\vec{\sigma}$ and a variable profile $\vec{x} = (x_i)_{i \in N}$, it holds that:*

- *$\vec{\sigma}$ is a dominant strategy equilibrium iff $[\![\mathrm{DSE}(\vec{x})]\!]^{\mathcal{A}}_{\mathscr{A}[\vec{x} \mapsto \vec{\sigma}]}(h) = 1$;*
- *$\vec{\sigma}$ is a Nash equilibrium iff $[\![\mathrm{NE}(\vec{x})]\!]^{\mathcal{A}}_{\mathscr{A}[\vec{x} \mapsto \vec{\sigma}]}(h) = 1$;*
- *$\vec{\sigma}$ is an m-resilient equilibrium iff $[\![\mathrm{RE}_m(\vec{x})]\!]^{\mathcal{A}}_{\mathscr{A}[\vec{x} \mapsto \vec{\sigma}]}(h) = 1$.*

Proof. These follow from the definitions of $\mathrm{util}^{\mathcal{A}}_i(\vec{\sigma})$ and the solution concepts. □

Using these constructions, we can now address the verification and synthesis problems for static incentives.

Theorem 5.2. *For $\zeta \in \{\mathrm{DSE}, \mathrm{NE}, \mathrm{RE}_m\}, m \in \{1, \ldots, n\}$, both ζ-S-E-INCENTIVE-VERIFICATION and ζ-S-A-INCENTIVE-VERIFICATION are 2EXPTIME-complete.*

Proof. We begin with ζ-S-E-INCENTIVE-VERIFICATION, noting that the approach for the dynamic counterpart is similar. To establish the upper bound, observe that model checking the following $\mathsf{SL}[\mathcal{F}]$ formula in the game $\mathcal{G}_T$ induced by T decides the ζ-S-E-INCENTIVE-VERIFICATION problem:

$$\exists \vec{\sigma} .[\zeta(\vec{\sigma}) \wedge (N, \vec{\sigma})\varphi].$$

By Proposition 5.1 and the semantics of the $\mathsf{SL}[\mathcal{F}]$ operators $\exists \vec{\sigma}$ and $\wedge$, we see that the satisfaction value of the above formula either represents the highest satisfaction value of φ that can be obtained by some ζ-equilibrium of $\mathcal{G}_T$, or it takes on the value of -1. Using this, we can compare the obtained satisfaction value with the threshold c to decide the answer to ζ-S-E-INCENTIVE-VERIFICATION. Membership in 2EXPTIME then straightforwardly follows from the fact that model checking an $\mathsf{SL}[\mathcal{F}]$ formula can be done in $(k+1) -$ EXPTIME for formulas φ where the number of alternations in strategy quantifiers is k [274].

For ζ-S-A-INCENTIVE-VERIFICATION, we can apply the same reasoning as above, only this time with the following $\mathsf{SL}[\mathcal{F}]$ formula:

$$\forall \vec{\sigma} .[\zeta(\vec{\sigma}) \rightarrow (N, \vec{\sigma})\varphi].$$

For hardness, we reduce from qualitative (strong) rational synthesis, which is the problem of deciding whether a strategy profile exists in a concurrent game such that the goal φ_0 of a player 0 is satisfied in some rational outcome of the game while holding this player's strategy fixed [275]. Here, the term 'rational' includes the solution concepts we consider, namely dominant strategy equilibria, Nash equilibria, and h-resilient equilibria. The qualitative component means that player goals are given by formulae expressed in LTL, which is a special case of $\mathsf{LTL}[\mathcal{F}]$ [274], where the following conditions apply:

- For all $s \in S, p \in \Phi$, $\ell(s, p) \in \{-1, 1\}$;
- For all players $i \in N$, γ_i is an LTL formula, i.e., an $\mathsf{LTL}[\mathcal{F}]$ formula where $\mathcal{F} = \{\neg, \vee, \wedge\}$.

The strong version of the rational synthesis problem quantifies universally over rational outcomes, i.e., it asks whether there exists a strategy for player 0 such that φ_0 is guaranteed to be satisfied for *all* rational outcomes for the remaining players. Since the (strong) rational synthesis problem is 2EXPTIME-complete for all solution concepts we consider [275], we can straightforwardly obtain the lower bound by reducing from standard rational synthesis to our variants of the E-INCENTIVE-VERIFICATION problem and from strong rational synthesis to the A-INCENTIVE-VERIFICATION problem by simply verifying an incentive scheme which applies no modifications to the game, where the formula to be

verified is given by φ_0 and the player whose strategy is fixed is an additional inconsequential dummy player who is added to the game. □

For the synthesis of a static incentive, we can proceed similarly to the emptiness problem. Essentially, we first (nondeterministically) guess an appropriate static incentive, then check that it solves the problem by employing the emptiness procedure introduced above.

Theorem 5.3. *For* $\zeta \in \{\mathrm{DSE}, \mathrm{NE}, \mathrm{RE}_m\}, m \in \{1, \ldots, n\}$, *both* ζ-S-E-INCENTIVE-SYNTHESIS *and* ζ-S-A-INCENTIVE-SYNTHESIS *are* 2EXPTIME*-complete.*

Proof. For the upper bound, we proceed by simply nondeterministically guessing an incentive scheme. Since g is assumed to be a rational number, the number of incentive schemes is exponential in the representation of the granularity. Static incentive schemes can be represented in space at most equal to that required for representing the valuation function ℓ, which is hence polynomial in the size of the input. Once this incentive scheme is guessed, we can run ζ-S-E-INCENTIVE-VERIFICATION for the existential version of incentive synthesis or ζ-S-A-INCENTIVE-VERIFICATION for the universal version to obtain a solution to our problems. Since guessing a static incentive scheme is in PSPACE and running an incentive verification procedure is in 2EXPTIME for both cases, the overall complexity of the procedure is in 2EXPTIME.

For the lower bound, we can apply the same reduction as used in Theorem 5.2, but with an additional modification to the game to account for the incentive design component. This can be handled by simply introducing a new propositional variable $q \notin \Phi$ to the game, which is guaranteed not to influence the decision making of agents in any way, and letting the incentive designer intervene only on this new variable, i.e., setting $U = \{q\}$. In this way, there exists an incentive scheme such that the satisfaction value of φ_0 is 1 on some/all ζ-equilibria of the game if and only if the answer to the rational synthesis problem is "yes." □

Given these results, we observe that in the case of static incentives, the incentive verification and synthesis problems have the same worst-case complexity. As our results in the following section will demonstrate, this is not necessarily the case when considering dynamic incentive schemes. This contrast, along with the fact that static incentive schemes are typically simpler to implement and communicate in practice, suggests that it may be better to use static incentives in cases where they are sufficient to implement a goal.

5.5 Dynamic Incentives

Especially in the context of dynamic incentive schemes, a key aspect of our approach is to embed the incentive designer into the game as a player, whose actions correspond to the application of incentives. To achieve this, we introduce a transformation which converts a given wCGS $\mathcal{G}$ into a modified game $\mathcal{G}'$, which naturally encodes the abilities of the incentive designer to modify the values of propositional variables as described above. Intuitively, the transformation works by interleaving actions of the incentive designer (which correspond to incentive schemes) and the actions of the players in the original game.[3] This is done because the incentive designer requires access to the history of the game in order to decide what incentives to implement on a given round, but this decision must be made *before* the next actions of the agents so that it applies on the round in which they take their actions. To perform this interleaving while preserving the satisfaction values of players' goals on the modified runs, we will first need to introduce the notion of run inflations and formula translations [276].

Given an $(\Phi, \vec{\ell})$-labelled run $L_1 = L(\rho; \Phi, \vec{\ell})$ and an $(\Phi', \vec{\ell}')$-labelled run $L_2 = L(\rho'; \Phi', \vec{\ell}')$ where $\Phi \subseteq \Phi'$, we say that L_2 is a *d-fold inflation* of L_1 if it is the case that for all $p \in \Phi$, we have $\ell^{dt'}(s^{dt'}, p) = \ell^t(s^t, p)$ for every $i \in \mathbb{Z}^+$.[4]

We say that a d-fold inflation L_2 of L_1 is *r-labelled* if $r \in \Phi'$ and for all $i, j \geq 0$, we have that $\ell^{j'}(s^{j'}, r) = 1$ if $j = di$ and $\ell^{j'}(s^{j'}, r) = -1$ otherwise. Intuitively, r-labelling is an indicator taking the value of 1 only at the states in the inflated run ρ' that correspond to states in the original run ρ, which will be useful for translating $\mathsf{LTL}[\mathcal{F}]$ formulae over the latter into formulae over the former.

Next, we define a translation function tr^d which transforms an $\mathsf{LTL}[\mathcal{F}]$ formula φ into another $\mathsf{LTL}[\mathcal{F}]$ formula $\mathrm{tr}^d(\varphi)$ that preserves the satisfaction value of φ over a d-fold inflation of any labelled run over which φ is evaluated. More precisely, let φ be an $\mathsf{LTL}[\mathcal{F}]$ formula over a set Φ of atomic propositions. The *d-fold translation* $\mathrm{tr}^d(\varphi)$ of φ is the $\mathsf{LTL}[\mathcal{F}]$ formula defined over the set $\Phi \cup \{r\}$, where r is a new atomic proposition, that is given by the following semantics:

- $\mathrm{tr}^d(p) = \mathbf{X}^{d-1}p$ for every $p \in \Phi$;

[3] This transformation, while conceptually simple, does come with the restriction that the incentive designer's action set must be finite and therefore cannot be used to model incentive schemes that take on a continuous range of values.

[4] Note that the usage of this term differs from that in [215, 277], because we only require that the states be indistinguishable from the perspective of the valuation functions.

- $\mathrm{tr}^d(f[\varphi_1, \ldots, \varphi_m]) = f[\mathrm{tr}^d(\varphi_1), \ldots, \mathrm{tr}^d(\varphi_m)]$;
- $\mathrm{tr}^d(\mathbf{X}\varphi) = \mathbf{X}^d\mathrm{tr}^d(\varphi)$;
- $\mathrm{tr}^d(\varphi_1 \,\mathbf{U}\, \varphi_2) = (r \rightarrow \mathrm{tr}^d(\varphi_1)) \,\mathbf{U}\, (r \rightarrow \mathrm{tr}^d(\varphi_2))$

where X^e represents an application of the $\mathbf{X}$ operator e times. The following is an adaptation of Lemma 1 in [277] to the setting of $\mathsf{LTL}[\mathcal{F}]$ formulae that we consider.

Lemma 5.4. *Let Φ and Φ'' be two disjoint sets of atomic propositions with $r \in \Phi''$. Let $L_1 = L(\rho; \Phi, \vec{\ell})$ be an $(\Phi, \vec{\ell})$-labelled run and suppose that $L_2 = L(\rho'; \Phi', \vec{\ell}')$ is an r-labelled, d-fold inflation of L_1 with $\Phi' = \Phi \cup \Phi''$. Then, for all $\mathsf{LTL}[\mathcal{F}]$ formulae φ over L_1, it holds that $[\![\varphi]\!]_{\rho}^{\Phi,\vec{\ell}} = [\![tr^d(\varphi)]\!]_{\rho'}^{\Phi',\vec{\ell}'}$.*

Proof. The proof goes by structural induction on the formula φ, that is, we show that for all $\mathsf{LTL}[\mathcal{F}]$ formulae φ and $i \geq 0$, it holds that $[\![\varphi]\!]_{\rho^{\geq t}}^{\Phi,\vec{\ell}} = [\![\mathrm{tr}^d(\varphi)]\!]_{\rho'^{\geq dt}}^{\Phi',\vec{\ell}'}$. Considering the base case, it is straightforward to see that $[\![p]\!]_{\rho^{\geq 1}}^{\Phi,\vec{\ell}} = [\![\mathrm{tr}^d(p)]\!]_{\rho'^{\geq d-1}}^{\Phi',\vec{\ell}'}$ by the definition of the d-fold inflation. Now, by the induction hypothesis and the definition of the semantics for $[\![f[\varphi_1,\ldots,\varphi_m]]\!]_{\rho^{\geq t}}^{\Phi,\vec{\ell}}$, it follows that

$$
\begin{aligned}
&[\![\mathrm{tr}^d(f[\varphi_1,\ldots,\varphi_m])]\!]_{\rho'^{\geq dt}}^{\Phi',\vec{\ell}'} = [\![f[\mathrm{tr}^d(\varphi_1),\ldots,\mathrm{tr}^d(\varphi_m)]]\!]_{\rho'^{\geq dt}}^{\Phi',\vec{\ell}'} \\
&= f([\![\mathrm{tr}^d(\varphi_1)]\!]_{\rho'^{\geq dt}}^{\Phi',\vec{\ell}'},\ldots,[\![\mathrm{tr}^d(\varphi_m)]\!]_{\rho'^{\geq dt}}^{\Phi',\vec{\ell}'}) \\
&= f([\![\varphi_1]\!]_{\rho^{\geq t}}^{\Phi,\vec{\ell}},\ldots,[\![\varphi_m]\!]_{\rho^{\geq t}}^{\Phi,\vec{\ell}}) = [\![f[\varphi_1,\ldots,\varphi_m]]\!]_{\rho^{\geq t}}^{\Phi,\vec{\ell}}.
\end{aligned}
$$

Moving on to the next operator, suppose that $\varphi = \mathbf{X}\psi$ for some $\mathsf{LTL}[\mathcal{F}]$ formula ψ and recall that by definition, $[\![\mathbf{X}\psi]\!]_{\rho^{\geq t}}^{\Phi,\vec{\ell}} = [\![\psi]\!]_{\rho^{\geq t+1}}^{\Phi,\vec{\ell}}$. By the semantics of the $\mathbf{X}$ operator and the induction hypothesis, we have

$$
\begin{aligned}
[\![\mathbf{X}\psi]\!]_{\rho^{\geq t}}^{\Phi,\vec{\ell}} &= [\![\psi]\!]_{\rho^{\geq t+1}}^{\Phi,\vec{\ell}} = [\![\mathrm{tr}^d(\psi)]\!]_{\rho'^{\geq d(t+1)}}^{\Phi',\vec{\ell}'} \\
&= [\![\mathbf{X}^d\mathrm{tr}^d(\psi)]\!]_{\rho'^{\geq dt}}^{\Phi',\vec{\ell}'} = [\![\mathrm{tr}^d(\mathbf{X}\psi)]\!]_{\rho'^{\geq dt}}^{\Phi',\vec{\ell}'}.
\end{aligned}
$$

Finally, for the until operator, we apply the same reasoning. Suppose that $\varphi = \varphi_1 \,\mathbf{U}\, \varphi_2$. Then, by the semantics of the $\mathbf{U}$ operator and the induction hypothesis,

we have

$$\begin{aligned}
[\![\varphi_1 \,\mathbf{U}\, \varphi_2]\!]^{\Phi,\vec{\ell}}_{\rho\geq t} &= \sup_{j\geq 0} \min\Big([\![\varphi_2]\!]^{\Phi,\vec{\ell}}_{\rho\geq t+j}, \min_{0\leq k<j} [\![\varphi_1]\!]^{\Phi,\vec{\ell}}_{\rho\geq t+k}\Big) \\
&= \sup_{j\geq 0} \min\Big([\![\mathbf{tr}^d(\varphi_2)]\!]^{\Phi',\vec{\ell}'}_{\rho'\geq d(i+j)}, \\
&\qquad \min_{0\leq k<j} [\![\mathbf{tr}^d(\varphi_1)]\!]^{\Phi',\vec{\ell}'}_{\rho'\geq d(i+k)}\Big) \\
&= \sup_{j\geq 0} \min\Big([\![\max(-r, \mathbf{tr}^d(\varphi_2))]\!]^{\Phi',\vec{\ell}'}_{\rho'\geq d(i+j)}, \\
&\qquad \min_{0\leq k<j} [\![\max(-r, \mathbf{tr}^d(\varphi_1))]\!]^{\Phi',\vec{\ell}'}_{\rho'\geq d(i+k)}\Big) \\
&= \sup_{j\geq 0} \min\Big([\![\max(-r, \mathbf{tr}^d(\varphi_2))]\!]^{\Phi',\vec{\ell}'}_{\rho'\geq dt+j}, \\
&\qquad \min_{0\leq k<j} [\![\max(-r, \mathbf{tr}^d(\varphi_1))]\!]^{\Phi',\vec{\ell}'}_{\rho'\geq dt+k}\Big) \\
&= \sup_{j\geq 0} \min\Big([\![r \to \mathbf{tr}^d(\varphi_2)]\!]^{\Phi',\vec{\ell}'}_{\rho'\geq dt+j}, \\
&\qquad \min_{0\leq k<j} [\![r \to \mathbf{tr}^d(\varphi_1)]\!]^{\Phi',\vec{\ell}'}_{\rho'\geq dt+k}\Big) \\
&= [\![\mathbf{tr}^d(\varphi_1 \,\mathbf{U}\, \varphi_2)]\!]^{\Phi',\vec{\ell}'}_{\rho'\geq dt}.
\end{aligned}$$

□

Together, the concepts of inflation and translation allow us to define the *incentive-augmented game* $\mathcal{G}'$, which is a modified version of a game $\mathcal{G}$ which includes the incentive designer as a player whose actions represent the imposition of different incentive schemes. Given a wCG $\mathcal{G} = (\Phi, N, (\vec{A}_i)_{i\in N}, S, s^0, \mathcal{T}, \ell, (\gamma_i)_{i\in N})$, the *incentive-augmented* version of $\mathcal{G}$ is defined as the wCG

$$\mathcal{G}' = (\Phi', N', (\vec{A}_i)_{i\in N'}, S', s^0, \mathcal{T}', \ell', (\gamma'_i)_{i\in N}),$$

where

- $\Phi' = \Phi \cup \{r\}$, where $r \notin \Phi$.
- $N' = N \cup \{n+1\}$, where player $n+1$ corresponds to the incentive designer;
- $\vec{A}_i$ remains unchanged for all $i \in N$, and $\vec{A}_{n+1} = \Theta$;
- $S' = S \cup S_\Theta$, where $S_\Theta = \{s_\theta | s \in S, \theta \in \Theta\}$;
- $\mathcal{T}' : S' \times \bigtimes_{i\in N'} \vec{A}'_i \to S'$ is defined in the following way. For all $s \in S, \vec{a}\,' \in \vec{A}\,'$ with $\vec{a}\,'_{n+1} = \theta \in \Theta$, we have $\mathcal{T}'(s, \vec{a}\,') = s_\theta$ and for all $s \in S_\Theta, \vec{a}\,' \in \vec{A}\,'$ with $\vec{a}\,'_{-n+1} = (a_i)_{i\in N}$, we have $\mathcal{T}'(s_\theta, \vec{a}\,') = \mathcal{T}(s, (a_i)_{i\in N})$;

- $\ell' : S' \times \Phi \to [-1, 1]$ is defined such that for all $s \in S, p \in \Phi$, we have $\ell'(s, p) = 0$ and for all $s_\theta \in S_\Theta$, we have

$$\ell'(s_\theta, p) = \begin{cases} \theta(p), & p \in U \\ \ell(v, p), & p \notin U. \end{cases}$$

  Moreover, for all $s \in S'$, we have

$$\ell'(s, r) = \begin{cases} -1, & s \in S \\ 1, & s \in S_\Theta; \end{cases}$$

- $\gamma'_i = \mathsf{tr}^2(\gamma_i)$.

This translation of goals is well-defined due to the addition of the r variable which is used to r-label runs in $\mathcal{G}'$. Intuitively, $\mathcal{G}'$ introduces a new player – the 'incentive player' – whose actions correspond to the application of incentives at each round of the original game. The incentive player's actions are taken in an alternating manner with the rest of the players in the states S, which are interleaved with the new states S_Θ such that at every odd timestep $t = 2k + 1, k \in \mathbb{N}$, the game will be in a state $s^t \in S$, whereas at every positive even timestep $t' = 2k, k \in \mathbb{Z}^+$, the game will be in a state $s^{t'}_\theta \in S_\Theta$. In this way, at states $s \in S$, only the incentive player's action determines the next state and at states $s_\Theta \in S_\Theta$, only the original players' actions (i.e., players $1, \ldots, n$) will determine the next state.

Due to the definition of the modified valuation function ℓ' and the 2-fold translation of the players' objectives, the incentive schemes chosen by the incentive player at each of their turns modify the satisfaction values of the players' objectives, influencing the players' strategic considerations.

Moreover, it is straightforward to see that once the incentive player's strategy has been fixed, this induces a new game with potentially different stable outcomes from the original game. Under this construction, the task of the incentive player can then be formulated as the problem of choosing a strategy such that some or all of the induced equilibria of the resulting game for the remaining players maximises the satisfaction value of the incentive designer's objective. The following result establishes the relationship between strategies for the incentive player in an incentive-augmented game and the application of dynamic incentive schemes in the original game.

Proposition 5.5. *Suppose that $\mathcal{G}$ is a weighted concurrent game, $\mathcal{G}'$ is its incentive-augmented version, φ is an $\mathsf{LTL}[\mathcal{F}]$ formula, and $(\vec{\sigma}', \mathcal{T})$ is a strategy profile in $\mathcal{G}'$.*[5] *Then, there is a strategy profile $\vec{\sigma}$ in $\mathcal{G}$ and a dynamic incentive scheme T such that*

$$[\![\varphi]\!]^{\mathcal{A}_T}_{\mathscr{A}[N\mapsto(\vec{\sigma})]} = [\![tr^2(\varphi)]\!]^{\mathcal{A}'}_{\mathscr{A}[N'\mapsto(\vec{\sigma}',\mathcal{T})]}.$$

Proof. We prove the claim by constructing the strategy profile $\vec{\sigma}$ and the dynamic incentive scheme T. Firstly, let $\vec{\sigma}$ be the strategy profile in $\mathcal{G}$ such that for all even-length histories $h' = s^1 s^1_{\theta^1} s^2 s^2_{\theta^2} \dots s^k s^k_{\theta^k}, k \in \mathbb{Z}^+$ that are generated by $\vec{\sigma}'$ in $\mathcal{G}'$, it holds that $\vec{\sigma}(h) = \vec{\sigma}'(h')$, where $h = s^1 s^2, \dots, s^k$. This strategy profile will be referred to as the *2-fold deflation* of the strategy profile $\vec{\sigma}'$, which extracts the actions that players take at every second round of the game $\mathcal{G}'$. Next, let T be the dynamic incentive scheme such that for any given history $h' = s^1 s^1_{\theta^1} \dots s^{k-1}_{\theta^{k-1}} s^k$ of a run ρ' in $\mathcal{G}'$ and the corresponding history $h = s^1 s^2 \dots s^k$ of a run ρ in $\mathcal{G}$ resulting from the execution of $\vec{\sigma}$, we have that $\mathcal{T}(h') = T(h)$. By the construction of the game $\mathcal{G}'$, we see that the incentive scheme θ^k corresponding to the chosen action $\mathcal{T}(h')$ is implemented in state $s^k_{\theta^k}$, that is, for all $p \in \Phi$, we have $\ell'(s^k_{\theta^k}, p) = \vec{\ell}_T(h) = \ell_{\theta^k}$. In other words, the satisfaction values for all variables on round $2k, k \in \mathbb{Z}^+$ of the game $\mathcal{G}'$ coincide with those on round k of the game $\mathcal{G}$.

Now, let ρ' be the run generated in $\mathcal{G}'$ by $(\vec{\sigma}', \mathcal{T})$, $\vec{\ell}' = \ell'\ell' \dots$, ρ be the run generated in $\mathcal{G}$ by $(\vec{\sigma})$, and $\vec{\ell}_T(\rho) = \ell_{\theta^1}\ell_{\theta^2} \dots$. Then, we can conclude that the $(\Phi', \vec{\ell}')$-labelled run $L_2 = L(\rho'; \Phi', \vec{\ell}')$ in $\mathcal{G}'$ is an r-labelled d-fold inflation of $L_1 = L(\rho; \Phi, \vec{\ell}_T(\rho))$ in $\mathcal{G}$ and hence, the result follows by the application of Lemma 5.4. □

Finally, we turn our attention to the dynamic incentive verification and synthesis problems. Again, we are able to make use of the expressive power of $\mathsf{SL}[\mathcal{F}]$ to define a suitable formula which can be model-checked to determine the solution to our problems.

Theorem 5.6. *For $\zeta \in \{\mathrm{DSE}, \mathrm{NE}, \mathrm{RE}_m\}, m \in \{1, \dots, n\}$, ζ-*D-E-INCENTIVE-VERIFICATION *and ζ-*D-A-INCENTIVE-VERIFICATION *are* 2EXPTIME*-complete.*

Proof. Regarding ζ-D-E-INCENTIVE-VERIFICATION, we use the construction $\mathcal{G}'$, which introduces the incentive designer as a player in the game, whose actions are interleaved with those of the original players N. Given this, we can solve the

[5] Here, $\mathcal{T}$ denotes the incentive player's strategy in the augmented game $\mathcal{G}'$, and is distinct from the transition function $\mathcal{T}$ of a wCGS.

incentive verification problem by model checking the following $\mathsf{SL}[\mathcal{F}]$ formula in the game $\mathcal{G}'$:

$$(n+1, \mathcal{T})\,[\exists \vec{\sigma}\,'.[\zeta(\vec{\sigma}\,') \wedge (N, \vec{\sigma}\,')\mathsf{tr}^2(\varphi)]],$$

where the strategy $\mathcal{T}$ is defined such that for all histories $h = s^1 s^2 \dots s^k \in \mathbb{H}$ of the original game $\mathcal{G}$ and a corresponding history $h' = s^1 s^1_{\theta^1} s^2 s^2_{\theta^2} \dots s^k_{\theta^k}$, we have that $\mathcal{T}(h') = T(h)$. By the construction of $\mathcal{G}'$, strategies for the original players in N are inconsequential on odd timesteps, and moreover, the formula $\zeta(\vec{\sigma}\,')$ is now defined in terms of the 2-fold translated versions of players' goals.

Thus, by Proposition 5.5, $\zeta(\vec{\sigma}\,')$ has a satisfaction value of 1 in the above formula if and only if the 2-fold deflation of $\vec{\sigma}\,'$ in the original game $\mathcal{G}_T$ is a ζ-equilibrium. Moreover, by Proposition 5.5 again, the satisfaction value of the subformula $(N, \vec{\sigma}\,')\mathsf{tr}^2(\varphi)$ in $\mathcal{G}'$ is equal to the satisfaction value of the formula $(N, \vec{\sigma}\,)\varphi$ in the game $\mathcal{G}_T$. Hence, model-checking this formula indeed solves the ζ-D-E-INCENTIVE-VERIFICATION problem. For all of our solution concepts, this formula has an alternation depth of 1, and hence, the formula can be model-checked in 2EXPTIME.

Similarly, for ζ-D-A-INCENTIVE-VERIFICATION, we model check the formula:

$$(n+1, \mathcal{T})\,[\forall \vec{\sigma}\,'.[\zeta(\vec{\sigma}\,') \rightarrow (N, \vec{\sigma}\,')\mathsf{tr}^2(\varphi)]],$$

which again has an alternation depth of 1 when converted into prenex normal form.

For hardness, we can apply the same reduction as used in the proof of Theorem 5.2, i.e., we reduce from qualitative rational synthesis and its strong variant to the ζ-D-E-INCENTIVE-VERIFICATION and ζ-D-A-INCENTIVE-VERIFICATION problems respectively. Again, we check a dynamic incentive scheme that does not modify any of the variables and use player 0's goal as the global formula to be checked, with a threshold $c = 1$. □

As with static incentive schemes, the incentive verification problem for dynamic incentive schemes remains in 2EXPTIME. However, the following result indicates that we do not necessarily obtain the same complexity when moving from static to dynamic incentive synthesis.

Theorem 5.7. *For* $\zeta \in \{\text{DSE}, \text{NE}, \text{RE}_m\}, m \in \{1, \ldots, n\}$, ζ-D-E-INCENTIVE-SYNTHESIS *is* 2EXPTIME*-complete and* ζ-D-A-INCENTIVE-SYNTHESIS *is in* 3EXPTIME *and is* 2EXPTIME*-hard.*

Proof. Starting with the existential version of the problem, we again make use of the construction $\mathcal{G}'$. The ζ-D-E-INCENTIVE-SYNTHESIS problem can be solved by model-checking the following $\mathsf{SL}[\mathcal{F}]$ formula:

$$\exists \mathcal{T}. [(n+1, \mathcal{T}) \exists \vec{\sigma}\,'. [\zeta(\vec{\sigma}\,') \wedge (N, \vec{\sigma}\,') \mathbf{tr}^2(\varphi)]].$$

By a similar argument as in the proof of Theorem 5.6, model-checking this formula solves the ζ-D-E-INCENTIVE-SYNTHESIS, and we can extract the dynamic incentive scheme T from a witness $\mathcal{T}$ to the outer existential quantifier of the above formula using the method outlined in the proof of Proposition 5.5. Noting that the quantifiers over $\mathcal{T}$ and $\vec{\sigma}\,'$ do not introduce another level of alternation, model checking this formula is also in 2EXPTIME.

For ζ-D-A-INCENTIVE-SYNTHESIS, we cannot group the quantifiers for incentive schemes and the candidate ζ-equilibria, which introduces another level of alternation in the representative $\mathsf{SL}[\mathcal{F}]$ formula:

$$\exists \mathcal{T}. \left[(n+1, \mathcal{T}) \forall \vec{\sigma}\,'. [\zeta(\vec{\sigma}\,') \rightarrow (N, \vec{\sigma}\,') \mathbf{tr}^2(\varphi)]\right].$$

This leads to a 3EXPTIME procedure for the universal incentive synthesis problem. For the lower bounds, one can use the same reduction as in Theorem 5.3, where we set up the game so that the incentive designer has no effect on the rational outcomes of the original rational synthesis problem. The addition of memory for the incentive designer does not impact their capabilities in this setup, so the arguments for the reduction carry through. □

5.6 Discussion

In this chapter, we have introduced an approach to designing incentives for rational agents with $\mathsf{LTL}[\mathcal{F}]$ goals in order to achieve a societal objective. We formalised the problems of verifying and synthesising static and dynamic incentive schemes for a range of solution concepts. We then proved that, for all solution concepts that we consider, the complexity of incentive verification is 2EXPTIME-complete, which is not harder than the corresponding qualitative rational verification problems. In practice, verification is double exponential

only in the sizes of the formulae representing the societal objectives, which are typically small for most applications. In the case of incentive synthesis, we showed it is 2ExpTime-complete in the static case; dynamic existential synthesis is likewise 2ExpTime-complete, and only dynamic universal synthesis escapes this bound, falling in 3ExpTime (with a 2ExpTime lower bound). We conjecture that this 3ExpTime result is a tight upper bound on dynamic universal synthesis.

While the high complexity of the problems studied here may be interpreted as negative results, they nonetheless reveal interesting differences in complexity between the static and dynamic cases and highlight the potential trade-offs between the two approaches – static incentives are conceptually simpler and generally easier to design, verify, and implement, while dynamic incentives are more powerful but come at the cost of increased complexity and potential implementation challenges. In particular, our static incentive synthesis has double exponential cost only in the size of the goals. However, there is also an additional exponential cost in the size of the representation of the granularity, which can be chosen by the incentive designer for their purposes. In practical applications such as setting interest rates or fixing minimum bidding increments in auctions, incentive designers typically use a rather coarse level of granularity, which would not incur a significant blowup in the algorithm's worst-case runtime. For dynamic incentive synthesis, there is a triple exponential cost in the size of the goals, and we again have an exponential cost in the size of the granularity, but the cost is the same as in verification for the remaining parts of the input. Thus, even with a potentially exponential gap between static and dynamic incentive synthesis, for relatively succinct goal specifications, there may not be a significant difference in runtimes between the two approaches.

For future work, a natural next step is to study the synthesis of incentive schemes with minimal modifications to the original game and to understand how incentive design problems differ in the cooperative setting, where agents can make binding agreements to cooperate with each other and play an agreed-upon set of strategies. Due to the expressivity of $\mathsf{SL}[\mathcal{F}]$, many additional requirements or constraints that one may wish to impose on the incentive schemes can also be naturally incorporated into the global goal in our framework, which enables a wide variety of incentive design problems to be solved using the approaches outlined here.

6

Endogenous Incentives: Agents Incentivising Each Other

To conclude Part I of this thesis, we introduce *Energy Reactive Modules Games* (ERMGs), an extension of Reactive Modules Games (RMGs) in which actions incur an energy cost (which may be positive or negative), and the choices that players make are restricted by the energy available to them. RMGs are a succinct representation of concurrent games using a modelling language to encode the rules of the game in terms of allowable actions by the players in different states. As we shall see in this chapter, a significant benefit of using such a language is that it allows us to extend the description language and semantics of the original model to account for resource-bounded agents in a natural and straightforward manner.

In ERMGs, each action is associated with an energy level update, which determines how their energy level is affected by the performance of the action. In addition, agents are provided with an initial energy allowance. This allowance plays a crucial role in shaping an agent's behaviour, as it must be taken into consideration when one is determining their strategy: agents may only perform actions if they have the requisite energy. We begin by studying rational verification for ERMGs and then introduce *Endogenous* ERMGs, where agents can choose to transfer their energy to other agents. This exchange may enable equilibria that are impossible to achieve without such transfers. We study

the decision problem of whether a stable outcome exists under both the Nash equilibrium and Core solution concepts.

Many variations of rational verification have now been studied [150, 268, 277, 278]. In this work, we introduce a variation of the problem in which agents act under energy resource bounds. Specifically, we assume agents are given some initial energy endowment, and subsequently, all actions that the agent performs are assumed to affect this endowment. Actions may generate energy (leading to an increase in the endowment) or consume energy (reducing the endowment). Crucially, agents can only perform actions for which they have sufficient energy. Although we refer to "energy", many possible interpretations are possible for the concept (e.g., money – although we do not investigate this interpretation). Note that energy plays a secondary role in agents' preferences: agents are concerned with achieving their goal, and energy affects preferences only indirectly (by affecting the actions they can perform). This is in contrast to, for example, [224], wherein agents were assumed to be attempting to achieve goals while minimising costs.

To model this setting, we introduce a variation of *Reactive Modules Games* (RMGs) [16]. In our new variation, individual agent actions (specified via guarded commands) are associated with an energy value, which may either increase or decrease the agent's energy endowment. We begin our study by showing that this framework, while providing a very natural platform through which to model resource-bounded multi-agent systems, can in fact be reduced to "classic" RMGs, and as a consequence, the key non-cooperative and cooperative decision problems in rational verification for our new setting are no harder than in the "classic" setting.

We then study an extension of our model called *Endogenous ERMGs*, in which agents may transfer energy to other agents. Such offers change the possible actions that agents may perform, and hence may change the underlying strategic structure of the game: for example, it may make sense for me to "donate" energy to another agent so that it will choose actions that are of benefit to me. This leads to a two-stage game, with a pre-play "offer" phase. Such an exchange may facilitate equilibria that are impossible to achieve without such an exchange. We study the decision problem of whether a stable outcome and offer profile exists under both the Nash equilibrium and Core solution concepts, again showing that the complexities of the associated problems for both solution concepts in Endogenous ERMGs are also 2EXPTIME-complete.

6.1 Related Work

Endogenous games and side-payments Jackson and Wilkie discuss the stability of transfers in their seminal work on endogenous games [279]. In their work, transfers are dependent on the outcome of the resulting game and thus allow agents to make conditional offers to each other. In the context of infinite games, however, energy values may constrain the set of strategies that are available to players in an ERMG, so transfers should be made *before* the game begins to be of use to the agents. This motivates the consideration of *unconditional* energy transfers in the setting we study. One could also extend our model to settings involving dynamic energy transfers, where each agent can transfer energy to other agents at every round of the game. We leave this as an open problem for future investigation. In addition, our model differs from the original endogenous games model in that the resources being transferred in our setting are not intrinsically valued or optimised by the agents.

Much work has been done on exploring pre-play negotiations and side payments in the context of strategic form games [280–284]. Of these, our work aligns most closely with the study conducted by Turrini (2016) [284]. These studies consider one-shot *strategic-form* or *Boolean games* as their focus. In contrast, RMGs are played for an infinite number of rounds, which allows agents' goals to be modelled using expressive logics like LTL. The agents' goals in [284] are given by Boolean formulae, as opposed to LTL formulae in our model. Furthermore, unlike the mentioned approaches, energy transfers in our model do not directly influence the utility functions of the recipient players. Instead, these transfers serve only to improve an agent's capability to achieve their goal. They are not inherently valued or optimised by the agents but are instrumentally beneficial for goal satisfaction. Thus, our study offers a different perspective and a complementary framework for examining resource transfers that primarily aid in goal achievement without being intrinsically valued or optimised by the agents.

Resource-bounded games and logics Resource-bounded games have attracted considerable attention and have been explored in various contexts. *Energy games* [285, 286] and their subsequent extensions are typically played in a two-player, turn-based, zero-sum setting. In *energy parity games* [287], the objective of Player 1 is to satisfy a qualitative parity condition while maintaining a positive energy level. This aligns closely with our model here, as LTL formulae can always be translated into parity conditions [288]. However, the game graphs

in [287] are singly-weighted, preventing a direct reduction of our games to theirs. Energy games with reachability [289] and ω-regular objectives [290] are also studied in the literature, but again focus only on two-player games. Energy games played on multi-weighted graphs are considered in [291, 292], but these settings only consider a quantitative condition, i.e., the players' energy levels. The work [293] is also relevant, studying the existence of *winning* strategies for a team of agents to achieve some LTL formula in one form of concurrent game structures. This is very similar to the notion of a beneficial deviation in the cooperative setting we consider, but we focus on the existence of strategy profiles which are *stable against deviations*.

Many logics have been developed for reasoning about resource-bounded games (see [294] for an extensive overview of such logics). For instance, pe-ATL [295] is an extension of the logic ATL, which can be used to reason about energy parity games involving multiple agents. Alechina et al. introduced RB-ATL [296, 297], another extension which assumed that resources could only be consumed and not replenished. This was then expanded by RB $\pm$ ATL(*) [298, 299], allowing for both resource consumption and production. Pertaining to our work, note that while (RB $\pm$)ATL* can encode statements about the core, it cannot express statements regarding the existence of Nash equilibria. Furthermore, it is not suitable for reasoning about side payments among agents.

Rational Verification This chapter builds upon the existing literature on rational verification and synthesis [16, 142, 266, 275]. While many studies have used RMGs as the model for rational verification (e.g., [16, 147, 277, 300]), few have addressed resource-boundedness. Electric Boolean games [301, 302] consider resource-bounded games like our approach. While these models study resource redistributions, they operate under the assumption that a central authority exists that is capable of redistributing energy endowments. In contrast, the key novel aspect of our model is that we assume agents are independent and can choose their redistributions, giving rise to strategic considerations regarding transfers.

6.2 Simple Reactive Modules with Energy

We use an extension of the *Simple Reactive Modules Language* (SRML) [146] to model agents, which we refer to as *modules*. SRML is a special case of the Reactive Modules framework developed by Alur and Henzinger [145]. An SRML module consists of:

1. An *interface*, which defines the module's name and the Boolean variables under the control of the module; and
2. Two sets of *guarded commands*, which define the choices available to the module at every state.

Interfaces are specified by the syntax `module` m_i `controls` Φ_i, where m_i is the name of the module and $\Phi_i \subseteq \Phi$ is the set of variables under its exclusive control. Guarded commands consist of two parts: a precondition for executing the command, known as the *guard*, and the actual *command*, which specifies how the value of (some of) the variables under the module's control are updated when the command is executed.

In the extension of SRML we work with in this chapter, we augment guarded commands with an additional *energy value*, which specifies how the module i's *energy level* at time t, denoted E_i^t, is changed if the command is executed. A positive energy value will increase available energy at time $t+1$; a negative value will decrease available energy. Thus, positive energy values correspond to gains in energy upon executing the command, whereas negative energy values correspond to costs/losses in energy. Given this, the general form of a guarded command in our augmented SRML is of the form [`energy`] `guard` $\rightsquigarrow$ `command` where `energy` is the associated energy cost/gain. More formally, a *guarded command* g for a module m_i controlling Boolean variables $\Phi_i \subseteq \Phi$ is an expression:

$$[e]\ \varphi \rightsquigarrow x_1' \coloneqq \psi_1; \cdots ; x_k' \coloneqq \psi_k,$$

where $e \in \mathbb{Z}$, φ and each ψ_j is a propositional logic formula over Φ, and every x_j' represents the value of the variable $x_j \in \Phi_i$ after the command is executed. The value of any variable in Φ_i that does not appear in a guarded command g is unchanged by the execution of g. We also require that $x_a' \neq x_b'$ for all $a \neq b \in \{1, \ldots, k\}$, i.e., no variable's value is reassigned twice in the same guarded command. For a guarded command $g = [e]\ \varphi \rightsquigarrow x_1' \coloneqq \psi_1; \cdots ; x_k' \coloneqq \psi_k$, $\eta(g) = e$ represents the change in energy level associated with g, $\mathsf{guard}(g) = \varphi$ represents the guard of g, and $\mathsf{evl}(g) = x_1' \coloneqq \psi_1; \cdots ; x_k' \coloneqq \psi_k$ represents the command of g. For an agent i with energy level E_i and a valuation $\vec{v}$, we say that a guarded command g_i for i is *enabled* if both $-E_i \leq \eta(g_i)$ and $\vec{v} \models \mathsf{guard}(g_i)$ and we write $\mathsf{enable}_i(\vec{v}, E_i)$ for the set of all guarded commands g_i which are enabled for i under $\vec{v}$.

For example, the guarded command $[-2]\ (p \vee q) \rightsquigarrow p' \coloneqq \bot; q' \coloneqq \top$ for a module m_i can be read as "if p or q are true and i has at least two units of energy, then *one of the actions available to* m_i is to set p to $\bot$ and q to $\top$, which incurs an energy cost of 2 units for m_i."

There are two kinds of guarded commands for a module: those used to *initialise* the variables under the module's control, and those used to subsequently *update* the variables. These are represented as sets of guarded commands `init` and `update`, respectively.

To ensure that agents always have a well-defined action available, we assume that in each agent's `init` set, at least one initial guarded command has a non-negative energy value. We also equip every module with a `skip` guarded command as part of its `update` set, which is always available and does nothing to the variables under its control. This can be explicitly written as $[e_i^{skip}]\ \top \rightsquigarrow \emptyset$, where $e_i^{skip} \in \mathbb{N}$ (including 0) and we use $\emptyset$ to denote a command which does nothing to the variables under the module's control. We require that the energy cost for the `skip` command be non-negative to ensure that at least one command is always available to every agent, which entails that runs are always well-specified.

Formally, an *SRML module*, m_i is given by a triple

$$m_i = (\Phi_i, \texttt{init}_i, \texttt{update}_i), \text{ where:}$$

- $\Phi_i \subseteq \Phi$ is the set of variables controlled by m_i;
- $\texttt{init}_i$ is a finite set of *initialisation guarded commands* for m_i; and
- $\texttt{update}_i$ is a finite set of *update guarded commands* for m_i.

An SRML arena is then simply a collection of agents, their representative modules, and the specification of each agent's initial and maximum energy values:

$$A = (N, \Phi, (m_i)_{i \in N}, (e_i^{max})_{i \in N}, (E_i^0)_{i \in N}),$$

where:

- $N = \{1, \ldots, n\}$ is a finite, non-empty set of *agents*;
- $\Phi = \bigcup_{i \in N} \Phi_i$ is a finite, non-empty set of *propositional variables*, where the sets Φ_i are all pairwise disjoint;

- $m_i = (\Phi_i, \texttt{init}_i, \texttt{update}_i)$ is an SRML module over Φ that defines the choices available to agent $i \in N$;
- $e_i^{max} \in \mathbb{N}$ is the *maximum energy capacity* of agent i; [1] and
- $E_i^0 \in \{0, \ldots, e_i^{max}\}$ is the *initial energy level* of i.

An agent module m_i for an agent i is defined by augmenting the set of original variables Φ_i under their control with a set of *command variables*, which are used to identify precisely which action the agent took at each point in time. Given this, the *agent module* for i takes the following general form as in Figure 6.1: (i) for any set $\Psi \subseteq \Phi$ of Boolean variables and a Boolean literal $v \in \{\top, \bot\}$, $\Psi \coloneqq v$ denotes the assignment where all variables in Ψ are set to v; (ii) for any ordered set $\Psi = \{p_1, \ldots, p_m\} \subseteq \Phi$ of Boolean variables and a binary string $B = b_1 b_2 \ldots b_m$, $\Psi' \coloneqq B$ denotes the assignment $p'_1 \coloneqq b_1;\ p'_2 \coloneqq b_2;\ \ldots;\ p'_m \coloneqq b_m$. Finally, for a given energy level E_i of an agent i, we let $\texttt{init}_i|_E$ and $\texttt{update}_i|_E$ denote the set of initial and update guarded commands for i respectively whose corresponding energy values are at least $-E$:

$$\begin{aligned} \texttt{init}_i|_E &\coloneqq \{g \in \texttt{init}_i \mid \eta(g) \geq -E\} \\ \texttt{update}_i|_E &\coloneqq \{g \in \texttt{update}_i \mid \eta(g) \geq -E\}. \end{aligned}$$

These sets will be useful in identifying the set of guarded commands that are available to an agent, given their current energy level at any point in a run.

$\texttt{module}\ m_i\ \texttt{controls}\ \Phi_i$
$\texttt{init}$
$[e_i^1]\ \top \rightsquigarrow \Phi'_i \coloneqq v_i^1$
$\cdots$
$[e_i^{m_1}]\ \top \rightsquigarrow \Phi'_i \coloneqq v_i^{m_1}$
$\texttt{update}$
$\left[e_i^{m_1+1}\right]\ \varphi_{m_1+1} \rightsquigarrow \Phi'_i \coloneqq v_i^{m_1+1}$
$\cdots$
$\left[e_i^{k_1}\right]\ \varphi_{k_1} \rightsquigarrow \Phi'_i \coloneqq v_i^{k_1}$
$\texttt{skip}$

Figure 6.1: A Reactive Module augmented with energy values.

[1] We assume that each agent is equipped with a device that has a finite capacity for storing energy.

6.3 Energy Reactive Modules Games

With these definitions in place, we can define the model for concurrent games that we focus on in this study. Formally, an *Energy Reactive Modules Game* (ERMG) is a structure

$$\mathcal{G} = (\mathcal{A}, \gamma_1, \ldots, \gamma_n),$$

where $\mathcal{A}$ is an SRML arena and γ_i is the LTL goal of agent $i \in N$. At the beginning of a game, each agent $i \in N$ selects an initial guarded command $g_i^0 \in \texttt{init}_i|_{E_i^0}$. The energy level of each agent i at timestep 1 is then updated as $E_i^1 = \min(E_i^0 + \eta(g_i^0), e_i^{max})$ and the initial valuation $\vec{v}_0$ of the variables in Φ is set according to the commands chosen, i.e., $\mathsf{evl}(g_i^0)$.[2] Then, at every step $t \in \mathbb{Z}^+$ of the execution, each agent i selects an enabled update guarded command $g_i^t \in \texttt{update}_i|_{E_i^t}$ to execute, which updates the values of the variables in Φ_i according to $\mathsf{evl}(g_i^t)$ and updates the agent's energy level from E_i^t to $E_i^{t+1} = \min(E_i^t + \eta(g_i^t), e_i^{max})$. This update gives rise to a new valuation $\vec{v}_t$, and then the next round proceeds in the same manner. This process repeats indefinitely, giving rise to a *run*, which in an ERMG is an infinite sequence of valuations $\rho = \vec{v}_0\vec{v}_1\ldots$ where for all $t \in \mathbb{N}$, we have that $\vec{v}_{t+1}$ is obtained from the execution of enabled guarded commands by all agents $i \in N$. Note that given the restrictions on allowable actions at each time point, each agent's energy level will always be a non-negative integer throughout any run.

A particular class of games which can be represented using ERMGs are *regular RMGs*, which are exactly the same as ERMGs, except that no energy values are involved. Regular RMGs can thus be specified by an arena $\mathcal{A} = (N, \Phi, (m_i)_{i\in N})$ and LTL goals $(\gamma_i)_{i\in N}$. Key decision problems in rational verification under the Nash equilibrium solution concept for regular RMGs have been well-studied in prior work [16].

We model strategies as deterministic Mealy machines, which are known to be sufficient for optimality in our setting [16]. A strategy for a player $i \in N$ with associated module $m_i = (\Phi_i, \texttt{init}_i, \texttt{update}_i)$ is a Mealy machine (i.e., a finite state machine with output) $\sigma_i = (Q_i, q_i^0, \delta_i, \tau_i)$, where Q_i is a finite set of machine states, q_i^0 is the initial state, and for all $q \in Q_i, \vec{v} \in \{0,1\}^{|\Phi|}$, we have:

- $\delta_i : Q_i \times \{0,1\}^{\Phi} \to Q_i$ is the strategy's deterministic state update function such that $\delta_i(q, \vec{v}) \neq q_i^0$, i.e., the initial state is never revisited;

[2] It is also possible to consider settings with unbounded energy capacities, but this is not physically realistic and is likely to lead to undecidable verification problems [303].

- $\tau_i(q, \vec{v}) \in \mathsf{enable}_i(\vec{v}, E_i)$ is the output function that specifies which enabled guarded command is selected by player i with energy level E_i, such that $\tau_i(q, \vec{v}) \in \mathtt{init}_i$ iff $q = q_i^0$, i.e., initial guarded commands are only selected at the start of the game.

For each player $i \in N$, we let Σ_i represent their set of possible strategies and write $\Sigma = \prod_{i \in N} \Sigma_i$ to represent the set of all *strategy profiles*, i.e., tuples of strategies for each player. Since we consider deterministic strategies in this setting, a strategy profile $\vec{\sigma}$ deterministically generates a run $\rho(\vec{\sigma}, \mathcal{G})$ in an ERMG $\mathcal{G}$, which consists of the infinite sequence of valuations generated by the execution of enabled guarded commands by modules at each time step.

Given this, we are now in a position to define preferences and utility functions over runs in ERMGs. We assume that a player $i \in N$ has the sole objective of satisfying their LTL goal γ_i. Given this, we say that player $i \in N$ *prefers* the run ρ over ρ', written $\rho \succeq_i \rho'$, if and only if $\rho' \models \gamma_i$ implies that $\rho \models \gamma_i$. This can be extended to the strict preference relation $\succ_i$ and the indifference relation $\sim_i$ in the usual manner (see [304]). Finally, since strategy profiles give rise to fixed runs in an ERMG, we can write $\vec{\sigma} \succeq_i \vec{\sigma}\,'$ as a shorthand for $\rho(\vec{\sigma}, \mathcal{G}) \succeq_i \rho(\vec{\sigma}\,', \mathcal{G})$. Notice that energy does not feature in the definition of preferences: agents are only secondarily concerned about energy consumption, in the sense that it places restrictions on the choices they can make.

Finally, we can define the first solution concept considered in this study. A strategy profile $\vec{\sigma}$ is a *Nash equilibrium* (or Nash-stable strategy profile) of an ERMG $\mathcal{G}$ if for all players $i \in N$ and all alternative strategies $\sigma_i' \in \Sigma_i$, it holds that $\vec{\sigma} \succeq_i (\vec{\sigma}_{-i}, \sigma_i')$. We let $\mathrm{NE}(\mathcal{G})$ denote the set of Nash Equilibria of an ERMG $\mathcal{G}$ and for an LTL formula φ, let $\mathrm{NE}_\varphi(\mathcal{G}) = \{\vec{\sigma} \in \mathrm{NE}(\mathcal{G}) : \rho(\vec{\sigma}, \mathcal{G}) \models \varphi\}$.

Example 6.1. Consider a scenario involving two robots, B_1 and B_2, and three locations: middle (M), left (L), and right (R). From each location, each robot can either stay in the same position by choosing the skip command or move to another location. From M, each robot can move to either L or R. However, from L or R, they can only return to M. A move from one location to another can only be initiated when both robots are at the same location. Initially, both robots are positioned at M. For B_1, the cost of each move is -1. For B_2, moving left costs -1 but moving right costs -2. B_1 (resp. B_2) is assigned to service L (resp. R) at least once. However, for the service to be carried out, both robots must be at the same location. After completing the service and returning to M, B_1 and B_2 receive an energy recharge equal to the amount they have spent. For example, if B_1 and

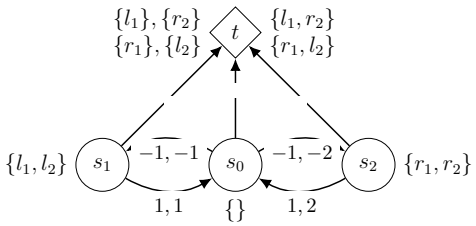

Figure 6.2: The diagram for Example 6.1. The losses/gains of energy for B_1 and B_2 are represented by edge labels. To maintain clarity, some labels are not shown. The nodes s_0, s_1, s_2 symbolise the positions of both robots in M, L, R respectively. The node t is a macrostate encompassing several states, as indicated by its labels. Each of these (macro)states also includes a self-loop edge (not shown) corresponding to the skip command.

B_2 both visit R and then return to M, B_1 gains 1 unit of energy while B_2 gains 2 units. For $i \in \{1, 2\}$, B_i is represented by the following module.

$$
\begin{array}{l}
\texttt{module}\ B_i\ \texttt{controls}\ \{l_i, r_i\} \\
\texttt{init} \\
\quad [0]\ \top \rightsquigarrow l_i' \coloneqq \bot; r_i' \coloneqq \bot \\
\texttt{update} \\
\quad [-1]\ \neg(l_i \vee r_i \vee l_{3-i} \vee r_{3-i}) \rightsquigarrow l_i' \coloneqq \top \\
\quad [-i]\ \neg(l_i \vee r_i \vee l_{3-i} \vee r_{3-i}) \rightsquigarrow r_i' \coloneqq \top \\
\quad [1]\ l_i \wedge l_{3-i} \rightsquigarrow l_i' \coloneqq \bot \\
\quad [i]\ r_i \wedge r_{3-i} \rightsquigarrow r_i' \coloneqq \bot \\
\texttt{skip}
\end{array}
$$

Each robot controls two variables l_i and r_i. If these variables are set to true, it indicates that the robot moves to the left or right. Let $E^0_{B_1} = e^{max}_{B_1} = 1, E^0_{B_2} = e^{max}_{B_2} = 2$. B_i's goal is given by $\gamma_{B_i} = \mathbf{F}\ s_i$. A graphical representation of the game is shown in Fig. 6.2. Consider a strategy profile which gives rise to the initial sequence $\{\}, \{l_1, l_2\}, \{\}, \{r_1, r_2\}$. This is a Nash equilibrium (NE), as the run satisfies $\gamma_{B_1} \wedge \gamma_{B_2}$, and the robots have enough energy to execute this strategy profile. Now, consider the sequence of assignments $\{\}, \{l_1, l_2\}, \{\}, \{r_1\}$. Although the robots also have enough energy to execute this run, it is not a NE. This is because B_2 can choose to set r_2 to true on the fourth round and achieve its goal.

6.4 Rational Verification in ERMGs

Before exploring mechanisms for agents to make energy transfers, it is helpful to first understand the complexity of rational verification in ERMGs. Using a poly-

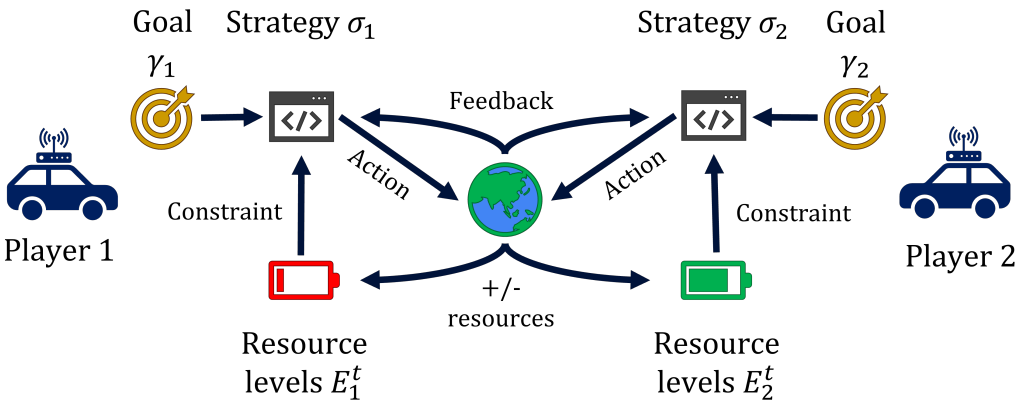

Figure 6.3: Illustration of the ERMG framework. Players possess a resource/energy level as well as a goal specified as an LTL formula. At each round of the game, players choose an action that is available to them, which updates the state of the game as well as their resource levels defined by a pre-specified cost function.

nomial time transformation from ERMGs into regular RMGs, we can establish that these rational verification problems are no harder than their counterparts in regular RMGs. Thus, we have the following results:

Proposition 6.1. NASH-MEMBERSHIP *for ERMGs is* PSPACE*-complete, while* E-NASH *and* A-NASH *for ERMGs are* 2EXPTIME*-complete.*

Proof. We begin by showing that ERMGs can be transformed into regular RMGs in polynomial time. Therefore, these problems can be solved by a reduction to their counterparts in regular RMGs without any energy costs. Given a game $\mathcal{G} = (\mathcal{A}, (\gamma_i)_{i \in N})$, we construct an RMG $\mathcal{G}^* = (\mathcal{A}^*, (\gamma_i)_{i \in N})$, where $\mathcal{A}^* = (N, \Phi^*, (m_i^*)_{i \in N})$ is formed by adding additional variables to keep track of the agents' energy values and augmenting the agent module corresponding to each agent $i \in N$ so that the execution of guarded commands updates these additional energy variables appropriately. Thus, the modifications included in $\mathcal{A}^*$ are:

- $\Phi^* = \Phi \cup \Phi^e$ is the modified set of arena variables, which consists of (i) the original set of variables Φ that are controlled by the agent modules and (ii) a set of variables $\Phi^e = \bigcup_{i \in N} \Phi_i^e$ that encode the energy values of each agent using the standard unsigned binary integer representation, where $|\Phi^e| = \sum_{i=1}^{n} \lceil \log_2(e_i^{max} + 1) \rceil$;
- $m_i^* = (\Phi_i \cup \Phi_i^e, \texttt{init}_i', \texttt{update}_i')$ is an SRML module over Φ^* that defines the choices available to agent $i \in N$.

Augmented agent modules m_i^* thus keep track of their own energy values explicitly, and these form explicit constraints on the actions which they are allowed to take at any given point in time, based on the history of actions taken thus far. Given this, the *augmented agent module* m_i^* for i takes the following general form:

$$
\begin{array}{l}
\texttt{module}\ m_i^*\ \texttt{controls}\ \Phi_i^*\ (=\Phi_i \cup \Phi_i^e) \\
\texttt{init} \\
\quad \top \rightsquigarrow \Phi_i' \coloneqq\ v_i^1;\ \Phi_i^{e'} \coloneqq \min\{\texttt{Bin}(e_i^{max}), \Phi_i^e +^B \texttt{Bin}(e_i^1)\} \\
\quad \cdots \\
\quad \top \rightsquigarrow \Phi_i' \coloneqq\ v_i^{m_1};\ \Phi_i^{e'} \coloneqq \min\{\texttt{Bin}(e_i^{max}), \Phi_i^e +^B \texttt{Bin}(e_i^{m_1})\} \\
\texttt{update} \\
\quad \varphi_{m_1+1} \wedge \left(\Phi_i^e \geq \texttt{Bin}(e_i^{m_1+1})\right) \rightsquigarrow \Phi_i' \coloneqq\ v_i^{m_1+1}; \\
\qquad\qquad\qquad\qquad \Phi_i^{e'} \coloneqq \min\{\texttt{Bin}(e_i^{max}), \Phi_i^e +^B \texttt{Bin}(e_i^{m_1+1})\} \\
\quad \cdots \\
\quad \varphi_{k_1} \wedge \left(\Phi_i^e \geq \texttt{Bin}(e_i^{k_1})\right) \rightsquigarrow \Phi_i' \coloneqq\ v_i^{k_1}; \Phi_i^{e'} \coloneqq \min\{\texttt{Bin}(e_i^{max}), \Phi_i^e +^B \texttt{Bin}(e_i^{k_1})\} \\
\texttt{skip}
\end{array}
$$

where the following notation is used:

- For an integer k, $\texttt{Bin}(k)$ denotes the binary representation of k using two's complement and, for $k \in \mathbb{N}$, $\texttt{Bin}^+(k)$ denotes the unsigned binary integer representation of k;
- $\geq$ is a function that compares two binary numbers in the natural way (returning either $\top$ or $\bot$);
- $+^B$ represents the addition of the binary number encoded by the values of the variables in Φ_i^e with the unsigned binary representation of an associated energy value.

This construction thus explicitly encodes the rules of an ERMG using the RMG framework, and it is clear that an infinite sequence of valuations ρ is a run in an ERMG $\mathcal{G}$ if and only if it is also a run in $\mathcal{G}^*$. Using this reduction, we have the following:

NASH-MEMBERSHIP: The upper bound is straightforward, as we can use the construction $\mathcal{G}^*$ to reduce from our problem to the MEMBERSHIP problem for regular RMGs, which is known to be PSPACE-complete [16]. Clearly, a strategy profile $\vec{\sigma}$ is a Nash equilibrium in $\mathcal{G}$ if and only if it is a Nash equilibrium in $\mathcal{G}^*$.

The lower bound follows by PSPACE-hardness of the MEMBERSHIP problem for regular RMGs, which can trivially be encoded as an ERMG where all energy values are 0.

E/A-NASH: We begin by considering E-NASH. For membership, we can construct the game $\mathcal{G}^*$ in polynomial time and then check E-NASH for regular RMGs on the formula φ, which is known to be 2EXPTIME-complete [16]. Since there is a one-to-one correspondence between valid strategy profiles in $\mathcal{G}$ and in $\mathcal{G}^*$ and their associated runs, any Nash equilibrium $\vec{\sigma}$ in $\mathcal{G}^*$ which satisfies φ must also be a Nash equilibrium in $\mathcal{G}$ which satisfies φ. Hence, E-NASH for regular RMGs given $\mathcal{G}^*$ and φ returns "yes" if and only if the answer to our decision problem is "yes." For hardness, we can simply reduce from E-NASH in regular RMGs to our problem by constructing an ERMG which is the same as the regular one, with the addition of setting all energy costs and values to 0. Moving on to the upper bound for the A-NASH problem, we can simply check E-NASH on $\mathcal{G}$ and the formula $\neg\varphi$. The answer is "yes" if and only if the answer to A-NASH is "no." For the lower bound, we reduce from A-NASH for regular RMGs in the same manner as with the E-NASH problem. □

Cooperative Games

A central assumption in non-cooperative game theory is that agents act independently. In cooperative games, however, we assume that agents are able to form binding agreements with others – that is, agents can form *coalitions*, who may deviate from an outcome if it is mutually beneficial to do so. In the games we consider, the question of whether a coalition has a beneficial deviation depends not only on their choices, but also on those of the remaining non-deviating players, i.e., coalitions face *externalities*. For this reason, we work under the assumption that coalitions are maximally pessimistic; that is, a deviation must be beneficial for the deviating coalition against *all* responses by the remaining players. This corresponds to the well-known α-core, which was introduced by Aumann in [98], and remains a key solution concept in cooperative game theory.

Under this assumption, a strategy profile $\vec{\sigma}$ is said to be *core-stable* in $\mathcal{G}$ if for all coalitions $C \subseteq N \neq \emptyset$ such that $\rho(\vec{\sigma}, \mathcal{G}) \models \bigwedge_{i \in C} \neg\gamma_i$ and partial strategy profiles $\vec{\sigma}'_C = (\sigma'_i)_{i \in C}$, there exists a response by the remaining players $\vec{\sigma}'_{-C} = (\sigma'_i)_{i \in N \setminus C}$ such that for some player $i \in C$, it holds that $\rho((\vec{\sigma}'_C, \vec{\sigma}'_{-C}), \mathcal{G}) \not\models \gamma_i$. We let $\text{CORE}(\mathcal{G})$ denote the set of core-stable strategy profiles in $\mathcal{G}$ and define $\text{CORE}_\varphi(\mathcal{G}) = \{\vec{\sigma} \in \text{CORE}(\mathcal{G}) : \rho(\vec{\sigma}, \mathcal{G}) \models \varphi\}$. In the same manner as with the

Nash equilibrium solution concept, we can now define the cooperative rational verification problems as follows:

E-CORE:
Given: Game $\mathcal{G}$, LTL goal φ.
Question: Is it the case that $\text{CORE}_{\varphi}(\mathcal{G}) \neq \emptyset$?

A-CORE:
Given: Game $\mathcal{G}$, LTL goal φ.
Question: Is it the case that $\text{CORE}_{\varphi}(\mathcal{G}) = \text{CORE}(\mathcal{G})$?

Example 6.2. To illustrate the difference between NE and the core, let us revisit Example 6.1. Consider a strategy profile where both robots remain at location M forever. This strategy profile is a NE (albeit a suboptimal one), as there is no *unilateral* deviation by B_i that satisfies γ_{B_i}: any change in B_i's strategy would result in the game transitioning to state t. However, this strategy profile is not in the core, as the (grand) coalition $\{B_1, B_2\}$ has a strategy profile that ensures the satisfaction of $\gamma_{B_1} \wedge \gamma_{B_2}$. In fact, all strategy profiles in the core satisfy $\gamma_{B_1} \wedge \gamma_{B_2}$. As a result, an A-NASH query with $\varphi = \gamma_{B_1} \wedge \gamma_{B_2}$ would return a negative answer, while an A-CORE query would return a positive one.

Because no results exist for cooperative rational verification in the context of regular RMGs, we solve these problems by reductions to and from concurrent game structures, for which these problems have been studied [151]. Thus, we establish here the first results for cooperative rational verification in the RMGs model as well as in our extension to settings with bounded resources.

Proposition 6.2. E-CORE *and* A-CORE *for ERMGs and regular RMGs are* 2EXPTIME*-complete.*

Proof. We begin by considering E-CORE. The upper bound is obtained by first converting the given ERMG into an LTL game and then model checking a suitable ATL* formula over the LTL game. Such games are modelled using *concurrent game structures* (CGS), which are specified by a tuple

$$M = (N, \Phi, S, (A_i)_{i \in N}, s^0, \mathsf{tr}, \lambda),$$

where

- $N = \{1, \ldots, n\}$ is a finite, non-empty set of *agents*;

- Φ is a finite, non-empty set of Boolean variables;
- S is a finite, non-empty set of *states*;
- For each $i \in N$, A_i is a finite, non-empty set of *actions* which are available to agent i. For each state $s \in S$, let $A_i(s) \subseteq A_i$ denote the set of actions available to agent i in s, $\vec{A} \coloneqq \bigtimes_{i \in N} A_i$, and $\vec{A}(s) \coloneqq \bigtimes_{i \in N} A_i(s)$;
- $s^0 \in S$ is the *initial state*;
- $\mathsf{tr} : S \times \vec{A} \to S$ is the deterministic *transition function* of the game;
- $\lambda : S \to \{0,1\}^{\Phi}$ is the *labelling function*, which assigns to each state $s \in S$ a set of Boolean variables which are true in that state.

An *LTL game* is then simply a CGS augmented with LTL goals for each agent $i \in N$, i.e., a structure

$$\mathcal{G}_L = (M, \gamma_1, \ldots, \gamma_n).$$

Please see [151] for more details about ATL* and LTL games. Given an ERMG $\mathcal{G} = (\mathcal{A}, \gamma_1, \ldots, \gamma_n)$, we first convert it into a regular RMG $\mathcal{G}^* = (\mathcal{A}^*, \gamma_1, \ldots, \gamma_n)$ using the construction described earlier in Section 6.4 and from this, we construct an LTL game

$$\mathcal{G}_L = (N, \Phi^*, S, (A_i)_{i \in N}, s^0, \mathsf{tr}, \lambda),$$

where:

- Each state $s \in S$ represents a unique valuation for the variables in Φ^*, i.e., an assignment for the original variables Φ and the energy level of each agent Φ^e;
- The action set A_i for each agent $i \in N$ contains one action for every guarded command in the specification of m_i^* and only the actions corresponding to guarded commands that are enabled in a particular state s are available at that state;
- The transition function tr encodes how the variables in Φ^* are updated by the choice of actions (which correspond to enabled guarded commands) by each player;
- The labelling function λ labels each state with the set of variables that are true in the valuation that the state represents.

A more explicit procedure for this construction can be found in Fig. 12 of [16], which computes a Kripke structure from a regular SRML arena that can easily be converted into a CGS. Given the LTL game $\mathcal{G}_L$, we follow Theorem 3 of [151] by model checking the following ATL* formula:

$$\bigvee_{W \subseteq N} \left(\langle\langle N \rangle\rangle \left(\varphi \wedge \bigwedge_{i \in W} \gamma_i \wedge \bigwedge_{j \in N \setminus W} \neg \gamma_j \right) \wedge \bigwedge_{L \subseteq N \setminus W} [\![L]\!] \bigvee_{j \in L} \neg \gamma_j \right),$$

which encodes the property that there is a set of players $W \subseteq N$ (the "winners") in $\mathcal{G}_L$ such that:

1. There is a strategy profile $\vec{\sigma}$ that induces a run in $\mathcal{G}_L$ which satisfies φ and the goals of the winners W, but does not satisfy the goals of the "losers" $N \setminus W$;

2. For all subsets L of the losing players $N \setminus W$, and all partial strategy profiles $\vec{\sigma}'_L$ for the players in L, there is some response $\vec{\sigma}'_{N \setminus L}$ by the remaining players to prevent at least one player in L from achieving their goal, i.e., L does not have a beneficial deviation from $\vec{\sigma}$.

It is straightforward to see that the above formula is true in the initial state s^0 of $\mathcal{G}_L$, if and only if the answer to E-CORE is "yes." For the complexity of model checking this formula, note that the construction of $\mathcal{G}_L$ requires an exponential number of states in the size of Φ^*, which is itself polynomial in the size of $\mathcal{G}$. Moreover, we must make a worst-case exponential number of calls to the model checking procedure, one for each possible group of winners. Now, ATL* model checking can be done in 2EXPTIME with respect to the size of the formula and polynomial time with respect to the size of the game structure [38]. Hence, model checking this formula runs in doubly exponential time with respect to φ and $(\gamma_i)_{i \in N}$ and singly exponential time with respect to the number of variables $|\Phi|$ in the original game $\mathcal{G}$.

For the lower bound, we can reduce from the E-CORE decision problem for LTL games, which is known to be 2EXPTIME-complete [151]. Given an LTL game $\mathcal{G}_L = (M, \gamma_1, \ldots, \gamma_n)$, we can easily construct an ERMG $\mathcal{G} = (\mathcal{A}, \gamma_1, \ldots, \gamma_n)$ with no energy values whose runs are in a one-to-one correspondence with those of $\mathcal{G}_L$ [300, Chapter 7]. Then, it is clear that the answer to E-CORE for the LTL game $\mathcal{G}_L$ will be "yes" iff the answer to E-CORE for the corresponding ERMG $\mathcal{G}$ is "yes."

The result for the A-CORE problem for ERMGs follows straightforwardly by observing that the answer to A-CORE for a game $\mathcal{G}$ and a formula φ is "yes" if and only if the answer to E-CORE for the game $\mathcal{G}$ and the formula $\neg\varphi$ is "no", i.e., there are no Core-stable outcomes in $\mathcal{G}$ that do not satisfy φ. Since 2EXPTIME is a deterministic complexity class, we have that A-CORE for ERMGs is 2EXPTIME-complete.

To see that the result also holds for the case of regular RMGs, note that the reduction to establish the upper bound requires converting the ERMG to a regular RMG before encoding it as an LTL game. Hence, for regular RMGs, we can skip the first step, which runs in polynomial time, and then proceed to apply the same argument. For the lower bound, observe that the ERMG constructed does not require us to use any energy values. Therefore, the same reduction can be used for regular RMGs. □

6.5 Endogenous ERMGs

We now turn to the primary focus of our study, which is to understand how energy offers can be strategically utilised by agents to achieve better outcomes in a stable manner. To this end, we first present a framework for modelling negotiations by introducing a pre-play offer phase, in which the agents can offer resource transfers to other agents before the game begins. Crucially, we maintain that such offers should be stable with respect to deviations by (coalitions of) players in the same way that strategies in the ensuing ERMGs are.

Formally, an *offer* by agent i is a function

$$\omega_i : N \setminus \{i\} \to \{0, \ldots, E_i^0\}, \text{ such that:}$$

1. $\mathsf{totoff}_i \coloneqq \sum_{j \neq i} \omega_i(j) \leq E_i^0$; and
2. for each agent $j \in N \setminus \{i\}$, we have $\omega_i(j) + E_j^0 \leq e_j^{max}$.

The first condition states that the total amount of energy an agent offers to other agents does not exceed their initial endowment, and the second condition states that offers must be made within the capacity constraints of their recipients. An offer thus specifies how much energy agent i proposes to transfer to each other agent in the game in the pre-play negotiation phase. Denote by Ω_i the set of all

valid initial offers for an agent $i \in N$. Then, an *offer profile* $\vec{\omega} = (\omega_1, \ldots, \omega_n)$ is simply a tuple of offers for each agent $i \in N$, and we write $\Omega = \bigtimes_{i \in N} \Omega_i$ for the set of all offer profiles. We write $\vec{\omega}^{\,0}$ for the *empty offer profile* in which no players offer any energy to any other player.

Given this, an *Endogenous Energy Reactive Modules Game* (EERMG) proceeds in the following manner:

Stage 1: Each agent $i \in N$ chooses a valid offer ω_i, giving rise to an offer profile $\vec{\omega}$.

Stage 2: Each agent chooses a strategy σ_i in the ERMG $\mathcal{G}^{\vec{\omega}} = (\mathcal{A}', \gamma_1, \ldots, \gamma_n)$, where A' is exactly the same as $\mathcal{A}$, except that for each $i \in N$, we update i's initial energy as $E_i^{0'} = \min(E_i^0 - \mathsf{totoff}_i + \sum_{j \neq i} \omega_j(i), e_i^{max})$, for each agent i to reflect the resource transfer offers made in $\vec{\omega}$. The game $G^{\vec{\omega}}$ is then played according to the strategy profile $\vec{\sigma} = (\sigma_i)_{i \in N}$.

Given this two-stage game, the notion of a stable outcome can be ambiguous. This is because agents make two types of decisions, the former affecting the latter. Here, we will assume that when agents reason about offer profiles, they consider the stable outcomes induced by such offers. We use the classical Nash equilibrium and core stability notions for each stage separately. This means that we require both the energy offers to be stable against the appropriate kind of deviations *and* the strategy profile in the resulting stage 2 game to be stable to deviations. Given that we consider both unilateral and coalitional deviations, this approach allows us to combine the different solution concepts in different ways for each stage. Specifically, this gives rise to four possible combinations: Nash-Nash, Nash-Core, Core-Nash, and Core-Core. This decoupling between stability concepts for the first and second stages of the game allows one greater flexibility in finding a suitable model for the situation that they are trying to capture, because agents may have differing capabilities for communicating with one another and coordinating their behaviours in different stages of a game.

Stable Offers

To define a notion of stability over offer profiles, we require a way for the agents to rank such offers in terms of preferability. However, this is difficult to specify precisely, because different offer profiles can induce different sets of stable strategy profiles in stage 2 of the game. Since we do not make assumptions

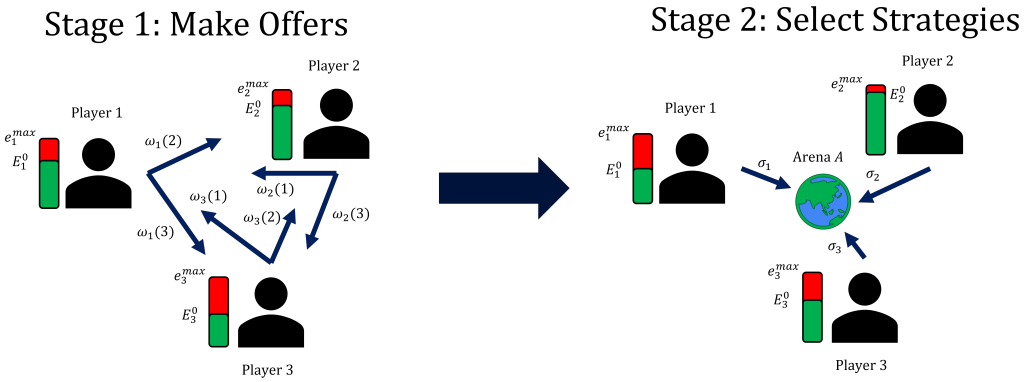

Figure 6.4: Illustration of the two-stages of an EERMG. In the first stage, players make offers of energy transfers to one another. In the second stage, players choose their strategies, and the game is played according to the chosen strategies.

about how the agents resolve the equilibrium selection problem, we propose a minimal preference relation $\succ_i^S$ for each agent $i \in N$ over offer profiles that we should expect such agents to have. This relation is defined based on whether (i) a stable stage 2 outcome exists, (ii) a stable stage 2 outcome exists which satisfies i's goal, and (iii) all (and at least one) stable stage 2 outcomes satisfy i's goal. More concretely, for an agent $i \in N$ and a stability concept $S \in \{\text{NASH}, \text{CORE}\}$, let $\emptyset_i^S := \{\vec{\omega} \in \Omega \mid S_{\gamma_i}(\mathcal{G}^{\vec{\omega}}) = \emptyset\}$, $\exists_i^S := \{\vec{\omega} \in \Omega \mid S(\mathcal{G}^{\vec{\omega}}) \neq S_{\gamma_i}(\mathcal{G}^{\vec{\omega}}) \neq \emptyset\}$, and $\forall_i^S := \{\vec{\omega} \in \Omega \mid S(\mathcal{G}^{\vec{\omega}}) = S_{\gamma_i}(\mathcal{G}^{\vec{\omega}}) \neq \emptyset\}$.

Given any offer profiles $\vec{\omega}_1 \in \emptyset_i^S, \vec{\omega}_2 \in \exists_i^S$, and $\vec{\omega}_3 \in \forall_i^S$, we define $\succ_i^S$ such that $\vec{\omega}_3 \succ_i^S \vec{\omega}_2 \succ_i^S \vec{\omega}_1$. Moreover, for any two offer profiles $\vec{\omega}, \vec{\omega}'$ in the same set ($\emptyset_i^S, \exists_i^S$, or $\forall_i^S$), we assume that i is indifferent between the two offer profiles, i.e., $\vec{\omega} \sim_i^S \vec{\omega}'$. This preference relation captures the intuition that an agent should (i) prefer an offer profile in which it is *possible* for their goal to be achieved in some stable outcome over one in which it is not possible to do so, and (ii) prefer an offer profile in which their goal is *guaranteed* to be achieved in any stable outcome over one in which it is merely possible.

With this, we can define Nash equilibrium in the usual way. An offer profile $\vec{\omega}$ is *Nash-S-stable* for the stage 1 negotiation phase if for all agents $i \in N$ and all alternative offers $\omega_i' \in \Omega_i$, it holds that $\vec{\omega} \succeq_i^S (\vec{\omega}_{-i}, \omega_i')$. For the case of core stability, the definition is not as immediate. The reason for this is that once a coalition $C \subseteq N$ deviates by making alternative offers, we assume that the remaining players $N \setminus C$ have the opportunity to *respond* and block the deviation from being profitable for all of the deviating players. In the context of offer profiles, however,

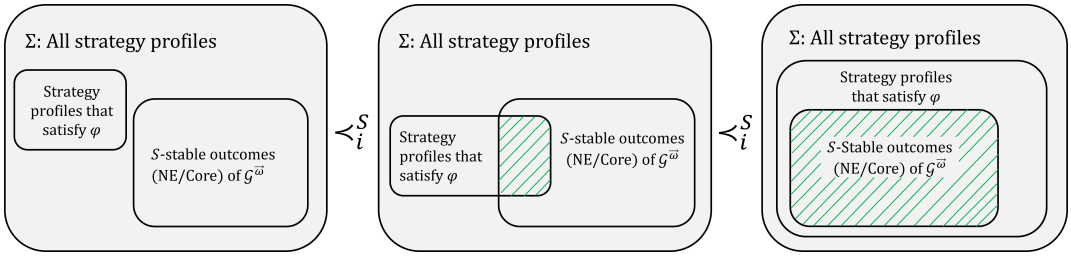

Figure 6.5: Visual illustration of the preference relation over games for a given player i in an ERMG that is induced by a solution concept S. A player prefers games in which their goal is achieved in *all* S-stable outcomes over those where it is achieved in some (but not all). Additionally, games in which it is possible for their goal to be achieved in an S-stable outcome are preferred to those in which this is impossible.

a deviation can change the set of offers the remaining players can respond with. In this study, we will assume that a response takes into account the new offers made by the deviating players. However, we will also assume that for both the deviating and responding players, the option of withdrawing all previously made offers and re-allocating the withdrawn offers is always available. Given this, we will say that an offer profile $\vec{\omega}$ is *Core-S-stable* for stage 1 if for all coalitions $C \subseteq N$ and alternative partial offer profiles $\vec{\omega}'_C = (\omega'_i)_{i \in C}$, there is some response $\vec{\omega}'_{-C} = (\omega'_i)_{i \in N \setminus C}$ such that for some $i \in C$, it holds that $\vec{\omega} \succeq_i^S (\vec{\omega}'_C, \vec{\omega}'_{-C})$.

Example 6.3. Consider the following modification of Example 6.1. Let $E^0_{B_1} = E^0_{B_2} = 0$ and for $i \in \{1, 2\}, \gamma_{B_i} = \mathbf{F}s_i \wedge \bigwedge_{j=0}^{2} \mathbf{G}(s_j \rightarrow \mathbf{X}\neg s_j)$, i.e., both B_1 and B_2 also do not want to stay in the same location twice in a row. We then introduce two additional robots B_3 and B_4 with $E^0_{B_3} = e^{max}_{B_3} = 2, E^0_{B_4} = e^{max}_{B_4} = 1$. Unlike the other robots, these additional robots are immobile and can only transfer their energy to others. Suppose that their goals are defined as $\gamma_{B_3} = \mathbf{X}s_1, \gamma_{B_4} = \mathbf{X}s_2 \vee \mathbf{X}t$. Consider the offer profile in which B_3 offers B_1 and B_2 1 unit each, and B_4 offers B_2 1 unit of energy. This profile is CORE-CORE-stable, as any possible offer deviation by B_3 does not result in γ_{B_3} being satisfied in all strategy profiles in the core of the second stage game (e.g., if B_3 withdraws the offer to B_2, B_4 can counter by also withdrawing its offer). On the other hand, in this scenario, there is no NASH-CORE-stable offer profile, since any valid offer profile will eventually enter a cycle of deviations.

Example 6.3 illustrates that stable offer profiles may not exist in EERMGs. This example suggests that some games are inherently unstable during the first stage. Therefore, the stability of the negotiation phase becomes a critical issue to address. In the following sections, we examine some decision problems related

to this and provide several algorithms for solving them. We demonstrate that solving such problems is no harder than standard rational verification.

Decision Problems

Now, we turn to the central question in this study, which is to determine whether a stable offer profile exists in a given EERMG. To this end, we introduce the following decision problems:

> S-OFFER-PREFERENCE:
> *Given*: Game $\mathcal{G}$, agent $i \in N$, offer profiles $\vec{\omega}_1, \vec{\omega}_2$, solution concept S for stage 2.
> *Question*: Is it true that $\vec{\omega}_1 \succeq_i^S \vec{\omega}_2$?

> S^1-S^2-OFFER-EXISTENCE:
> *Given*: Game $\mathcal{G}$, solution concept S^1 for stage 1, solution concept S^2 for stage 2.
> *Question*: Does there exist an S^1-S^2-stable offer profile $\vec{\omega}$ in $\mathcal{G}$?

It is worth noting that, in general, an EERMG is not guaranteed to have a NASH-S- or CORE-S-stable offer profile, since the preference relations $\succ_i^S$ implicitly group offer profiles into more than two categories, thus allowing "deviation cycles" to exist.

Turning to S-OFFER-PREFERENCE, for the upper-bound, we employ an algorithm that simply runs a stable profile check over the given offer profiles. For the lower bound, we reduce the problem of deciding whether an S-stable strategy profile exists in a regular RMG, hereafter called the S-NON-EMPTINESS problem.

Theorem 6.3. *For $S \in \{\text{NASH}, \text{CORE}\}$, S-OFFER-PREFERENCE is 2EXPTIME-complete.*

Proof. For membership, we begin by showing that Algorithm 1 with the given solution concept S decides our problem. In line 3 of the algorithm, we observe that if $A_2 =$ "No", then $\vec{\omega}_2 \in \emptyset_i^S$ and hence, it is impossible that $\vec{\omega}_2 \succ_i^S \vec{\omega}_1$. Hence, the answer to S-OFFER-PREFERENCE must be "Yes." If $A_2 =$ "Yes" and $A_1 =$ "No", then clearly $\vec{\omega}_2 \succ_i^S \vec{\omega}_1$ so we should return "No." If both A_1 and A_2 are "Yes", then we can run A-S with $\varphi = \gamma_i$ to settle the matter. If $A_3 =$ "Yes", then we know that $\vec{\omega}_1 \in \forall_i^S$ and similarly for A_4. If the answer is "No", then we know

instead that the offer profile is in $\exists_i^S$. The only case in which $\vec{\omega}_2 \succ_i^S \vec{\omega}_1$ is when $A_3 =$ "No" and $A_4 =$ "Yes." Since the worst-case runtime of both E-S and A-S is doubly exponential in the size of the players' goals $(\gamma_i)_{i \in N}$ and the input formula φ (c.f. Proposition 13 in [16] and the proof of Proposition 6.2) and we only call these algorithms at most two times each, Algorithm 1 runs in doubly exponential time with respect to the size of the players' goals and the input formula, and at most exponential time with respect to the size of the arena.

For hardness, we reduce from the problem of deciding whether an S-stable strategy profile exists in a regular RMG $\mathcal{G} = (\mathcal{A}, \gamma_1, \ldots, \gamma_n)$, which we know to be 2ExpTime-complete by Propositions 6.1 and 6.2. From the arena $\mathcal{A}$, we construct an SRML arena

$$\mathcal{A}_E = (N_E, \Phi_E, (m_i)_{i \in N_E}, (e_i^{max})_{i \in N_E}, (E_i^0)_{i \in N_E}),$$

where

- $N_E = \{1, \ldots, n+2\}$;
- $\Phi_E = \Phi \cup \{q_1, q_2\}$, where $\Phi_{n+1} = \{q_1\}$ and $\Phi_{n+2} = \{q_2\}$;
- For all $i \in N$, the module m_i remains unchanged, and the remaining modules are given below:

```
module m_{n+1} controls Φ_{n+1}
init
  [0] ⊤ ⇝ q_1 := ⊥
update
skip

module m_{n+2} controls Φ_{n+2}
init
  [0] ⊤ ⇝ q_2 := ⊥
  [−1] ⊤ ⇝ q_2 := ⊤
update
skip
```

- $e_i^{max} = 0$ for $i \in \{1, \ldots, n\}$ and $e_{n+1}^{max} = e_{n+2}^{max} = 1$;
- $E_i^0 = 0$ for $i \in \{1, \ldots, n\}$, $E_{n+1}^0 = 1$, and $E_{n+2}^0 = 0$.

Given this, we construct an ERMG $\mathcal{G}_E = (\mathcal{A}_E, \gamma_1, \ldots, \gamma_n, \gamma_{n+1}, \gamma_{n+2})$, where all of the original goals $(\gamma_i)_{i \in N}$ remain unchanged, and $\gamma_{n+1} = \gamma_{n+2} = q_2$.

Now, let $\vec{\omega}^*$ be the offer profile in which player $n+1$ offers one unit of energy to player $n+2$ and all other offer values are 0. The proof concludes by showing that the answer to E-S in the original RMG $\mathcal{G}$ is "yes" if and only if the answer to S-OFFER-PREFERENCE with $i = n+1$, $\vec{\omega}_1 = \vec{\omega}_0$, and $\vec{\omega}_2 = \vec{\omega}^*$ is "no."

To see this, note first that the offer profile $\vec{\omega}^*$ enables player $n+2$ to select the initial command $[-1]\ \top \rightsquigarrow q_2 \coloneqq \top$, which is not otherwise enabled under $\vec{\omega}_0$. Since player $n+2$ prefers to choose this initial command rather than the one which sets $q_2 \coloneqq \bot$ on the first round, it holds that $\vec{\omega}^* \succ^S_{n+1} \vec{\omega}_0$ if and only if there is an S-stable strategy profile in the original game $\mathcal{G}$.

To see the forward implication of this statement, first observe that since player $n+2$ does not have enough energy to set $q_2 \coloneqq \top$ under $\vec{\omega}_0$, there is no S-stable strategy profile in $\mathcal{G}_E^{\vec{\omega}_0}$ which satisfies γ_{n+1}. Secondly, the assumption $\vec{\omega}^* \succ^S_{n+1} \vec{\omega}_0$ implies that $\vec{\omega}^* \in \exists^S_{n+1} \cup \forall^S_{n+1}$, which is only possible if there is some S-stable strategy profile $\vec{\sigma}_E = (\sigma_1, \ldots, \sigma_{n+2})$ in $\mathcal{G}_E$. Now, because the actions of the players N and $\{n+1, n+2\}$ are completely independent of each other, we can infer from $\vec{\sigma}_E$ the existence of an S-stable strategy profile $\vec{\sigma} = (\sigma_1, \ldots, \sigma_n)$ in $\mathcal{G}$.

For the backward implication, suppose that $\vec{\sigma} = (\sigma_1, \ldots, \sigma_n)$ is an S-stable strategy profile in $\mathcal{G}$. Then, consider the strategy profile $\vec{\sigma}_E = (\sigma_1, \ldots, \sigma_n, \sigma_{n+1}, \sigma_{n+2})$ in $\mathcal{G}_E$, where σ_{n+1} is trivial because m_{n+1} is deterministic, and σ_{n+2} sets $q_2 \coloneqq \top$ on the first round, after which their only choice is `skip`. Clearly, it holds that $\vec{\sigma}_E \in S(\mathcal{G}_E^{\vec{\omega}^*})$, since no player (or coalition of players) in $\{1, \ldots, n\}$ has a beneficial deviation by assumption and $\rho(\vec{\sigma}_E, \mathcal{G}_E^{\vec{\omega}^*}) \models q_2 = \gamma_{n+1} = \gamma_{n+2}$. Moreover, as we have argued above, we know that $S_{\gamma_{n+1}}(\mathcal{G}_E^{\vec{\omega}_0}) = \emptyset$ and hence, $\vec{\omega}^* \succ^S_{n+1} \vec{\omega}_0$. □

Using this result, we can settle the complexity of the remaining problems. We show that they are also 2EXPTIME-complete, and hence, reasoning about the existence of stable offers with desirable properties is no harder than the underlying rational verification problems. The upper bound for the NASH-S-OFFER EXISTENCE problem is established by checking, for every agent, whether there is no beneficial deviation from the given strategy profile, which can be done by employing a suitable variant of LTL synthesis. The lower bound again uses a reduction from the S-NON-EMPTINESS problem.

Theorem 6.4. *For $S \in \{\text{NASH}, \text{CORE}\}$, NASH-$S$-OFFER-EXISTENCE is 2EXPTIME-complete.*

Proof. For membership, we use Algorithm 2 with S for the stage 2 solution concept. This procedure returns "Yes" if and only if there is a Nash-S-stable offer profile. Since each agent has a finite amount of initial energy, the number of offers each agent can make is exponential in their initial energy level E_i^0 and the number of players n. Hence, the number of valid offer profiles is exponential in $\sum_{i \in N} E_i^0$ and n. Line 2 requires iterating through an exponential number of potential offer deviations, and line 3 requires making a call to Algorithm 1, which runs in worst-case doubly exponential time with respect to the size of the players' goals and singly exponential time with respect to the number of players and the initial energy values. Hence, the overall procedure is in 2EXPTIME.

For hardness, we reduce from the problem of deciding whether an S-stable strategy profile exists in a regular RMG $\mathcal{G}$, which we know to be 2EXPTIME-complete by Propositions 6.1 and 6.2. Given a regular RMG $\mathcal{G} = (\mathcal{A}, \gamma_1, \ldots, \gamma_n)$ where $\mathcal{A} = (N, \Phi, (m_i)_{i \in N})$, we first construct an arena

$$\mathcal{A}_E = (N_E, \Phi_E, (m_i)_{i \in N_E}, (e_i^{max})_{i \in N_E}, (E_i^0)_{i \in N_E}),$$

where:

- $N_E = \{1, \ldots, n+2\}$;
- $\Phi_E = \Phi \cup \{q_1, q_2, q_3\}$, where $\Phi_{n+1} = \{q_1, q_2\}$ and $\Phi_{n+2} = \{q_3\}$;
- For all $i \in N$, the module m_i remains unchanged, and the remaining modules are given below:

```
module m_{n+1} controls Φ_{n+1}
init
  [0] ⊤ ⇝ q_1 := ⊥, q_2 := ⊥
  [−1] ⊤ ⇝ q_1 := ⊤, q_2 := ⊥
  [−2] ⊤ ⇝ q_1 := ⊥, q_2 := ⊤
update
skip

module m_{n+2} controls Φ_{n+2}
init
  [0] ⊤ ⇝ q_3 := ⊥
  [−2] ⊤ ⇝ q_3 := ⊤
update
skip
```

- $e_i^{max} = 0$ for $i \in \{1, \ldots, n\}$ and $e_{n+1}^{max} = e_{n+2}^{max} = 2$;

- $E_i^0 = 0$ for $i \in \{1, \ldots, n\}$, $E_{n+1}^0 = 1$, and $E_{n+2}^0 = 1$.

From this, we construct an ERMG $\mathcal{G}_E = (\mathcal{A}_E, \gamma_1, \ldots, \gamma_n, \gamma_{n+1}, \gamma_{n+2})$, where $\gamma_{n+1} = q_2 \vee q_3$ and $\gamma_{n+2} = q_1 \wedge \neg q_2$. We will show that there exists an S-stable strategy profile in the regular RMG $\mathcal{G}$ if and only if there is no Nash-S-stable offer profile in $\mathcal{G}_E$.

For the forward direction, suppose that there exists an S-stable strategy profile $\vec{\sigma} \in S(\mathcal{G})$ in the original RMG. We will use this to show the existence of a beneficial deviation for some player from all valid offer profiles, of which there are four:

1. $\vec{\omega}^{\,0}$: the empty offer profile;
2. $\vec{\omega}^{\,1}$: the offer profile in which player $n+1$ offers 1 unit of energy to player $n+2$ and player $n+2$ makes no offers;
3. $\vec{\omega}^{\,2}$: the offer profile in which player $n+2$ offers 1 unit of energy to player $n+1$ and player $n+1$ makes no offers;
4. $\vec{\omega}^{\,3}$: the offer profile in which players $n+1$ and $n+2$ offer 1 unit of energy to each other.

These are the only valid offer profiles because players $\{1, \ldots, n\}$ have no energy capacity, so no offers can be made by or to them. We will show that there is a cycle of deviations of the form $\vec{\omega}^{\,0} \to_{n+1} \vec{\omega}^{\,1} \to_{n+2} \vec{\omega}^{\,2} \to_{n+1} \vec{\omega}^{\,3} \to_{n+2} \vec{\omega}^{\,0}$, where $\vec{\omega} \to_i \vec{\omega}\,'$ denotes the fact that player i can change their offer ω_i under $\vec{\omega}$ to an offer ω_i' such that $\vec{\omega}\,' = (\vec{\omega}_{-i}, \omega_i') \succ_i^S \vec{\omega}$. Firstly, consider $\vec{\omega}^{\,0}$. Under this offer profile, player $n+1$ does not have enough energy to set $q_2 \coloneqq \top$ and player $n+2$ does not have enough energy to set $q_3 \coloneqq \top$. Hence, it is impossible for player $(n+1)$'s goal to be achieved, so $\vec{\omega}^{\,0} \in \emptyset_{n+1}^S$. Now, consider $\vec{\omega}^{\,1}$ and the strategy profile $\vec{\sigma}_{\,1} = (\sigma_1, \ldots, \sigma_n, \sigma_{n+1}, \sigma_{n+2})$ in which players 1 to n play their corresponding strategies in $\vec{\sigma}$, player $n+1$ sets $q_1 \coloneqq \bot, q_2 \coloneqq \bot$ in the first round, and player $n+2$ sets $q_3 \coloneqq \top$ in the first round. It is easy to see that $\rho(\vec{\sigma}_{\,1}, \mathcal{G}_E^{\vec{\omega}^{\,1}}) \models \gamma_{n+1}$. Moreover, it holds that $\vec{\sigma}_{\,1} \in S_{\gamma_{n+1}}(\mathcal{G}_E^{\vec{\omega}^{\,1}})$, because: (1) players 1 to n have no beneficial deviations by assumption, (2) player $n+1$ has no other options, and (3) player $n+2$ does not get their goal satisfied regardless of what they do. Therefore, we have that $\vec{\omega}^{\,1} \in \exists_{n+1}^S \cup \forall_{n+1}^S$, so $\vec{\omega}^{\,0} \to_{n+1} \vec{\omega}^{\,1}$.

By a similar line of argument, it is straightforward to check that $\vec{\omega}^{\,1} \in \emptyset_{n+2}^S$, since there is no way for player $n+2$ to get their goal satisfied under this offer

profile. However, under $\vec{\omega}^{\,2}$, player $n+1$ has enough energy to choose the initial guarded command that sets $q_1 \coloneqq \top$ and $q_2 \coloneqq \bot$, hence satisfying γ_{n+2}. Moreover, constructing a strategy profile $\vec{\sigma}_{\,2}$ using this selection along with $\vec{\sigma}$ in the same manner as above gives us an S-stable strategy profile which satisfies γ_{n+2}. This is because under $\vec{\omega}^{\,2}$, player $n+1$ has no chance of getting their goal satisfied no matter what they do. Thus, we have $\vec{\omega}^{\,2} \in \exists^S_{n+2} \cup \forall^S_{n+2}$ and hence, $\vec{\omega}^{\,1} \to_{n+2} \vec{\omega}^{\,2}$.

Proceeding in the same manner, we can see that $\vec{\omega}^{\,2} \in \emptyset^S_{n+1}$, whereas under $\vec{\omega}^{\,3}$, player $n+1$ now has the option of setting $q_1 \coloneqq \bot$ and $q_2 \coloneqq \top$ on the first round. This satisfies γ_{n+1} and we can construct an S-stable strategy profile $\vec{\sigma}_{\,3} \in S_{\gamma_{n+1}}(\mathcal{G}_E^{\vec{\omega}^{\,3}})$ as above, using the fact that now player $n+2$ has only one available option. Therefore, we have $\vec{\omega}^{\,3} \in \exists^S_{n+1} \cup \forall^S_{n+1}$ and hence, $\vec{\omega}^{\,2} \to_{n+1} \vec{\omega}^{\,3}$.

Finally, observe that $\vec{\omega}^{\,3} \in \emptyset^S_{n+2}$. This is because assuming that player $n+2$ chooses their first initial guarded command on the first round, player $n+1$ would never choose their second initial guarded command over their third one if they had enough energy, because the latter leads to their goal γ_{n+1} being satisfied whereas the former does not. From $\vec{\omega}^{\,3}$, player $n+2$ can withdraw their offer to take us back to $\vec{\omega}^{\,0}$, under which the option for player $n+1$ to select their third initial guarded command is taken away. In this case, player $n+1$ has no chance of achieving their goal, so we can construct a strategy profile $\vec{\sigma}_{\,4}$ in the same manner as above, where player $n+1$ selects the initial guarded command that sets $q_1 \coloneqq \top$ and $q_2 \coloneqq \bot$ and player $n+2$ only has one available option. Clearly, we have $\vec{\sigma}_{\,4} \in S_{\gamma_{n+2}}(\mathcal{G}_E^{\vec{\omega}^{\,0}})$, so $\vec{\omega}^{\,0} \in \exists^S_{n+2} \cup \forall^S_{n+2}$. Thus, we have $\vec{\omega}^{\,3} \to_{n+2} \vec{\omega}^{\,0}$, so none of the possible offer profiles are Nash-S-stable.

For the backward direction of the claim, we will assume that there is no S-stable strategy profile in $\mathcal{G}$ and then show that there must exist a Nash-S-stable offer profile. Given this assumption, it is clear that for any strategy profile $\vec{\sigma}$ in $\mathcal{G}_E$, some player $i \in \{1, \ldots, n\}$ (or coalition of players $C \subseteq \{1, \ldots, n\}$) will have a beneficial deviation no matter what players $n+1$ and $n+2$ do. Thus, under all valid offer profiles, there is no S-stable strategy profile in $\mathcal{G}_E$, so $\Omega = \emptyset^S_i$ for all $i \in N_E$. Hence, not only does there exist a Nash-S-stable offer profile, but in fact all valid offer profiles are Nash-S-stable. □

Finally, we turn to the cooperative setting and study the existence of core-stable offer profiles. Here, we establish the upper bound with a reasoning similar

to the one for Nash-S-Offer-Existence, and reduce from S-NON-EMPTINESS to obtain the lower bound.[3]

Theorem 6.5. *For $S \in \{\text{NASH}, \text{CORE}\}$,* CORE-$S$-OFFER-EXISTENCE *is* 2EXPTIME*-complete.*

Proof. For membership, observe that Algorithm 3 decides our problem by directly applying the definition of core-stability for offer profiles. As argued in the proof of Theorem 6.4, the number of valid offer profiles is at most exponential in $\sum_{i \in N} E_i^0$ and n. Line 2 requires iterating through an exponential number of subsets $C \subseteq N$ and for each of these subsets, a number of partial offer profiles $\vec{\omega}'_C$ that is exponential in $\sum_{i \in C} E_i^0$ and $|C|$. Line 3 again requires iterating through an exponential number of responses $\vec{\omega}'_{-C}$. Line 4 requires making a call to Algorithm 1, which we know runs in doubly exponential time with respect to the size of the goals $(\gamma_i)_{i \in N}$ and singly exponential time with respect to the number of players and the initial energy values.

For hardness, we reduce from the E-S decision problem for regular RMGs, which by Propositions 6.1 and 6.2 is 2EXPTIME-complete. Given an RMG $\mathcal{G} = (\mathcal{A}, \gamma_1, \ldots, \gamma_n)$, we first construct an augmented arena

$$\mathcal{A}_E = (N_E, \Phi_E, (m_i)_{i \in N_E}, (e_i^{max})_{i \in N_E}, (E_i^0)_{i \in N_E}),$$

where

- $N_E = \{1, \ldots, n+3\}$;
- $\Phi_E = \Phi \cup \{q_1, q_2, q_3, q_4, q_5, q_6\}$, where $\Phi_{n+1} = \{q_1, q_4\}, \Phi_{n+2} = \{q_2, q_5\}$, and $\Phi_{n+3} = \{q_3, q_6\}$;
- For all $i \in N$, the module m_i remains unchanged, and the remaining modules are given below:

$$\begin{array}{l}
\texttt{module}\ m_{n+1}\ \texttt{controls}\ \Phi_{n+1} \\
\texttt{init} \\
\quad [0]\ \top \rightsquigarrow q_1 := \bot, q_4 := \bot \\
\quad [-2]\ \top \rightsquigarrow q_1 := \top, q_4 := \bot \\
\quad [-2]\ \top \rightsquigarrow q_1 := \top, q_4 := \top \\
\texttt{update} \\
\texttt{skip}
\end{array}$$

[3]Note that although we reduce from the same decision problem in these three cases, the constructions required in each case are very distinct, due to the differences in the considered solution concepts as well as the fact that the S^1-S^2-OFFER-EXISTENCE problems require one to consider the set of all possible offer profiles.

```
module m_{n+2} controls Φ_{n+2}
init
  [0] ⊤ ⇝ q_2 := ⊥, q_5 := ⊥
  [−2] ⊤ ⇝ q_2 := ⊤, q_5 := ⊥
  [−2] ⊤ ⇝ q_2 := ⊤, q_5 := ⊤
update
skip

module m_{n+3} controls Φ_{n+3}
init
  [0] ⊤ ⇝ q_3 := ⊥, q_6 := ⊥
  [−2] ⊤ ⇝ q_3 := ⊤, q_6 := ⊥
  [−2] ⊤ ⇝ q_3 := ⊤, q_6 := ⊤
update
skip
```

- $e_i^{max} = 0$ for $i \in \{1, \ldots, n\}$ and $e_{n+1}^{max} = e_{n+2}^{max} = e_{n+3}^{max} = 3$;
- $E_i^0 = 0$ for $i \in \{1, \ldots, n\}$ and $E_{n+1}^0 = E_{n+2}^0 = E_{n+3}^0 = 1$.

From this arena, we can construct an ERMG

$$\mathcal{G}_E = (\mathcal{A}_E, \gamma_1, \ldots, \gamma_n, \gamma_{n+1}, \gamma_{n+2}, \gamma_{n+3}),$$

where $(\gamma_i)_{i \in N}$ remain unchanged, $\gamma_{n+1} = q_1 \vee q_6, \gamma_{n+2} = q_2 \vee q_4$, and $\gamma_{n+3} = q_3 \vee q_5$. We will show that there exists an S-stable strategy profile in $\mathcal{G}$ if and only if there is no Core-S-stable offer profile in $\mathcal{G}_E$.

For the forward direction, assume that $\vec{\sigma} = (\sigma_1, \ldots, \sigma_n)$ is an S-stable strategy profile in $\mathcal{G}$. We will show that for every valid offer profile $\vec{\omega} \in \Omega$, there exists a coalition $C \subseteq N$ who has a beneficial deviation from $\vec{\omega}$. To show this, we can consider the possible energy distributions that result from any valid offer profile, of which there are ten. Note that because the players in $\{1, \ldots, n\}$ have no energy capacity, we can characterise such distributions by 3-tuples, which represent the initial energy that players $n+1, n+2$, and $n+3$ will begin with. Thus, the possible resulting distributions can be written as $(3,0,0), (2,1,0), (2,0,1), (1,2,0), (1,1,1), (1,0,2), (0,3,0)$
$(0,2,1), (0,1,2), (0,0,3)$. We can sort these into the following groups:

$$O_1 = \{(3,0,0), (2,1,0), (2,0,1)\}$$
$$O_2 = \{(1,2,0), (0,3,0), (0,2,1)\}$$
$$O_3 = \{(1,0,2), (0,1,2), (0,0,3)\}$$
$$O_4 = \{(1,1,1)\}.$$

Additionally, we let $\hat{\Omega}_i$ denote the set of offer profiles in $\mathcal{G}_E$ whose resulting distribution of initial energy values is contained in O_i. We will proceed by systematically showing that for every outcome in each of these groups, there is a coalition who has a beneficial deviation.

Beginning with O_1, consider any $\vec{\omega}^{\,1} \in \hat{\Omega}_1$. Under this offer profile, player $n+1$ can select any of their initial guarded commands and will choose either the second or third one, because this guarantees the satisfaction of their goal $\gamma_{n+1} = q_1 \vee q_6$. Additionally, the players $n+2$ and $n+3$ have only one option, which is to initialise both of the variables under their control to $\bot$. Moreover, since $\vec{\sigma} = (\sigma_1, \ldots, \sigma_n)$ is S-stable in $\mathcal{G}$, we can infer the existence of an S-stable strategy profile $\vec{\sigma}_{\,E} = (\sigma_1, \ldots, \sigma_n, \sigma_{n+1}, \sigma_{n+2}, \sigma_{n+3})$ in $\mathcal{G}_E^{\vec{\omega}^{\,1}}$ where player $n+1$ chooses either their second or third initial guarded command and the other two additional players choose their only options. Because the actions of the players in $\{1, \ldots, n\}$ and $\{n+1, n+2, n+3\}$ are completely independent of each other, it is clear that $\vec{\sigma}_{\,E}$ will be S-stable in $\mathcal{G}_E^{\vec{\omega}^{\,1}}$. Moreover, any S-stable outcome in $\mathcal{G}_E^{\vec{\omega}^{\,1}}$ will satisfy γ_{n+1} because the only ones that do not are those in which player $n+1$ chooses their first initial guarded command, but in those cases, player $n+1$ has a beneficial deviation which the remaining players cannot block. In other words, it holds that $\vec{\omega}^{\,1} \in \forall^S_{n+1}$. Secondly, we can see that depending on player $(n+1)$'s choice to choose either their second or third initial command, player $(n+2)$'s goal $\gamma_{n+2} = q_2 \vee q_4$ may or may not be satisfied. Hence, we have that $\vec{\omega}^{\,1} \in \exists^S_{n+2}$. Finally, it is clear that no matter which S-stable strategy profile is chosen, player $(n+3)$'s goal will not be satisfied and hence, we have $\vec{\omega}^{\,1} \in \emptyset^S_{n+3}$. Summing up, we have shown that $\hat{\Omega}_1 \subseteq \forall^S_{n+1} \cap \exists^S_{n+2} \cap \emptyset^S_{n+3}$.

Now, we will show that the coalition $\{n+2, n+3\}$ has a beneficial deviation from any $\vec{\omega}^{\,1} \in \hat{\Omega}_1$. Consider the possible outcomes that could arise if both players in $\{n+2, n+3\}$ decided instead to offer their initial energies to player $n+2$. The possible responses by player $n+1$ would result in one of the outcomes in O_2. Thus, we can show that no matter which outcome in O_2 player $n+1$ picks as a response, players $n+2$ and $n+3$ will be better off. This can be seen by following the same line of argument as above but applied to an offer profile $\vec{\omega}^{\,2} \in \hat{\Omega}_2$, from which we can conclude that $\hat{\Omega}_2 \subseteq \emptyset^S_{n+1} \cap \forall^S_{n+2} \cap \exists^S_{n+3}$. Thus, both players $n+2$ and $n+3$ are guaranteed to benefit from this deviation, no matter the response of player $n+1$.

We can repeat this argument again, showing that the coalition $\{n+1, n+3\}$ has a beneficial deviation from any $\vec{\omega}^{\,2} \in \hat{\Omega}_2$ by both offering their initial energies to player $n+3$. The possible responses by player $n+2$ would lead to one of the

outcomes in O_3. Moreover, it is easy to verify using the same argument as above that $\hat{\Omega}_3 \subseteq \exists^S_{n+1} \cap \emptyset^S_{n+2} \cap \forall^S_{n+3}$. Hence, both players $n+1$ and $n+3$ are guaranteed to benefit by deviating in this manner, no matter the response of player $n+2$.

Next, we can close the loop by showing that the coalition $\{n+1, n+2\}$ has a beneficial deviation from any $\vec{\omega}^{\,3} \in \hat{\Omega}_3$ by both offering their initial energies to player $n+1$. The possible responses by player $n+3$ would result in one of the outcomes in O_1 and as we have already shown, it is the case that $\hat{\Omega}_1 \subseteq \forall^S_{n+1} \cap \exists^S_{n+2} \cap \emptyset^S_{n+3}$. Hence, both players $n+1$ and $n+2$ are guaranteed to benefit from this deviation.

Finally, we consider any offer profile $\vec{\omega}^{\,4} \in \hat{\Omega}_4$. Clearly, no player in $\{n+1, n+2, n+3\}$ is able to achieve their goal under this outcome, so $\hat{\Omega}_4 \subseteq \bigcap_{i \in \{n+1,n+2,n+3\}} \emptyset^S_i$. It is easy to see that any two of the three players will have a beneficial deviation which results in one of the outcomes in $O_1 \cup O_2 \cup O_3$ and hence, we conclude that no stable offer profile exists.

For the reverse direction, assume that there is no S-stable strategy profile in $\mathcal{G}$. Then, we see that $\Omega = \emptyset^S_i$ for all $i \in N_E$. Hence, all offer profiles are Core-S-stable. □

We remark that although the complexity of the decision problems related to EERMGs may be the same as those of the cooperative and non-cooperative rational verification problems, this largely inherits from the 2ExpTime upper-bound due to the fact that players' goals are expressed in LTL. It is also possible to consider fragments of LTL such as safety/reachability or GR(1) formulae for the players' goals, under which the complexity of rational verification is significantly reduced [150]. In some of these special cases, it is likely to be the case that the problem of deciding whether a stable offer profile exists is more complex than the underlying rational verification decision problems. We do not consider these here, but note that the reductions in Theorems 6.4 and 6.5 can easily be adapted to these special cases, since they do not rely on the full expressivity of LTL.

Algorithms 1-3 are utilised to establish upper bounds on the complexity of the decision problems introduced in Section 6.5. In particular, Algorithm 1 decides whether an offer profile $\vec{\omega}_1$ is weakly preferred to offer profile $\vec{\omega}_2$ by an agent i under solution concept S. Algorithm 2 decides whether there is a Nash-S-stable offer profile in a given EERMG under solution concept S. Finally, Algorithm 3 decides whether there is a Core-S-stable offer profile in a given EERMG under solution concept S.

Algorithm 1 S-OFFER-PREFERENCE

Input: Game $\mathcal{G}$, agent $i \in N$, offer profiles $\vec{\omega}_1, \vec{\omega}_2$, solution concept S

```
A_1 ← E-S(G^{ω⃗_1}, γ_i)
A_2 ← E-S(G^{ω⃗_2}, γ_i)
if A_2 = "No" then
    return "Yes"
else if A_1 = "No" then
    return "No"
else
    A_3 ← A-S(G^{ω⃗_1}, γ_i)
    A_4 ← A-S(G^{ω⃗_2}, γ_i)
    if A_3 = "No" and A_4 = "Yes" then
        return "No"
    else
        return "Yes"
    end if
end if
```

Algorithm 2 NASH-S-OFFER

Input: Game $\mathcal{G}$, stage 2 solution concept S

```
Guess offer profile ω⃗
for all i ∈ N and ω'_i ∈ Ω_i do
    if (ω⃗_{-i}, ω'_i) ≻_i^S ω⃗ then
        return "No"
    end if
end for
return "Yes"
```

6.6 Discussion

In this chapter, we have introduced ERMGs, a model of concurrent games with resource-constrained players using the reactive modules framework, and settled the complexity of both the key cooperative and non-cooperative rational verification questions for this model, showing that they are no harder than for regular RMGs. This therefore expands the range of practical applications which can be succinctly modelled using the reactive modules framework to analyse situations related to bounded resources, with almost no additional complexity cost. We then introduced Endogenous ERMGs, which include a pre-play phase that allows players to make energy offers to one another, along with a notion of preferences over offer profiles. This preference relation enables us to study

Algorithm 3 CORE-S-OFFER

Input: Game $\mathcal{G}$, stage 2 solution concept S

Guess offer profile $\vec{\omega}$
for all $C \subseteq N$ **and** $\vec{\omega}'_C \in \Omega_C$ **do**
  Guess response $\vec{\omega}'_{-C} \in \Omega_{-C}$
  if $(\vec{\omega}'_C, \vec{\omega}'_{-C}) \succ_j^S \vec{\omega}$ for all $j \in C$ **then**
    return "No"
  end if
end for
return "Yes"

the stability of offers in both cooperative and non-cooperative settings. We then settled the problem of deciding whether a stable offer profile exists in a given game under all combinations of stability concepts, which remains 2EXPTIME-complete. This initial study of the Endogenous ERMG model thus sheds light on the strategic considerations an agent faces when their actions in one stage of a game may affect the possible rational outcomes in the second stage of the game.

One natural extension of this work is to study how a top-down incentive designer [207, 272] could modify the energy levels of agents to shape the set of resulting stable outcomes. Relatedly, one could take an approach akin to *parameterised* resource-bounded ATL [299, 305], which can express formulae about the minimal amount of energy a coalition needs to achieve a particular goal, by asking whether energy subsidies or taxes can be introduced to enable specific outcomes. Finally, a direct next step would be to study whether some or all stable offer profiles induce games with (un)desirable properties. This becomes especially pertinent when considering the presence of malicious agents who might collude by exchanging energies to bring about an undesirable outcome. These results on the existence of stable offer profiles lay the groundwork for these future developments.

Part II

Incentives for Boundedly Rational Agents

7

A Model of Bounded Rationality

> *It is not enough, however, for our ideas to be rational—coherent and logical; they must be* applicable. *In other words, to be of value they must tell us about the world we experience and live in.*
>
> C. Robert Mesle, *Process-Relational Philosophy* (p. 15) [306]

In the second part of this thesis, we turn our attention to the ways in which real-world agents respond to incentives. In particular, we focus on the concept of bounded rationality, which seeks to describe the ways in which agents deviate from normative standards of rationality due to various limitations that are placed on their abilities to achieve their goals. At the end of Part I, we concluded by exploring a model in which agents were bounded by a limited resource known as an energy level that constrained the actions that they were able to take at any point in time. However, the agents were not intrinsically motivated to pursue or acquire energy. In contrast, this chapter directly models the trade-offs between goal acquisition and the expenditure of a kind of 'cognitive resource' measured in information-theoretic terms. This approach falls within the scope of an increasingly prominent family of models of decision making known as *resource-rational* models [307], which have been found to be a powerful tool for modelling the decision making of humans and other biological agents. While the focus of study so far has been on multi-agent systems and reasoning about

how incentives interact with equilibria, the following three chapters turn their attention to the single-agent setting. In taking this approach, we highlight less conventional ways that the decision making of a single boundedly rational agent can be influenced by their own limitations and the features of the environment. To conclude this part, we return to the multi-agent setting, expanding the model developed in this chapter to include interactions between agents.

Accurately predicting and modelling the decisions of real-world agents, from humans to AI systems, remains a fundamental challenge across cognitive science, neuroscience, and artificial intelligence, including the alignment of AI systems to human preferences. While machine learning methods and capabilities have advanced significantly, progress in modelling real-world decision making under cognitive and informational constraints has been comparatively slower. To address this gap, we introduce the *Path Divergence Objective* (PDO), an information-theoretic model of boundedly rational decision making in stochastic, partially observable environments. The PDO is derived from physical principles, modelling the inherent costs of information processing in model-based planning and belief updating for embodied agents. This framework enables us to model key features observed in real-world agent behaviour, such as curiosity-driven exploration, novelty-seeking, and the intention-behaviour gap. By adjusting a single parameter, the PDO can describe a continuous spectrum of decision making strategies, ranging from highly irrational to perfectly rational. This flexibility makes the PDO applicable to a wide range of scenarios, including modelling biological organisms, simulating interactions between agents with varying degrees of bounded rationality, addressing AI alignment challenges, and designing AI systems that interact more effectively with humans.

7.1 Related Work

The concept of bounded rationality, originally developed to model human decision making [56, 308], has broader applications in modelling any teleological physical system. This includes not only humans but also AI systems and other biological entities. The universality of this approach stems from the fact that all physical systems operate within thermodynamic constraints, converting available energy into useful work [65, 309–312].

Our proposed framework builds upon and generalises an information-theoretic model of bounded rationality [313, 314], which focuses on the informational cost

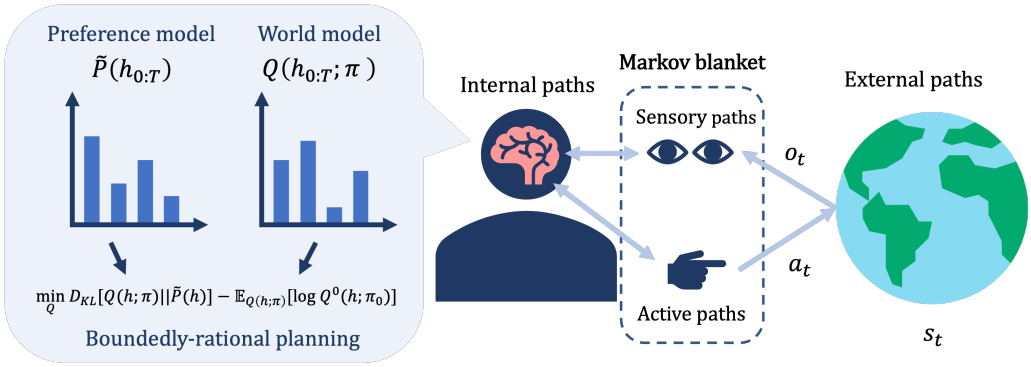

Figure 7.1: Illustration of the framework. The agent possesses an internal model, which is decomposed into a world model and a preference model, minimising the PDO. The agent's interface to the external world is known as its *Markov blanket* [65, 315, 316], which is comprised of active and sensory paths, which are state trajectories that mediate the interactions between internal and external paths.

of finding a good policy. This model offers a principled approach to modelling decision making in complex, uncertain environments, naturally capturing trade-offs between exploitation and exploration. We anticipate its applicability to a wide range of agents with varying internal structures and intelligence levels, from individual neurons to advanced AI systems, providing a unified framework for understanding decision making across different scales of complexity. Figure 7.1 gives an overview of the framework.

Bounded Rationality The PDO formalises and extends Simon's original concept of bounded rationality [308] in three key ways: (1) it introduces partial observability to information-theoretic models of bounded rationality [313, 314, 317, 318] to enable a derivation of the origins of curious or information-seeking behaviour; (2) it completes the bridge to active inference models [70, 319], demonstrating features like information-seeking behaviour which are a common feature of such models; and (3) it complements models such as rational inattention [320–322] by modelling the costs of deliberation and planning in dynamic, sequential decision making problems, modelling how agents balance information processing costs with preference satisfaction over time.

Our model can also be viewed through the lens of resource-rationality [307, 323, 324], which is centered on the idea of viewing boundedly rational agents as making "rational use of limited cognitive resources" for decision making, and has been found to predict various facets of human decision making better than existing models [325, 326]. Closely related is the notion of computational

rationality [327, 328], which is concerned with the computational costs associated with deliberation and decision making.

Active Inference and Divergence Objectives The PDO shares conceptual foundations with active inference, a framework for modelling perception and action based on free energy minimisation [70, 319, 329–331]. Indeed, one view of this work is as a derivation of a broader class of active inference or divergence objectives from the starting point of bounded rationality [332], which includes several existing objectives such as the Expected Free Energy [333, 334], the Free Energy of the Expected Future [335], and Action Perception as Divergence minimisation [336].

Reinforcement Learning and Control Theory While traditional Reinforcement Learning (RL) focuses on maximising expected rewards [62], recent work has explored information-theoretic objectives in RL and control theory [337–341].

KL-divergence or mutual information regularisation has seen increasingly widespread adoption in reinforcement learning as a practical approach to updating an agent's policy [342–345]. Another complementary perspective is to start from an information-theoretic objective, such as path entropy [346] or empowerment [347], and incorporate reward-seeking biases as constraints or regularisers. In both approaches, one arrives at a family of objectives that encode the inherent trade-offs between goal attainment and information processing/acquisition.

Similarly, the PDO offers a principled framework for incorporating these ideas into partially observable settings, which, to our knowledge, has not been studied as an RL objective. Our approach may provide a theoretical foundation for understanding how RL agents might balance exploration and exploitation in a way that more closely mimics human decision making, potentially leading to more robust and adaptive AI systems.

The main contributions of this chapter include (1) the derivation and introduction of the Path Divergence Objective, a novel framework for modelling bounded rationality in partially observable environments; (2) an analysis of the PDO through various decompositions to understand the decision making trade-offs underlying PDO-minimisation; (3) an algorithm to compute the PDO in certain environments for policy selection; and (4) a comparative analysis of the PDO with expected utility maximisation and Expected Free Energy, illustrating novel insights and predictions provided by our approach.

7.2 Preliminaries

POMDPs, Policies, and World Models A *Partially Observable Markov Decision Process* [62, 348] is a tuple $\mathcal{M} = (S, A, \Omega, O, T, p, I, \mathcal{U})$, where:

1. S is a finite set of *states*;
2. A is a finite set of *actions*;
3. Ω is a finite set of *observations*;
4. $O : A \times S \rightharpoonup \Delta(\Omega)$ is the partial *observation likelihood function*;
5. $T \in \mathbb{Z}^+$ is a finite *time horizon*;
6. $p : S \times A \to \Delta(S)$ is the *probabilistic transition function*;
7. $I \in \Delta(S)$ is the *initial state distribution*;
8. $\mathcal{U} : \mathbb{H} \to \mathbb{R}$ is the *history utility function* which models the agent's preferences, where $\mathbb{H}$ is the set of all histories $h^{0:t} = s^0 o^0 a^0 s^1 \dots s^t o^t, t \in \{1, \dots, T\}$.

We similarly use the notation $s^{0:t}, o^{0:t}$, and $a^{0:t-1}$ to denote state, observation, and action trajectories respectively up to time t. A *policy function* $\pi : \mathbb{O} \to \Delta(A)$ maps each observation history of the agent to a probability distribution over their actions. We let Π denote the set of all policies.

In a POMDP, an agent does not have direct access to the true state of the environment. Instead, it receives observations that provide partial information about the state. The agent's goal is to maximise the expected utility of its trajectories. The goal, as defined here, accounts for a wide range of special cases commonly encountered in reinforcement learning, such as the expected sum of discounted rewards [62] and non-Markovian rewards generated by, e.g., a reward machine [349]. The framework can also be generalised to infinite-horizon settings, but for simplicity, we consider a finite horizon. In order to compute expected rewards in $\mathcal{M}$, we define the *reach probability* of a history h under a policy π as

$$p(h; \pi) := I(s^0) \cdot \left(\prod_{\tau=0}^{T-1} O(o^\tau | a^{\tau-1}, s^\tau) \cdot \pi(a^\tau | o^{0:\tau}) \cdot p(s^{\tau+1} | s^\tau, a^\tau) \right) \cdot O(o^T | s^T),$$

Additionally, we assume that agents possess a probabilistic world model $Q(h^{0:t}; \pi)$, which captures their beliefs about the past and future. This can be written in a similar form to the reach probability as

$$Q(h^{0:t}\,;\,\pi) := Q(s^0) \cdot \left(\prod_{\tau=0}^{t-1} Q(o^\tau|s^\tau) \cdot Q(s^{\tau+1}|s^\tau, a^\tau) \cdot \pi(a^\tau|o^{0:\tau}) \right) \cdot Q(o^t|s^t).$$

We let $\mathcal{Q}$ denote the set of possible world models for an agent, noting that the set of policies Π available to an agent is contained within $\mathcal{Q}$, as policies can be interpreted as beliefs about one's own actions, conditional on different observation histories. We generally assume a perfect recall setting, in which the agent is able to perfectly remember all of its past actions and observations.

Value functions Perhaps the central concept of interest in control theory and RL is the (objective) [1] *value function*, which measures the expected reward/utility to-go for the agent from a given time point t until the end of the episode, under the policy π. Formally, the value function of the agent is given by

$$\mathcal{V}(o^{0:t}; \pi) = \mathbb{E}_{p(h^{0:T} \,|\, o^{0:t}; \pi)} \left[\mathcal{U}(h^{0:T}) \right],$$

where $h^{0:T} = h^{0:t}\, a^t s^{t+1} \ldots s^T o^T$. By $\mathcal{V}(\pi) = \mathcal{V}(\emptyset; \pi)$, we denote the total expected utility under π. In practice however, an agent only has access to their own model Q, which is used to calculate their expectations of how the future will unfold. Hence, their optimisation problem must be carried out with respect to their *subjective value function*, which we write (with a slight abuse of notation) as

$$V(Q, \pi) := \mathbb{E}_{Q(h^{0:T}; \pi)} \left[\mathcal{U}(h^{0:T}) \right].$$

Belief Updating Given a history of observations $o^{0:t}$ and a policy π, an agent may infer the latent causes of its sensory data by attempting to compute the posterior probability $P(s^{0:t}|o^{0:t}; \pi)$. However, even with perfect knowledge of the generative process giving rise to the agent's observations, computing such a posterior exactly is computationally intractable for more complex generative models, due to the need to evaluate the *model evidence* $P(o^{0:t}; \pi)$, which involves computing a sum over all possible trajectories that could have given rise to the sequence of observations. Thus, in practice, this generative model is usually

[1] We use the term "objective" to mean that the expectation of an agent's utility is computed with respect to the true reach probability, rather than a subjective estimate of it.

approximated by optimising a variational approximation $Q(h;\pi)$ to the generative model $P(h;\pi)$, which is usually chosen from some tractable family of distributions $\mathcal{Q}$ that admits a factorisation over states, policies, and observations as defined above [70]. A common form of this process is known as variational Bayesian inference, which aims to approximate the posterior P by minimising a functional known as the *variational free energy* (VFE) or the negative *evidence lower bound* (ELBO) [350, 351], which is an upper bound on the surprise, i.e., the negative log probability of the observed data:

$$\begin{aligned}
&\mathrm{D}_{\mathrm{KL}}\left[Q(s^{0:t}|o^{0:t};\pi) \,||\, P(s^{0:t}|o^{0:t};\pi)\right] \\
&= \mathbb{E}_{Q(s^{0:t}|o^{0:t};\pi)}\left[\log Q(s^{0:t}|o^{0:t};\pi) - \log P(s^{0:t}|o^{0:t};\pi)\right] \\
&= \mathbb{E}_{Q(s^{0:t}|o^{0:t};\pi)}\left[\log Q(s^{0:t}|o^{0:t};\pi) - \log P(s^{0:t},o^{0:t};\pi) + \log P(o^{0:t};\pi)\right] \\
&= \underbrace{\mathbb{E}_{Q(s^{0:t}|o^{0:t};\pi)}\left[-\log P(s^{0:t},o^{0:t};\pi)\right] - H[Q(s^{0:t}|o^{0:t};\pi)]}_{\text{variational free energy}} \underbrace{-\left(\log P(o^{0:t};\pi)\right)}_{\text{surprise}}. \qquad (7.1)
\end{aligned}$$

Finding an approximation $\hat{Q}$ to the posterior by minimising the VFE can thus be written as

$$\hat{Q}(s^{0:t}|o^{0:t};\pi) \approx \arg\min_{Q\in\mathcal{Q}} \mathbb{E}_{Q(s^{0:t}|o^{0:t};\pi)}\left[-\log P(s^{0:t},o^{0:t};\pi)\right] - H[Q(s^{0:t}|o^{0:t};\pi)]. \qquad (7.2)$$

Here, we again simplify the complexity introduced by computing such an approximation and assume that the agent's variational approximation Q does indeed minimise their VFE and can thus perform exact Bayesian inference, i.e., $\hat{Q} = P = p$. The following chapter will focus on exploring the consequences of relaxing this assumption, which as we will see, gives rise to a rich yet simple model of belief updating.

7.3 Path Divergence Objective

Here, we introduce the PDO, outlining its derivation and discussing some of its properties. In this framework, we make two additional assumptions: (1) the agent has a *prior world model* $Q^0(h;\pi^0)$, which represents their default or habitual model of the world and policy at a given point in time before doing any information processing[2]. Written in this way, the 'cognitive effort' can be read as

[2]We assume throughout this text that the prior model has full support over all trajectories.

the mental exertion required to update or overcome one's habitual or instinctual beliefs and behaviours [352]; (2) observing that information processing incurs a cost [353], and that agents expend effort when computing a posterior to improve the value function, we assume that this expenditure reduces the agent's utility linearly, and that the cost incurred can be measured by the Kullback-Leibler (KL) divergence [352, 354].[3] Given these, we can write the optimisation problem that the agent appears to be solving as

$$\max_{Q \in \mathcal{Q}} \mathbb{E}_{Q(h^{0:T};\pi)} \left[\mathcal{U}(h^{0:T}) \right] - \frac{1}{\beta} \mathrm{D}_{\mathrm{KL}} \left[Q(h^{0:T};\pi) \,||\, Q^0(h^{0:T};\pi^0) \right], \tag{7.3}$$

for some $\beta > 0$, with the convention that a choice of Q implicitly entails a choice of policy π as well. Now, suppose that we define a probability distribution $\tilde{P}(h^{0:T}) := \frac{\exp\left(\beta \mathcal{U}(h^{0:T})\right)}{Z(\beta;\mathcal{U})}$, where we let $Z(\beta;\mathcal{U}) := \sum_{h^{0:T'} \in \mathbb{H}^T} \exp\left(\beta \mathcal{U}(h^{0:T'})\right)$. This may be recognised as a softmax function or Boltzmann distribution over the utility function $\mathcal{U}$. We call this distribution the *preference model*, as it is another way of representing the agent's preferences in the form of a probability distribution. Then, re-writing the problem using the preference model, we have the following result:

Lemma 7.1. *The optimisation problem in 7.3 is equivalent to the following optimisation problem:*

$$\min_{Q \in \mathcal{Q}} D_{KL} \left[Q(h^{0:T};\pi) \,||\, \tilde{P}(h^{0:T}) \right] - \mathbb{E}_{Q(h^{0:T};\pi)} \left[\log Q^0(h^{0:T};\pi^0) \right]. \tag{7.4}$$

Proof. Recall the original optimisation problem:

$$\max_{Q \in \mathcal{Q}} \mathbb{E}_{Q(h^{0:T};\pi)} \left[\mathcal{U}(h^{0:T}) \right] - \frac{1}{\beta} \mathrm{D}_{\mathrm{KL}} \left[Q(h^{0:T};\pi) \,||\, Q^0(h^{0:T};\pi^0) \right]. \tag{7.5}$$

Now, defining the preference model as

$$\tilde{P}(h^{0:T}) := \frac{\exp\left(\beta \mathcal{U}(h^{0:T})\right)}{Z(\beta;\mathcal{U})},$$

we can rearrange this for $\mathcal{U}$, and we obtain

$$\mathcal{U}(h^{0:T}) = \frac{1}{\beta} \cdot \log \left[\tilde{P}(h^{0:T}) \cdot Z(\beta;\mathcal{U}) \right]. \tag{7.6}$$

[3] For an axiomatic justification of the latter assumption, we refer the reader to [314]. As a more heuristic argument, we note that modelling the informational cost associated with cognitive processes can be related to thermodynamics [313], which by Landauer's principle is known to imply a minimum energy expenditure requirement [310, 353], and is hence costly for the agent.

Using this, we obtain the equivalent problem

$$\begin{aligned}
&\max_{Q\in\mathcal{Q}} \frac{1}{\beta}\mathbb{E}_{Q(h^{0:T};\pi)}\left[\log \tilde{P}(h^{0:T}) + \log Z(\beta;\mathcal{U})\right] - \frac{1}{\beta}\mathrm{D}_{\mathrm{KL}}\left[Q(h^{0:T};\pi) \mid\mid Q^0(h^{0:T};\pi^0)\right] \\
&= \max_{Q\in\mathcal{Q}} \mathbb{E}_{Q(h^{0:T};\pi)}\left[\log \tilde{P}(h^{0:T}) + \log Z(\beta;\mathcal{U}) - \log Q(h^{0:T};\pi) + \log Q^0(h^{0:T};\pi^0)\right] \\
&= \min_{Q\in\mathcal{Q}} \mathbb{E}_{Q(h^{0:T};\pi)}\left[\log Q(h^{0:T};\pi) - \log \tilde{P}(h^{0:T}) - \log Z(\beta;\mathcal{U}) - \log Q^0(h^{0:T};\pi^0)\right] \\
&= \min_{Q\in\mathcal{Q}} \mathrm{D}_{\mathrm{KL}}\left[Q(h^{0:T};\pi) \mid\mid \tilde{P}(h^{0:T})\right] - \mathbb{E}_{Q(h^{0:T};\pi)}\left[\log Q^0(h^{0:T};\pi^0)\right],
\end{aligned}$$

where the last equality follows because $Z(\beta;\mathcal{U})$ is a constant. □

Thus, we see that the objective for a boundedly rational agent can be viewed as finding a model (including a policy) that minimises the KL divergence between its 'prediction' model and its 'preference' model $\tilde{P}$, with an additional cross entropy term that acts as a penalty for large deviations from Q^0 to Q. In other words, one can think of the KL divergence term as the expected excess surprise when the agent wishfully believes that trajectories are distributed according to $\tilde{P}$, when its actual belief is Q. This may be instantiated by the ability of more complex agents to imagine alternative preferred outcomes, and then subsequently update their models and take actions to resolve the discrepancy with their current beliefs.

Definition 7.1. The **Path Divergence Objective** (PDO) for an agent i in a POMDP $\mathcal{M}$ given a prior policy π^0 and a posterior policy π is given by:

$$G(Q;Q^0) := \mathrm{D}_{\mathrm{KL}}\left[Q(h^{0:T};\pi) \mid\mid \tilde{P}(h^{0:T})\right] - \mathbb{E}_{Q(h^{0:T};\pi)}\left[\log Q^0(h^{0:T};\pi^0)\right]. \tag{7.7}$$

The divergence term can be interpreted metacognitively, in that it represents the expected excess surprisal when the agent falsely believes that their belief is $\tilde{P}(h^{0:T})$ when it is actually $Q(h^{0:T};\pi)$. This objective is closely related to the Free Energy of the Expected Future [335] and the Action Perception Divergence objectives [336], which naturally model the trade-offs between taking actions to resolve uncertainty about the world and maximising rewards.[4] Also closely related is the notion of expected free energy [355, 356] and its recursive variant – the sophisticated free energy [330] – which we compare against in Section 7.4.

Decomposition of the PDO. The PDO can be decomposed in several ways, which sheds light on its connections to active inference and intrinsic motivation in reinforcement learning [70, 319, 357, 358]. Firstly, interpreting (negative)

[4] For a detailed study of such trade-offs, we refer the reader to the expositions in [332, 335].

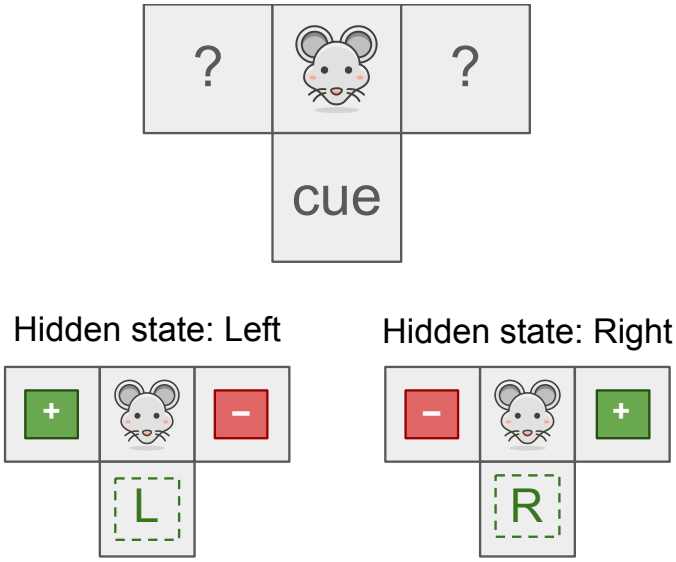

Figure 7.2: Schematic representation of the T-Maze environment. The maze consists of a start position from which two goal arms (left and right) extend, along with a third *cue* arm (bottom). The maze randomly starts in one of two states: a reward in the left arm and a punishment in the right arm, or vice versa. This state is initially hidden from the agent, but the information about the hidden state is positioned in the cue arm. The agent can visit two locations in a single experiment. We set the reward to +1 and the punishment to -4, and the cost of visiting the cue to C_{cue}.

expected utility as an energy, the PDO is an upper bound on expected energy minus entropy:

$$\begin{aligned} G(Q;Q^0) &= \underbrace{\mathbb{E}_{Q(h^{0:T};\pi)}\left[-\log \tilde{P}(h^{0:T})\right]}_{\text{-Value (Energy)}} + \underbrace{\mathrm{D}_{\mathrm{KL}}\left[Q(h^{0:T};\pi) \,||\, Q^0(h^{0:T};\pi^0)\right]}_{\text{Divergence from prior}} \\ &\geq \underbrace{\mathbb{E}_{Q(h^{0:T};\pi)}\left[-\log \tilde{P}(h^{0:T})\right]}_{\text{-Value (Energy)}} - \underbrace{H\Big[\,Q(h^{0:T};\pi)\,\Big]}_{\text{Entropy}}, \end{aligned}$$

where the inequality follows because the cross-entropy is always non-negative. Furthermore, decomposing the divergence term in the PDO reveals a natural decomposition in terms of *epistemic value*, *pragmatic value*, and an *intention-behaviour gap*, all of which have been robustly empirically observed in human behaviour [359–361].

Theorem 7.2. *If* $\tilde{P}(s^{0:T}|o^{0:T},a^{0:T}) = Q(s^{0:T}|o^{0:T},a^{0:T})$*, then the divergence term in the*

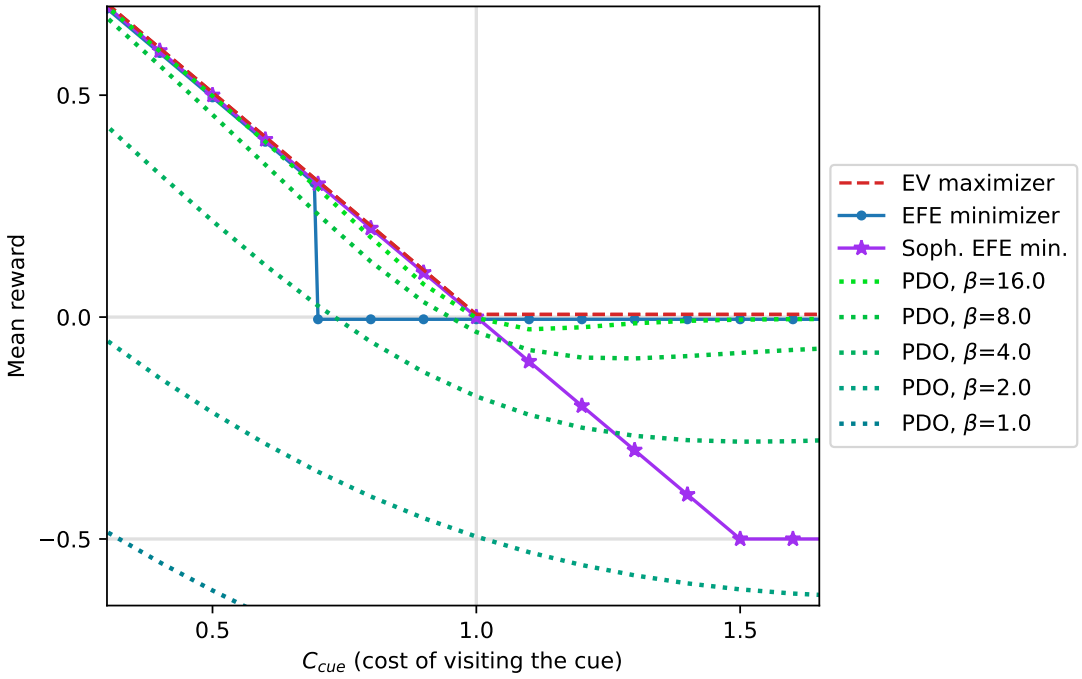

Figure 7.3: A plot of the mean reward obtained by several decision making models depending on C_{cue}. The EV maximiser plays the optimal strategy: when the cost of visiting the cue is over 1.0, it is optimal to do nothing. The PDO models various degrees of rationality (β) and smoothly approaches this optimum for $\beta \to \infty$; for $\beta \to 0$, this would correspond to playing π^0 (here a uniform policy). The EFE and Sophisticated EFE exhibit a different type of bounded rationality and decide suboptimally in different ranges: EFE stops visiting the cue when the C_{cue} is larger than its information gain term. Sophisticated EFE visits the cue even for $1.0 < C_{cue} < 1.5$, over-estimating the value of the cue information, and then for $C_{cue} > 1.5$ it still learns the value of the cue *indirectly* by visiting a random arm and inferring the cue from there, correcting the action on the second turn to get mean reward -0.5. (Note: the plot lines are slightly offset to minimise overlap.)

PDO can be decomposed as:

$$D_{KL}\left[Q(h^{0:T};\pi) \,||\, \tilde{P}(h^{0:T})\right] = \underbrace{-\mathbb{E}_{Q(o^{0:T},a^{0:T};\pi)}\left[D_{KL}\left[Q(s^{0:T}|o^{0:T},a^{0:T}) \,||\, Q(s^{0:T}|a^{0:T})\right]\right]}_{\textit{Epistemic Value}}$$
$$\underbrace{+\mathbb{E}_{Q(s^{0:T},a^{0:T};\pi)}\left[D_{KL}\left[Q(o^{0:T}|s^{0:T},a^{0:T}) \,||\, \tilde{P}(o^{0:T}|a^{0:T})\right]\right]}_{\textit{Pragmatic Value}} + \underbrace{D_{KL}\left[Q(a^{0:T};\pi) \,||\, \tilde{P}(a^{0:T})\right]}_{\textit{Intention-Behaviour Gap}}.$$

Proof. We can write the divergence term $\mathrm{D}_{\mathrm{KL}}\left[Q(h^{0:T};\pi) \,||\, \tilde{P}(h^{0:T})\right]$ under the

assumption that $\tilde{P}(s^{0:T}|o^{0:T}, a^{0:T}) = Q(s^{0:T}|o^{0:T}, a^{0:T})$ as follows:

$$
\begin{aligned}
&\mathbb{E}_{Q(h^{0:T};\pi)}\Big[\log Q(s^{0:T}|a^{0:T}) + \log Q(o^{0:T}|s^{0:T}, a^{0:T}) + \log Q(a^{0:T}) \\
&\qquad\qquad - \log \tilde{P}(s^{0:T}|o^{0:T}, a^{0:T}) - \log \tilde{P}(o^{0:T}|a^{0:T}) - \log \tilde{P}(a^{0:T})\Big] \\
=&\mathbb{E}_{Q(h^{0:T};\pi)}\Big[\log Q(s^{0:T}|a^{0:T}) + \log Q(o^{0:T}|s^{0:T}, a^{0:T}) + \log Q(a^{0:T}) \\
&\qquad\qquad - \log Q(s^{0:T}|o^{0:T}, a^{0:T}) - \log \tilde{P}(o^{0:T}|a^{0:T}) - \log \tilde{P}(a^{0:T})\Big] \\
=&-\underbrace{\mathbb{E}_{Q(o^{0:T},a^{0:T};\pi)}\Big[\mathrm{D}_{\mathrm{KL}}\Big[Q(s^{0:T}|o^{0:T}, a^{0:T}) \,||\, Q(s^{0:T}|a^{0:T})\Big]\Big]}_{\text{Epistemic Value}} \\
&+\underbrace{\mathbb{E}_{Q(s^{0:T},a^{0:T};\pi)}\Big[\mathrm{D}_{\mathrm{KL}}\Big[Q(o^{0:T}|s^{0:T}, a^{0:T}) \,||\, \tilde{P}(o^{0:T}|a^{0:T})\Big]\Big]}_{\text{Pragmatic Value}} \\
&+\underbrace{\mathrm{D}_{\mathrm{KL}}\Big[Q(a^{0:T};\pi) \,||\, \tilde{P}(a^{0:T})\Big]}_{\text{Intention-Behaviour Gap}}.
\end{aligned}
$$

□

The condition in Theorem 7.2 can be interpreted as the assumption that agents' preferences are only defined over components of their interface with the environment, i.e., their Markov blanket, and not directly over underlying states of the world. In the case of metacognitive agents [362], preferences may not be restricted only to one's observations or actions, but could also be defined over one's own internal world model. This is the main subject of focus in the subsequent chapter.

Unpacking this decomposition intuitively, we observe the following:

1. The *epistemic value*, also known as the expected information gain [363–365], scores the expected reduction in uncertainty about the state trajectory before and after knowing the observation trajectory. Notice that since the distributions which are being compared are conditional on the chosen action trajectory, the agent has a bias towards *active data sampling* to advance their understanding about the underlying state of the world [366–368].

2. The *pragmatic value* similarly scores the expected divergence between the agent's predictions about their own observations and their preferences over the same [334]. We can additionally decompose the pragmatic value term further as

$$
\mathbb{E}_{Q(s^{0:T},a^{0:T};\pi)}\Big[-H[Q(o^{0:T}|s^{0:T}, a^{0:T})] - \mathbb{E}_{Q(o^{0:T}|s^{0:T},a^{0:T})}\Big[\log \tilde{P}(o^{0:T}|a^{0:T})\Big]\Big].
$$

The first term can be interpreted as an entropy-regulariser which motivates the agent to seek out diverse or novel experiences [346, 369], while the second term can be interpreted as an expected utility over observations, conditional on the agent's actions.

3. The *intention-behaviour gap*, or value-action gap, can be interpreted as capturing the difference between an agent's preferences over their own actions and what their expectations over the same, given the posterior policy [361]. Such a gap is one contributor towards the experience of cognitive dissonance [370, 371] or predictive dissonance [372], which agents will attempt to minimise under this decomposition. The situation of this term amongst the epistemic and pragmatic value components may partially explain why individuals do not always act in a way consistent with their stated preferences, that is, the epistemic or pragmatic benefits of acting in a certain manner may outweigh the intention-behaviour gap induced by such behaviour.

For the remainder of this chapter, we will focus on the setting where the agent's world model in relation to the hidden states and their observations is fixed, and only their policy is allowed to vary. This is a common assumption that is often focused on in control theory, reinforcement learning, and game theory. For this reason, we will write $G(\pi; \pi^0)$ as a special case of the PDO under this restriction.

Theorem 7.3. *Under the assumption that the world model Q is fixed for the transition and observation likelihood functions, the Path Divergence Objective can be expressed in the following recursive forms:*

(a) With $\tilde{P}$ over the full path:

$$G(\pi; \pi^0) = \mathbb{E}_{Q(s^0, o^0)} G^0_{\mathrm{P}}(\pi | s^0, o^0; \pi^0), \textit{ where} \tag{7.8}$$

$$G^t_{\mathrm{P}}(\pi | h^{0:t}; \pi^0) = D_{KL}\left[\pi(a^t|o^{0:t}) \;||\; \pi^0(a^t|o^{0:t})\right] + \mathbb{E}_{Q(a^t, s^{t+1}, o^{t+1}|h^{0:t};\pi)}\left[G^{t+1}_{\mathrm{P}}(\pi | h^{0:t+1}; \pi^0)\right] \tag{7.9}$$

$$G^T_{\mathrm{P}}(\pi | h^{0:T}; \pi^0) = -\log \tilde{P}(h^{0:T}). \tag{7.10}$$

(b) With $\tilde{P}$ as conditionals: *For any decomposition of $\tilde{P}$ into a chain of conditional distributions of the form*

$$\tilde{P}(h^{0:t}) = \prod_{t=0}^{T-1} \tilde{P}^t(a^t, s^{t+1}, o^{t+1} | h^{0:t}),$$

we can express the PDO as

$$G(\pi;\pi^0) = \mathbb{E}_{Q(s^0,o^0)} G^0_{\mathrm{C}}(\pi|s^0,o^0;\pi^0), \textit{ where} \tag{7.11}$$

$$\begin{aligned} G^t_{\mathrm{C}}(\pi|h^{0:t};\pi^0) &= D_{KL}\left[\pi(a^t|o^{0:t}) \,||\, \pi^0(a^t|o^{0:t})\right] \\ &\quad + \mathbb{E}_{Q(a^t,s^{t+1},o^{t+1}|h^{0:t};\pi)}\Big[G^{t+1}_{\mathrm{C}}(\pi|h^{0:t+1};\pi^0) - \log \tilde{P}^t(a^t,s^{t+1},o^{t+1}|h^{0:t})\Big] \end{aligned} \tag{7.12}$$

$$G^T_{\mathrm{C}}(\pi|h^{0:T};\pi^0) = 0. \tag{7.13}$$

(c) Markovian preferential distribution: *Assuming that $\tilde{P}^t$ from (b) only depends on the previous state and the observation history, i.e. $\tilde{P}^t(a^t,s^{t+1},o^{t+1}|h^{0:t}) = \tilde{P}^t(a^t,s^{t+1},o^{t+1}|o^{0:t},s^t)$, we have*

$$G(\pi;\pi^0) = \mathbb{E}_{Q(s^0)Q(o^0|s^0)}\left[G^0_{\mathrm{M}}(\pi|o^0,s^0;\pi^0)\right] \textit{ , where} \tag{7.14}$$

$$\begin{aligned} G^t_{\mathrm{M}}(\pi|o^{0:t},s^t;\pi^0) &= D_{KL}\left[\pi(a^t|o^{0:t}) \,||\, \pi^0(a^t|o^{0:t})\right] + \mathbb{E}_{\pi(a^t|o^{0:t})Q(s^{t+1}|s^t,a^t)Q(o^{t+1}|s^{t+1})}\Big[\\ &\quad - \log \tilde{P}^t(a^t,s^{t+1},o^{t+1}|o^{0:t},s^t) + G^{t+1}_{\mathrm{M}}(\pi|o^{0:t+1},s^{t+1};\pi^0)\Big], \end{aligned} \tag{7.15}$$

$$G^T_{\mathrm{M}}(\pi|o^{0:T},s^T;\pi^0) = 0. \tag{7.16}$$

Proof. Recall the PDO for planning:

$$G(\pi;\pi^0) = \mathbb{E}_{Q(h^{0:T};\pi)}\Big[\log Q(h^{0:T};\pi) \;-\; \log \tilde{P}(h^{0:T}) \;-\; \log Q(h^{0:T};\pi^0)\Big].$$

With the usual factorisation

$$\begin{aligned} \log Q(h^{0:T};\pi) = \log Q(s^0,o^0) + \sum_{t=0}^{T-1}\Bigg(&\log \pi(a^t \mid o^{0:t}) + \log Q(s^{t+1} \mid s^t,a^t) \\ &+ \log Q(o^{t+1} \mid s^{t+1})\Bigg), \\ \log Q(h^{0:T};\pi^0) = \log Q(s^0,o^0) + \sum_{t=0}^{T-1}\Bigg(&\log \pi^0(a^t \mid o^{0:t}) + \log Q(s^{t+1} \mid s^t,a^t) \\ &+ \log Q(o^{t+1} \mid s^{t+1})\Bigg), \end{aligned}$$

their difference cancels the world-model terms, yielding

$$G(\pi;\pi^0) = \mathbb{E}_{Q(h^{0:T};\pi)}\left[\sum_{t=0}^{T-1}\Big(\log \pi(a^t \mid o^{0:t}) - \log \pi^0(a^t \mid o^{0:t})\Big) - \log \tilde{P}(h^{0:T})\right]. \tag{7.17}$$

Now define recursively, for $t = T, T-1, \dots, 0$,

$$G_{\mathrm{P}}^{T}(\pi|h^{0:T};\pi^0) \coloneqq -\log \tilde{P}(h^{0:T}), \tag{7.18}$$

$$G_{\mathrm{P}}^{t}(\pi|h^{0:t};\pi^0) \coloneqq \mathrm{D}_{\mathrm{KL}}\left[\pi(a^t|o^{0:t}) \,||\, \pi^0(a^t \mid o^{0:t})\right] \tag{7.19}$$

$$+ \mathbb{E}_{Q(a^t,s^{t+1},o^{t+1}|h^{0:t};\pi)}\left[G_{\mathrm{P}}^{t+1}(\pi|h^{0:t+1};\pi^0)\right]. \tag{7.20}$$

A backward induction shows that

$$\mathbb{E}_{Q(h^{0:t};\pi)}\Big[G_{\mathrm{P}}^{t}(\pi \mid h^{0:t};\pi^0)\Big] = \mathbb{E}_{Q(h^{0:T};\pi)}\Bigg[\sum_{\tau=t}^{T-1}\Big(\log \pi(a^\tau \mid o^{0:\tau}) - \log \pi^0(a^\tau \mid o^{0:\tau})\Big) - \log \tilde{P}(h^{0:T})\Bigg]. \tag{7.21}$$

The base case $t = T$ is immediate from 7.20. For the inductive step, we can expand 7.20, take expectations under $Q(h^{0:t};\pi)$, apply the tower property to the inner expectation over (a^t, s^{t+1}, o^{t+1}), and use the induction hypothesis at $t+1$. Setting $t = 0$ and taking expectation over $Q(s^0, o^0)$ recovers 7.17, which establishes part (a).

For part (b), writing the preference as a product of conditionals $\tilde{P}(h^{0:T}) = \prod_{t=0}^{T-1} \tilde{P}^t(a^t, s^{t+1}, o^{t+1} \mid h^{0:t})$, we obtain

$$-\log \tilde{P}(h^{0:T}) = -\sum_{t=0}^{T-1} \log \tilde{P}^t(a^t, s^{t+1}, o^{t+1} \mid h^{0:t}).$$

Substituting this into 7.17 and repeating the dynamic-programming construction above simply moves the $-\log \tilde{P}^t(\cdot)$ term from the terminal condition into the per-timestep recursion, yielding the defined G_{C}^t with terminal condition $G_{\mathrm{C}}^T \equiv 0$.

For part (c), under the stated Markovian dependence $\tilde{P}^t(a^t, s^{t+1}, o^{t+1} \mid h^{0:t}) = \tilde{P}^t(a^t, s^{t+1}, o^{t+1} \mid o^{0:t}, s^t)$ and the world-model factorisation $Q(a^t, s^{t+1}, o^{t+1} \mid h^{0:t};\pi) = \pi(a^t \mid o^{0:t})\, Q(s^{t+1} \mid s^t, a^t)\, Q(o^{t+1} \mid s^{t+1})$, the expectation in part (b) factorises as in the statement, which proves the G_{M}^t recursion. □

7.4 Algorithmic and Experimental Results

We propose an algorithm to find the policy $\hat{\pi}$ that minimises the PDO for a certain wide class of preference models $\tilde{P}$ and for environments with perfect recall. We implemented the algorithm as an extension to the PyMDP [373] library for active inference in finite-horizon POMDPs, and we demonstrate the properties of the PDO on a standard T-Maze environments with a cue [374, 375], comparing it to expected value maximisation (EV), Expected Free Energy minimisation (EFE) and its sophisticated version [330].

Algorithm For Computing the PDO

A *perfect recall environment* is one where the agent observes and remembers not just all its observations but also all its actions, i.e. any reachable sequence $o^{0:t}$ uniquely determines the only sequence of $a^{0:t-1}$ that may have led to it. Each action sequence may lead to multiple observation sequences (non-determinism), and there may be unreachable observation sequences.

Theorem 7.4. *Assume that condition (c) of Theorem 7.3 holds, that is $\tilde{P}$ can be decomposed into factors $\tilde{P}^t$ such that $\tilde{P}(h^{0:t}) = \prod_{t=0}^{T-1} \tilde{P}^t(a^t, s^{t+1}, o^{t+1}|h^{0:t})$ and $\tilde{P}^t$ only depends on the previous state and the observation history, i.e. $\tilde{P}^t(a^t, s^{t+1}, o^{t+1}|h^{0:t}) = \tilde{P}^t(a^t, s^{t+1}, o^{t+1}|o^{0:t}, s^t)$. Then, we can compute the policy $\hat{\pi}$ minimising $G(\pi; \pi^0)$ for any given perfect recall environment, any such $\tilde{P}^t$, and any given π^0, in time $\mathcal{O}(|\mathbb{O}| \cdot |S| \cdot (T_{\tilde{P}^t} + T_Q))$,[5] where S is the set of all states, $\mathbb{O}$ is the set of all reachable sequences of observations (within the time horizon) and their prefixes, and $T_{\tilde{P}^t}$ and T_Q are the times required to evaluate $\tilde{P}^t$ and Q, respectively.*

Proof. First, define $G_{\mathrm{M}'}$, a variant of G_{M} where we condition not on the last state, but rather on a random variable ξ^t representing a distribution (belief) over the last state:

$$
\begin{aligned}
G(\pi; \pi^0) &= \mathbb{E}_{Q(o^0)}\Big[G^0_{\mathrm{M}'}(\pi|o^0, \xi^0 = Q(s^0|o^0); \pi^0)\Big], \text{ where} \\
G^t_{\mathrm{M}'}(\pi|o^{0:t}, \xi^t; \pi^0) &= \mathbb{E}_{\pi(a^t|o^{0:t})}\Big[\log \pi(a^t|o^{0:t}) - \log \pi^0(a^t|o^{0:t}) + \mathbb{E}_{Q(\xi^{t+1}|\xi^t, a^t)Q(o^{t+1}|\xi^{t+1})}\Big[\\
&\quad - \mathbb{E}_{s^{t+1}\sim\xi^{t+1}} \log \tilde{P}(a^t, s^{t+1}, o^{t+1}|o^{0:t}) + G^{t+1}_{\mathrm{M}'}(\pi|o^{0:t+1}, \xi^{t+1}; \pi^0)\Big]\Big] \\
&= \mathbb{E}_{\pi(a^t|o^{0:t})}\Big[\log \pi(a^t|o^{0:t}) + F^t(a^t, o^{0:t}, \xi^t, \pi^0)\Big] \\
G^T_{\mathrm{M}'}(\pi|o^{0:T}, \xi^T; \pi^0) &= 0.
\end{aligned}
$$

Here $F^t(a^t, o^{0:t}, \xi^t, \pi^0)$ merely collects all the terms in the preceding equation except the first. Notably, it does not depend on π and can be evaluated for every individual a^t independently.

The algorithm to find $\hat{\pi}$ proceeds as if evaluating $G^0_{\mathrm{M}'}(\pi|o^0, \xi^0)$ by expanding it recursively, finding the optimal $\hat{\pi}$ along the way and returning it, along with the final value of G. We start with several observations before stating the algorithm.

Observe that in the evaluation tree, $G^t_{\mathrm{M}'}$ is only evaluated once for any given $o^{0:t}$, and $\pi(a^t|o^{0:t})$ only appears in that one evaluation, and moreover $\pi(a^t|o^{0:t})$

[5] We use $\mathcal{O}$ big O (worst-case time complexity) notation so as not to confuse it with the observation set O.

can be chosen independently from π for all other observations. Further observe that $G_{\mathrm{M}'}$ can in fact be minimised by minimising each $G_{\mathrm{M}'}(o^{0:t}, \xi^t)$ independently, since $G_{\mathrm{M}'}$ only appears as a positive term in other $G_{\mathrm{M}'}(\dots)$s, and the value of ξ^{t+1} passed down the recursion does *not* depend on π but rather is conditioned on a single action a^t.

Therefore, $\pi(a^t|o^{0:t})$ can be optimised locally after first calculating all values of $G^{t+1}_{\mathrm{M}'}(\pi|o^{0:t+1}, \xi^{t+1}; \pi^0)$. The $\pi(a^t|o^{0:t})$ minimising $\mathbb{E}_{\pi(a^t|o^{0:t})}\Big[\log \pi(a^t|o^{0:t}) + F^t(a^t, o^{0:t}, \xi^t, \pi^0)\Big]$ is the Boltzmann distribution where F plays the role of the expected energy of the action:

$$\hat{\pi}(a^t|o^{0:t}) = \frac{e^{-F^t(a^t, o^{0:t}, \xi^t, \pi^0)}}{Z^t(o^{0:t}, \xi^t, \pi^0)}, \tag{7.22}$$

where $Z^t(o^{0:t}, \xi^t, \pi^0)$ is a distribution normalisation constant.

The algorithm is then as follows: Traverse the tree of evaluating $G_{\mathrm{M}'}$ recursively. While evaluating the tree node $G^t_{\mathrm{M}'}(\pi|o^{0:t}, \xi^t; \pi^0)$, first evaluate the quantity $G^{t+1}_{\mathrm{M}'}(\pi|o^{0:t+1}, \xi^{t+1}; \pi^0)$ for all a^t and o^{t+1} recursively, combining the returned partial policies $\hat{\pi}$. Then set $\hat{\pi}(a^t|o^{0:t})$ according to equation (7.22), and return the updated policy along with the (directly computed) value of $G^t_{\mathrm{M}'}(\pi|o^{0:t}, \xi^t; \pi^0)$.

The runtime follows from visiting each $o^{0:t} \in \mathbb{O}$ only once, and each evaluation does $\mathcal{O}(|S| \cdot (T_{\tilde{P}^t} + T_Q))$ work. The algorithm is efficient since every algorithm without further assumptions on $\tilde{P}^t$ and Q needs to evaluate them on all observation sequences, otherwise we can engineer $\tilde{P}^t$ and Q that would encode an exceedingly high reward in the omitted branch. □

Complexity. Theorem 7.4 expresses the runtime in terms of $|\mathbb{O}|$. In the worst case, the number of observation prefixes grows as $|\mathbb{O}| \leq \sum_{t=0}^{T} |\Omega|^t = \mathcal{O}(|\Omega|^T)$, so Algorithm 4 runs in time exponential in the horizon. This dependence is unavoidable if one insists on outputting an explicit history-dependent policy table $o^{0:t} \mapsto \hat{\pi}(\cdot|o^{0:t})$, since that table has worst-case size at least $|\mathbb{O}| \cdot |A|$ (one action distribution for each reachable observation prefix).

Note that this algorithm can also work for a "full path" formulation similar to Theorem 7.3(a) if $\tilde{P}$ only depends on the observation sequence and the last state (i.e. $\tilde{P}(h^{0:T}) = \tilde{P}(a^{0:T-1}, o^{0:T}, s^T)$), as $\tilde{P}^t$ can be assumed to be trivial (e.g. uniform) for all $t < T$, and only have nontrivial $\tilde{P}^T(a^{T-1}, s^T, o^T|o^{0:T-1}) = \tilde{P}(a^{0:T-1}, o^{0:T}, s^T)$

Algorithm 4 COMPUTE-PDO-POLICY (perfect recall)

Input: World model Q (belief update and predictive likelihood), preference model $\tilde{P}$ (e.g. via factors $\tilde{P}^t$), prior policy π^0, horizon T **Output**: Policy $\hat{\pi} = \arg\min_\pi G(\pi; \pi^0)$ and objective value $G(\hat{\pi}; \pi^0)$

Initialise an empty policy table $\hat{\pi}$
$G \leftarrow 0$
for all o^0 with $Q(o^0) > 0$ **do**
  $\xi^0 \leftarrow Q(s^0|o^0)$
  $G \leftarrow G + Q(o^0) \cdot \text{SOLVE}(0, o^0, \xi^0)$
end for
return $(\hat{\pi}, G)$

function SOLVE$(t, o^{0:t}, \xi^t)$
if $t = T$ **then**
  return 0
end if
for all $a^t \in A$ **do**
  $F[a^t] \leftarrow -\log \pi^0(a^t|o^{0:t})$
  for all $o^{t+1} \in \Omega$ with $Q(o^{t+1}|\xi^t, a^t) > 0$ **do**
    $\xi^{t+1} \leftarrow Q(s^{t+1}|\xi^t, a^t, o^{t+1})$
    $g \leftarrow \text{SOLVE}(t+1, (o^{0:t}, o^{t+1}), \xi^{t+1})$
    $c \leftarrow -\sum_{s^t, s^{t+1}} Q(s^t, s^{t+1} \mid \xi^t, a^t, o^{t+1}) \log \tilde{P}^t(a^t, s^{t+1}, o^{t+1} \mid o^{0:t}, s^t) + g$
    $F[a^t] \leftarrow F[a^t] + Q(o^{t+1}|\xi^t, a^t) \cdot c$
  end for
end for
$Z \leftarrow \sum_{a^t \in A} \exp\{-F[a^t]\}$
for all $a^t \in A$ **do**
  $\hat{\pi}(a^t|o^{0:t}) \leftarrow \exp\{-F[a^t]\}/Z$ (cf. Eq. 7.22)
end for
return $-\log Z$

(note that due to the perfect recall assumption, past actions are implied by the past observations).

Optimal policy search. We propose and implement an algorithm to compute a PDO-minimising policy under the following assumptions: (1) an environment with perfect recall of actions, i.e., every *reachable* observation sequence $o^{0:t}$ uniquely determines the sequence of actions $a^{0:t-1}$ that has led to it. Secondly, a decomposition of $\tilde{P}$ into temporal factors $\tilde{P}^t$ such that we have $\tilde{P}(h^{0:T}) = \prod_{t=0}^{T-1} \tilde{P}^t(a^t, s^{t+1}, o^{t+1}|a^{0:t-1}, o^{0:t}, s^t)$.

The algorithm computes the optimal policy π minimising $G(\pi; \pi^0)$ for any such environment, any given π^0, and any $\tilde{P}^t$ as above, in time $\mathcal{O}(|\mathbb{O}| \cdot |S| \cdot (T_{\tilde{P}^t} + T_Q))$

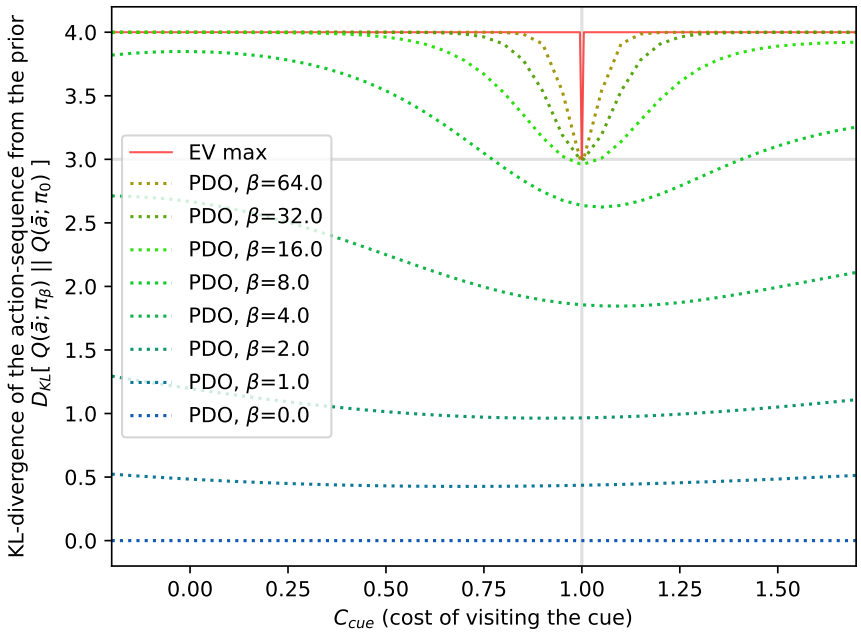

Figure 7.4: The divergence of the action-sequence distribution under Q when playing according to π_β vs according to π^0. Note that $\beta = 0$ implies playing π^0 (here a uniform policy). A fully deterministic policy (i.e. perfect control) generally requires 4 bits of information relative to the uniform action distribution, that is, 2 bits on each round. For $C_{\text{cue}} = 1.0$, there are two equally good courses of action (do nothing twice, or explore the cue and go to the reward location) so a Maximum Entropy policy only requires 3 bits of control even for expected value maximisation ("EV max" in the plot).

, where S is the set of all states, $\mathbb{O}$ is the set of all prefixes of reachable sequences of observations, and $T_{\tilde{P}^t}$ and T_Q are the times required to evaluate $\tilde{P}^t$ resp Q.

Experimental demonstration of the PDO. We study properties of the PDO on a standard T-Maze environment with a cue [374, 375]. This is a simple and commonly-used environment for studying cognition, information-seeking, and decision making under uncertainty. See Figure 7.2 for a description of the environment.

Figure 7.3 compares the expected reward of π under various models of decision making: the PDO for various values of β, the expected value maximising policy, and two other models of agency and information-seeking under uncertainty: the Expected Free Energy (EFE) [376] and the Sophisticated Expected Free Energy [330]. Figure 7.4 shows the control cost (KL to the prior) as the cost of visiting the cue is varied, for agents with different values of β. Note that the primary goal here is not to try to maximise the expected value, but rather to

study the qualitative differences between the models. These results demonstrate how varying the β parameter affects the behaviour of the agent, which is notably distinct from that of an expected free energy minimising policy. In particular, the PDO minimising policy exhibits a smooth variation in the mean reward as the cost of visiting the cue is varied, whereas the (sophisticated) EFE minimising policy exhibits a sharp transition in behaviour at a certain cost of visiting the cue, beyond which the mean reward remains constant. The experiments were carried out with the PyMDP library [373], adding our own implementation of the PDO and an expectation-maximising algorithm into the framework.

7.5 Discussion

In this chapter, we have introduced the Path Divergence Objective, a novel objective for modelling boundedly rational model-based planning in partially observable environments. Derived from an information-theoretic model of bounded rationality, the PDO balances reward-seeking behaviour with information processing constraints, parameterised by a single "rationality" parameter β. We have then demonstrated how to naturally decompose the PDO into epistemic value, pragmatic value, and intention-behaviour gap under a certain condition on the preference and world models, and derived an algorithm for computing PDO-optimal policies in perfect recall environments. Importantly, the PDO converges to expected value maximisation as β approaches infinity, establishing a clear link to classical decision theory [60].

Future research directions include connecting the PDO with applications in behavioural modelling, incentive design, AI alignment, and game theory. We aim to develop more scalable algorithms using MCTS-like approaches and function approximators, and empirically compare the PDO's behaviour against existing RL and POMDP algorithms. This flexible, theoretically-grounded framework opens up new possibilities for developing robust AI systems and advancing our understanding of biological cognition. Additional work could investigate learning dynamics under the PDO and develop more detailed models incorporating additional cognitive structures, potentially inspiring novel directions in AI research and cognitive modelling. Current limitations include, e.g., not accounting for the cost of learning world model parameters, inferring posteriors, and imperfect plan execution. We hope that further development of the PDO will lead to a versatile toolset for analysing and designing decision-makers to accommodate a wide range of cognitive constraints and real-world scenarios.

8

Boundedly Rational Belief Change

> *[H]uman reason is both biased and lazy. Biased because it overwhelmingly finds justifications and arguments that support the reasoner's point of view, lazy because reason makes little effort to assess the quality of the justifications and arguments it produces.*
>
> Mercier and Sperber, *The Enigma of Reason* (p. 9) [377]

In this chapter, we explore the role of incentives in the belief formation process of boundedly rational agents, drawing on a vast array of literature from the cognitive sciences on the ways in which human belief updating does not conform to the standard Bayesian paradigm. Indeed, the human mind is capable of extraordinary achievements, yet it often appears to work against itself. It actively defends its cherished beliefs even in the face of contradictory evidence, conveniently interprets information to conform to desired narratives, and selectively searches for or avoids information to suit its various purposes. Despite these behaviours deviating from common normative standards for belief updating, we argue that such 'biases' are not inherently cognitive flaws, but rather an adaptive response to the significant pragmatic and cognitive costs associated with revising one's beliefs. This chapter introduces a formal framework that attempts to model the influence of these costs on our belief updating mechanisms.

We treat belief updating as a motivated variational decision, where agents weigh the perceived 'utility' of a belief against the informational cost required

to adopt a new belief state, quantified by the Kullback-Leibler divergence from the prior to the variational posterior. We perform computational experiments to demonstrate that simple instantiations of this resource-rational model can be used to qualitatively emulate commonplace human behaviours, including confirmation bias and attitude polarisation. In doing so, we suggest that this framework makes steps towards a more holistic account of the motivated Bayesian mechanics of belief change and provides practical insights for predicting, compensating for, and correcting deviations from desired belief updating processes.

The framework presented here is the natural complement to the model developed in the preceding chapter. Whereas Chapter 7 applied informational costs to the problem of policy selection (penalising deviations from a default policy via the Kullback-Leibler divergence between the agent's policy and a reference prior), we now turn to the analogous question of how such costs govern the revision of beliefs. The difference lies in the object to which that measure is applied: policies in Chapter 7 and belief distributions here. This structural parallel reflects the broader thesis that a unified, information-theoretic account of bounded rationality must address both the deliberative and the epistemic dimensions of cognition.

Humanity faces an increasingly paradoxical epistemic problem. Never before have people been able to obtain so much information so quickly, yet at the same time, many societies have become increasingly polarised and paralysed by conflicting narratives. A feature that is common to several manifestations of this predicament, including public health crises and conspiracy theorising, is the presence of actors who tenaciously defend beliefs long after the balance of evidence supporting them has shifted. What, exactly, makes changing our minds so hard?

A natural approach to answering this question may begin by supposing a normative standard or benchmark against which to compare the actual processes of human belief change. The primary normative model for rational belief updating is Bayesian reasoning. According to this model, probabilistic beliefs should be adjusted proportionally to the strength of evidence according to Bayes' rule. However, the persistent discrepancy between the Bayesian standard and actual human belief updating raises questions about whether our epistemic processes are inherently irrational, or if something is missing from the traditional picture [378].

The Bayesian paradigm largely remains silent on *why* humans fall short of its ideals, primarily due to its assumptions that belief-revision is cost-free and that the driver of epistemic processes should be probabilistic coherency [379]. In practice, revising one's beliefs incurs metabolic costs, cognitive effort, and crucially, pragmatic risks and opportunities. A scientist retracting a cherished hypothesis, a politician breaking ranks with their party, or a public figure admitting error each pay tangible costs that a cost-free Bayesian calculus does not account for. Without a principled way to model the effect of such costs on belief revision, apparent "irrationalities" including confirmation bias [380], motivated reasoning [381], attitude polarisation [382], and belief persistence [383] seem like fundamental flaws in human cognition.

We argue that such apparent deviations from Bayesian norms are adaptive responses of agents operating under motivational considerations and real resource constraints. In particular, we formalise cognitive belief-change costs using the KL divergence to quantify informational distances between belief states, representing the 'informational work' required for belief state transitions. Our approach also integrates social and pragmatic factors. Beliefs are influenced by identity, social status, and interpersonal dynamics; fears of ostracism or admitting errors can increase resistance to change. Our hope is that through such modelling, we can take steps towards developing general frameworks that explain not just isolated sources of non-Bayesian belief updating, but also the inherent trade-offs between competing considerations associated with changing our minds.

Contributions and Structure

Our primary contribution is the proposal of a variational cost functional for belief revision that models the influence of pragmatic affordances and cognitive costs on human belief updating. Secondly, we present results from simplified computational experiments demonstrating how varying conservatism and likelihood weighting parameters qualitatively exhibit phenomena such as confirmation bias, evidence search asymmetries, and attitude polarisation. Finally, we discuss the implications of our model and promising future directions.

8.1 Related Work

> *There is considerable evidence that people are more likely to arrive at conclusions that they want to arrive at, but their ability to do so is constrained by their ability to construct seemingly reasonable justifications for these conclusions.*
>
> — Ziva Kunda, *The Case for Motivated Reasoning* [381]

Decision-Theoretic Models of Belief Updating

Drawing on frameworks of decision making, human belief revision is increasingly being treated as a value-based decision [384–386]. On this view, beliefs are updated not purely based on their accuracy, but are associated with a *utility*. The utility of holding a belief is derived from the outcomes it leads to, which can be internal (emotional comfort, positive feelings) or external (acceptance within a community, job opportunities) [386]. This is supported by several arguments highlighting the centrality of affect in decision making and belief-updating [387–389]. Certain beliefs may give rise to utility in proportion to how well they track or predict reality, in which case, there is an incentive for the agent to seek truthful beliefs. Other beliefs may demand that the agent confabulates an elaborate yet tenuous narrative that coincidentally supports their desired conclusion. In other words, a belief's usefulness can be orthogonal to its truthfulness.

Cognitive Costs

In addition to the pragmatic incentives that shape belief updating, there are unavoidable costs that any agent must pay to change their beliefs. These costs have been studied from the perspective of bounded/resource/computational rationality, where the presence of some form of cost associated with cognition is explicitly modelled in an agent's decision making [56, 307, 313, 314, 327, 328, 352, 390]. Belief updating can be understood as a thermodynamic process involving transitions between mental states, where each transition incurs unavoidable dissipative costs [310]. These costs arise from fundamental physical principles governing information processing in biological systems at the level of neural computation and metabolic energy expenditure [312, 391].

The transition between belief states involves both work-like and heat-like components. The work-like component corresponds to the directed shift of the belief state, while heat dissipation occurs in the form of entropy production during the transitions between belief states in finite amounts of time [392]. The total dissipation produced by belief updating can be quantified as the difference between the reversible work theoretically possible and the actual work captured during the transition process. This represents the unavoidable cost of finite-time belief changes [393].

This thermodynamic perspective helps to explain why, other considerations being equal, rapid belief changes tend to be more costly and inefficient compared to gradual updates. The system must balance the speed of belief revision against the increased dissipative costs of rapid change. This intuition can be made more precise using concepts from finite-time thermodynamics [392]. The total entropy production in a sequence of step-equilibrations is bounded by $\Delta S^u \geq \frac{L^2}{2K}$, where L is the thermodynamic length of the belief change pathway and K is the number of intermediate equilibration steps [394]. Thus, increasing the number of steps decreases the lower bound on total entropy production, permitting more efficient pathways of belief change. The brain appears to possess several remarkable features that aid in minimising these costs. For instance, the efficient coding hypothesis suggests that neural representations of sensory information are structured to minimise the number of neuronal spikes required to transmit a given signal [395].

Understanding these fundamental thermodynamic constraints provides insight into why belief change can be so difficult even in the presence of contradictory evidence. The brain must carefully balance the energetic and informational costs of updating against the potential benefits. It is unclear precisely how significantly the thermodynamic costs of belief change contribute to this effect, and it would be worthwhile empirically investigating how such considerations can contribute to and explain belief inertia.

Social Costs

Human beliefs serve not only as internal models of the world but also as social signals and commitments. In active inference and variational learning frameworks, agents update beliefs to minimise surprise or prediction error, yet these updates occur in a social context where beliefs fulfil both epistemic (truth-seeking) and social-coordination functions [377, 396–401]. Believing (or

disbelieving) certain propositions can grant individuals emotional comfort or group acceptance, independent of the belief's accuracy [386]. This dual role means that an agent's posterior after observing new evidence is not determined by epistemic considerations alone, but also by the expected social and personal utilities associated with holding particular beliefs [381, 386, 402, 403]. Consequently, standard Bayesian updates, which are focused purely on data and prior likelihood, are often tempered by an additional motive: to align with valued identities and norms that confer utility on the belief state [377, 397, 399, 403]. This insight echoes the idea that *all thinking is "wishful" thinking* to some extent, with motivational imperatives modulating inferential processes [384]. The free energy minimised during belief updating thus implicitly includes not just accuracy-related (surprisal) terms, but also pragmatic terms capturing the work required to overcome cognitive inertia and social repercussions [398, 404].

Changing one's mind can threaten group affiliations and invite real or perceived social sanctions (e.g., loss of status, trust, or membership) [396]. Beliefs often function as markers of group identity, so revising a key belief may signal disloyalty or value misalignment, incurring social costs like ostracism or ridicule. Anticipation of such costs creates a strong deterrent to belief revision, especially for identity-linked beliefs maintained by tight-knit communities and normative expectations [377, 398]. Indeed, social norms enforce a kind of epistemic conformity: individuals internalise the expectation that they "ought" to hold certain beliefs to remain in good standing [396, 403]. From a decision-analytic perspective, the utility of a belief therefore includes not only its truth-tracking benefits, but also its social payoff. A false or unfounded belief might persist if it brings social acceptance or emotional relief, whereas a truthful belief might be resisted if it carries stigma or existential dread. Accordingly, belief change in social contexts resembles a form of motivated reasoning: agents are inclined to arrive at the conclusions they want (or need) to reach, as long as they can justify them to themselves and others [381]. Here, "wants" are not arbitrary whims, but structured by social identity and normative pressures—what one wants to believe is often what one's group wants one to believe. An agent will unconsciously search for justifications to retain beliefs that serve valued social goals (e.g. solidarity, consistency, pride) and discount evidence that threatens those goals. The free energy landscape is warped by social potential energy. Certain directions of belief change appear steep (costly) due to the interpersonal consequences associated with them.

Empirical research supports these principles. For instance, people consistently overestimate the severity of the social sanctions they will face for changing a politically charged belief, leading to excessive self-censorship and public conformity [405]. In one set of their studies, U.S. partisans expected far more backlash from their in-group if they voiced a dissenting opinion than what actually materialised, with an average overestimation effect size of $d \approx 0.87$. These inflated expectations of ostracism or punishment (sometimes stemming from an egocentric bias in social perspective-taking) make belief revision seem riskier than it truly is. Accordingly, individuals often stick to publicly defending their prior attitudes, even when privately grappling with contrary evidence. Social psychologists refer to this pattern as identity-protective cognition, wherein reasoning processes bend to protect the agent from the social identity costs of admitting error. The effect can become self-reinforcing. If everyone fears speaking up or changing their mind, the apparent unanimity of belief within the group remains unchallenged, further raising the perceived cost of dissent. Yet research also shows that these perceived social costs are malleable. Prompting individuals to reflect on their past loyalty and contributions to the group can reassure them that a change of mind will not irrevocably brand them as "disloyal," thereby significantly reducing their concern about sanction and encouraging more open expression of revised beliefs.

Beliefs are multi-functional cognitive tools that balance accuracy, utility, and inertia. They must at once represent the world (epistemic accuracy), support our emotional needs and moral values, and coordinate with our social milieu (utility), all while minimising drastic revisions that incur cognitive and social "work" (inertia). This perspective prepares us to interpret classic phenomena—confirmation bias, selective exposure to information, and attitude polarisation, not as inexplicable failures of rationality, but as strategic trade-offs given the agent's objectives. An agent facing high costs for belief change will rationally exhibit a kind of conservatism. Agents will favour information that confirms existing beliefs and avoids provoking costly updates. Indeed, a confirmation bias in information-seeking can be seen as an adaptive strategy to preserve high-utility beliefs by selectively attending to congruent evidence and filtering out challenges. Experimental studies of selective exposure document that people spend more time with news and arguments that align with their preexisting attitudes than with those that contradict them, even when source credibility is controlled [406]. By skimming "friendly" evidence, individuals reduce the likelihood of encountering data that would demand painful social readjustments or internal value conflicts. Similarly, communities may become

polarised when each side's beliefs carry their own social rewards—members of opposing groups double down on group-consistent narratives, bolstering internal cohesion at the expense of cross-group accuracy. Over time, this self-reinforcing selection and interpretation of evidence drives group attitudes further apart, as each group lives in a bubble where maintaining their version of reality is pragmatically advantageous [398]. The following sections will explore how confirmation bias in evidence appraisal, asymmetrical information search, and polarisation dynamics emerge naturally once we acknowledge that changing one's mind is not "free." It incurs variational costs, paid in both cognitive effort and social capital, which a resource-rational mind navigates by carefully weighing when belief change is truly worth the price.

Confirmation Bias

Confirmation bias manifests through selective attention mechanisms, as shown in recent experimental work [407, 408]. Westerwick et al. demonstrated that when selecting political information online, participants spent more time with content matching their existing views regardless of source quality [406]. This bias emerged from participants' choices rather than the content itself. Building on this, Talluri et al. revealed that making a categorical choice selectively enhanced sensitivity to subsequent evidence consistent with that choice, similar to attentional cueing effects [408]. Prat-Ortega and de la Rocha proposed a neural mechanism for this bias [407], suggesting that choices direct feature-based attention to amplify processing of choice-consistent sensory evidence while suppressing inconsistent information. Together, these findings indicate confirmation bias operates through early attentional selection rather than just later-stage decision processes.

Motivated Reasoning

Confirmation bias is a type of motivated reasoning, a process where information processing is biased towards achieving desired outcomes rather than accuracy alone [381]. Motivations can be accuracy-driven, encouraging unbiased reasoning, or directional, prompting strategies that reinforce existing beliefs, identity, or preferred conclusions. However, motivated reasoning remains constrained by plausibility; people select cognitive processes, such as memory retrieval and

interpretation, that justify favoured conclusions rather than inventing implausible beliefs [409–411].

Individuals revise beliefs asymmetrically, giving more weight to confirmatory or emotionally favourable evidence than to equally informative negative evidence [412]. Motivated reasoners also selectively trust or avoid sources based on alignment with their views, effectively assigning lower reliability to disconfirming information. This acts like a biased Bayesian filter, reducing the impact of contradictory evidence on belief updating [413].

Consequently, shared evidence can polarise rather than unify groups with opposing and even similar priors. When interpreting balanced evidence through a biased lens, individuals' initial beliefs often become more extreme, exacerbating attitude polarisation [398, 414].

Biased Reasoning

Two recent frameworks have been proposed to explain some of the biases that occur in reasoning: coherence-based reasoning (CBR) [415] and belief-consistent information-processing (BCIP) [416]. Coherence-based reasoning posits that individuals strive to maintain a consistent and interconnected set of beliefs, minimising cognitive dissonance. More specifically, in CBR, a constraint-satisfaction network settles into an attractor by bidirectionally reshaping both beliefs and incoming information to maximise overall coherence. Crucially, strongly activated priors are harder to dislodge, so they often anchor the attractor state. Similarly, belief-consistent information processing describes the tendency to favour information that aligns with pre-existing beliefs, a process that is less cognitively demanding than evaluating and integrating contradictory evidence. BCIP is a special case of CBR's coherence construction under conditions of dominant priors [417].

These two frameworks can be reconciled with our account through the lens of cognitive economy. Our variational cost framework formalises this principle by suggesting that altering one's beliefs incurs a cognitive cost, quantified by the KL divergence, which measures the informational distance between prior and posterior beliefs. Moreover, the weight of other firmly held beliefs can be explained by the presence of costs associated with revising more strongly held beliefs. For example, the expected cost associated with modifying a fundamental belief such as "I make correct assessments of the world" would be increased

levels of doubt about the reliability of one's assessments, potentially leading to greater general levels of uncertainty and the accompanying negative affect that often arises.

In this sense, both coherence-based reasoning and belief-consistent information processing can be viewed as cognitive strategies that minimise the costs that we describe. By maintaining coherence and selectively processing information, individuals reduce the "informational work" required to update their mental models of the world, thereby avoiding the significant pragmatic and cognitive expenditures associated with belief revision. In essence, these frameworks highlight different facets of the same underlying drive to manage cognitive resources efficiently, where the perceived utility of a belief is weighed against the inherent costs of mental reorganisation.

8.2 A Motivated Variational Belief Change Model

As a starting point, we take inspiration from *variational inference*, which underpins the mathematical formalism of the Free Energy Principle and Active Inference [65, 418]. In the standard Bayesian paradigm, the goal is to infer a posterior belief $p(s \mid o)$ about the state of the world $s \in S$ given an observation o, where S is a finite set of latent states. According to Bayes' rule, finding this posterior requires one to compute the model evidence or marginal likelihood $p(o)$, which is an intractable problem in general. Variational inference aims to reformulate this problem by recasting it as an *optimisation problem* over a variational family $\mathcal{Q}$ of probability distributions. The objective function of this optimisation problem is the negative *evidence lower bound* (ELBO) or *variational free energy* (VFE) [351, 419], and is given by

$$F[q(s), o] = -\underbrace{\mathbb{E}_{q(s)}[\log p(o \mid s)]}_{\text{Accuracy}} + \underbrace{\mathrm{D}_{\mathrm{KL}}\left[q(s) \,||\, p(s)\right]}_{\text{Complexity}} \tag{8.1}$$

$$= \underbrace{-\mathbb{E}_{q(s)}[\log p(s, o)]}_{\text{Energy}} - \underbrace{H[q(s)]}_{\text{Entropy}}. \tag{8.2}$$

The decomposition in Equation 8.1 highlights a key tension between the expected log likelihood of the observation (accuracy) and the KL divergence from the prior to the variational posterior (complexity). In particular, this complexity

acts as a *regulariser* on the agent's posterior beliefs, penalising models that differ more from the prior.

Active Inference (AIF) extends the free energy principle (FEP) to recognise the role of the *actions* that agentic systems can perform to influence the environment and, vicariously, their observations [70, 319, 356]. Cast here in the variational perspective, the objective function that is posited to drive decision making in AIF is the *expected free energy* (EFE), which is a functional of a policy (sequence of actions) π, and is given by

$$G(\pi) = -\underbrace{\mathbb{E}_{q(s,o|\pi)}\left[\mathrm{D}_{\mathrm{KL}}\left[q(s \mid o, \pi) \mid\mid q(s \mid \pi)\right]\right]}_{\text{Epistemic value}} - \underbrace{\mathbb{E}_{q(o|\pi)}[\log \tilde{p}(o)]}_{\text{Pragmatic value}}, \tag{8.3}$$

where $\tilde{p}(o)$ is a probability distribution representing the agent's preferences over their own observations, commonly known as a *prior preference* [70]. Here, we propose to extend this picture by considering the implications of assuming that agents have *preferences about their own beliefs*, and develop a mathematical framework for describing the ensuing implications for belief updating. In other words, we suggest that the C matrix, which is used in the active inference literature to parameterise the preference prior [70, 319, 373], can be extended to be defined over the agent's own beliefs as well, rather than only observations/states.

For the purposes of this chapter, we focus on the mechanisms of belief change that drive agents. In particular, we are interested in the mapping from sensory states to belief states. We investigate the consequences of assuming that this mapping comprises two key components. The first component is a preference satisfaction component, here represented by an "expected utility" term, which can be related to prior preferences by a softmax transformation [420]. The second component is a direct cost for belief updating, which is quantified by the KL divergence from the agent's prior beliefs to their posterior beliefs.

We will further assume that agents' preferences are grounded only in internal paths (of observations and beliefs), and not directly in external paths (of states), which is in concordance with an affect-driven view on motivation [421, 422]. Under these assumptions, an agent's preferences can be described mathematically by a *utility functional* $\mathcal{U} : \mathcal{Q} \times \mathcal{O} \to \mathbb{R}$ defined over observations and *beliefs*, but not external states.

Given this, our core proposal is to model an agent's belief updating processes as a variational optimisation process that maximises the following functional of beliefs and observations, which we call the *Motivated Variational Free*

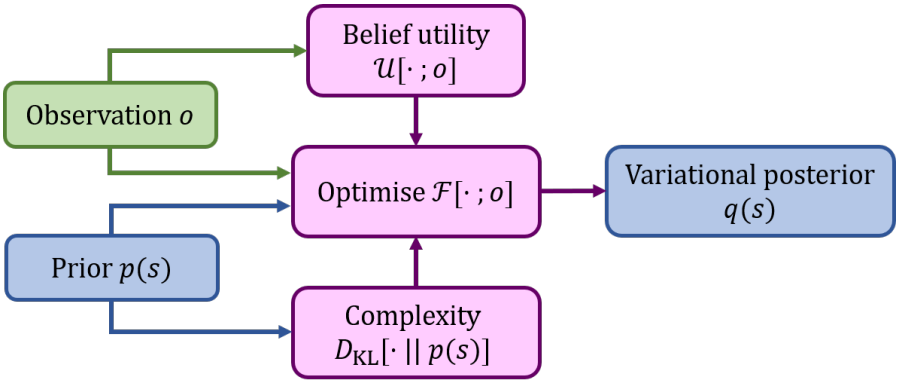

Figure 8.1: Schematic depicting the components involved in our proposed model of belief updating. Prior beliefs and observations act as inputs to a process that optimises a balance between a belief utility functional and a complexity term, measuring the KL divergence of the posterior relative to the prior beliefs.

Energy (MVFE):

$$\mathcal{F}[q(s), o] = \underbrace{\mathcal{U}[q(s), o]}_{\text{belief utility}} - \lambda \underbrace{\mathrm{D}_{\mathrm{KL}}\left[q(s) \,||\, p(s)\right]}_{\text{complexity}}, \tag{8.4}$$

where $\lambda > 0$ is a parameter determining the relative strength of the cost of belief updating to the belief utility term. The belief utility term may or may not depend on the accuracy of the model, and depending on its form, can give rise to different belief updating behaviours. Under this model, taking $\lambda \to 0$ induces a belief update that is purely driven by belief utility, which is akin to assuming that the agent is able to instantaneously and effortlessly convince themselves of whatever they wish to believe. On the other hand, taking $\lambda \to \infty$ increases the cost of updating to the point that the agent is no longer able to change their mind, under any circumstances.

Importantly, one particular form for the belief utility that we investigate is a linear combination of what we term an *affective utility* and a weighted expected log-likelihood or *accuracy* term, which takes the following form:

$$\mathcal{U}[q(s), o] = \underbrace{U[q(s), o]}_{\text{affective utility}} + \alpha \underbrace{\mathbb{E}_{q(s)}[\log p(o|s)]}_{\text{accuracy}}, \tag{8.5}$$

where $\alpha \geq 0$ is a *likelihood weighting* parameter, which determines the extent to which the agent's final belief distribution explains the data it has observed. A higher value of α can be interpreted as a stronger desire to arrive at beliefs that explain the observed data well. Moreover, for constant affective utility functions and $\alpha = \lambda = 1$, we recover the VFE as a special case of "accuracy-motivated" belief updating. In the following section, we study the predictions made by adopting the belief utility functional given in Equation 8.5.

8.3 Optimal Belief Updates with Linear Affective Utility Case

As an illustrative example that will be used in the computational experiments described in the following section, we derive the closed-form optimal posterior that maximises the MVFE introduced in Section 8.2 for the case of linear affective utilities.

Under this form, we have

$$U[q(s), o] \;=\; \sum_{s \in S} c_s \, q(s), \tag{8.6}$$

with coefficients $c_s \in \mathbb{R}$ capturing the valence of believing state s.

We maximise the MVFE in Equation 8.4 under the normalisation constraint $\sum_s q(s) \;=\; 1$. Introducing a Lagrange multiplier $\eta \in \mathbb{R}$ gives the augmented Lagrangian

$$\mathcal{L}(q, o, \eta) = \sum_{s \in S} c_s \, q(s) + \alpha \, \mathbb{E}_q[\log p(o \mid s)] - \lambda \, \mathrm{D}_{\mathrm{KL}} \left[q(s) \,||\, p(s) \right] + \eta \Big(1 - \sum_s q(s) \Big). \tag{8.7}$$

Stationarity with respect to each $q(s)$ yields

$$0 = \frac{\partial \mathcal{L}}{\partial q(s)} = c_s \;+\; \alpha \, \log p(o \mid s) \;-\; \lambda \left[1 + \log \frac{q(s)}{p(s)} \right] \;-\; \eta. \tag{8.8}$$

Solving (8.8) for $q(s)$ and exponentiating, we obtain

$$q(s) = p(s) \exp\left[\frac{1}{\lambda} \Big(c_s + \alpha \, \log p(o \mid s) - \eta \Big) - 1 \right] \tag{8.9}$$

$$\propto p(s) \exp\left[\tfrac{1}{\lambda} \Big(c_s + \alpha \, \log p(o \mid s) \Big) \right]. \tag{8.10}$$

Normalising with the partition function

$$Z(o) \;:=\; \sum_{s' \in S} p(s') \, \exp\left[\tfrac{1}{\lambda} \Big(c_{s'} + \alpha \, \log p(o \mid s') \Big) \right], \tag{8.11}$$

we arrive at the optimal variational posterior

$$q^\star(s) = \frac{p(s) \exp\left[\lambda^{-1} \Big(c_s + \alpha \, \log p(o \mid s) \Big) \right]}{Z(o)}. \tag{8.12}$$

In what follows, we use the convenient closed form solution for the optimal posterior in Equation 8.12 to conduct a series of illustrative simulations and experiments.

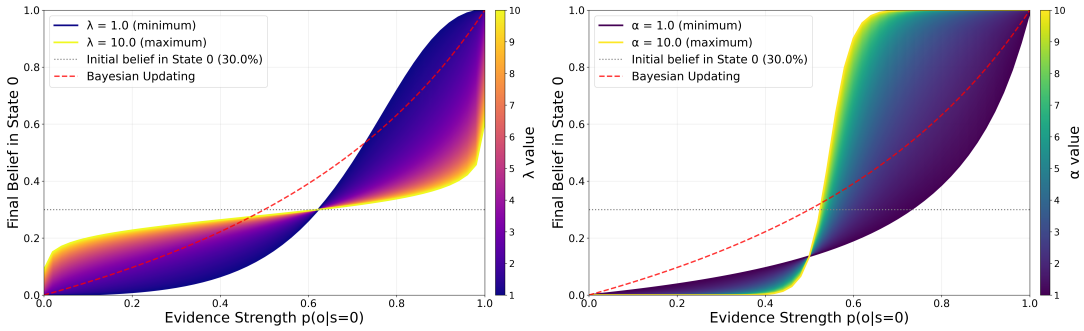

Figure 8.2: Plots depicting how agents with different conservatism parameters λ and likelihood weight parameters α respond to evidence that confirms or contradicts their belief preferences to varying degrees. Left: Final belief of the probability $q(s = 0)$ of state 0 occurring as the evidence strength (in the form of a likelihood $p(o|s = 0)$) varies from 0 to 1 for different values of λ. Right: Final belief $q(s = 0)$ as evidence strength varies for different values of α.

8.4 Experiments and Results

To study the implications of our proposed model on how motivated agents update their beliefs, we conducted a series of minimal experiments using categorical distributions.[1] In all simulations, we consider how a single piece of evidence presented in the form of a likelihood distribution may be selected and subsequently influence the belief updating process of a single agent, who we assume is endowed with a linear affective utility function and forms beliefs by optimising the MVFE given in Equation 8.4. In particular, we show that under our model, some important features of human belief-updating are qualitatively recovered. Moreover, our model can be used as a framework to generate testable predictions about and simulations of human behaviour in different scenarios.

Reactions to Good vs. Bad News

How do different agents react to differing degrees of good vs bad news? In our first set of experiments, we aim to understand and illustrate the effects of varying the strength of evidence, the conservatism parameter λ and the likelihood weight parameter α on belief updating. In this scenario, an agent begins with an initial prior over the outcomes of a Bernoulli random variable (i.e., a biased coin flip) specified by $p(s = 0) = 0.3$, and receives evidence of varying strengths

[1]The code for generating the experimental results can be found at `https://github.com/dkhyland/motivated-variational-belief-updating`.

in the form of a likelihood $p(o \mid s)$. For Bernoulli random variables, we will assume that the hidden state s may take the values 0 or 1. The agent updates their beliefs to maximise the objective in Equation 8.4 under the belief utility given in Equation 8.5.

In Figure 8.2, we plot the final belief in state 0 as we vary the evidence strength $p(o \mid s = 0)$ between 0 and 1 along the x axis and the values of λ (left) and α (right) as a spectrum for $\lambda \in [1, 10]$ and $\alpha \in [1, 10]$, along with the Bayesian update. From the left plot, we observe that higher values of λ lead to updates that are closer to the prior, whereas lower values of λ lead to updates that are more sensitive to the affective utility. From the right plot, the opposite effect is observed – higher likelihood weights lead to more sensitivity to the evidence, and lower likelihood weights increase sensitivity to the affective utility.

Evidence Selection

How do the relative strengths of belief utility and conservatism affect the selection of evidence? In this experiment, we demonstrate the presence of a form of confirmation bias in our model, and seek to understand how different components of the model affect the selection of evidence in our motivated agent. In particular, recall that a crucial tenet within active inference is that agents are active sense-makers, selecting evidence to resolve uncertainty in both specific and non-specific manners in order to develop a better model of the world and achieve their ultimate objectives. In this experiment, we extend this notion to include the motivated selection of evidence to either confirm or disconfirm an agent's preferences.

We consider what happens when an agent with a linear affective utility who prefers to believe that $q(s = 0) = 1$ is presented with two pieces of evidence. We study two different scenarios for what these pieces of evidence may be. In Scenario 1, the first piece of evidence (Evidence A) is *'confirmatory'*, in the sense that it provides evidence for the agent's desired belief, but is further (induces updates with a larger KL divergence) from the agent's prior compared to the second piece of evidence (Evidence B). Evidence B is *'contradictory'*, in the sense that it is evidence against the agent's desired belief but is closer to the agent's prior. In Scenario 2, Evidence A has both a higher affective utility and induces updates with a lower KL from the prior to the posterior. In Figure 8.3, the objective value is plotted as we sweep across values of $\lambda \in [0.1, 100]$. In Scenario 1, we observe a threshold where the agent switches from selecting the confirmatory evidence

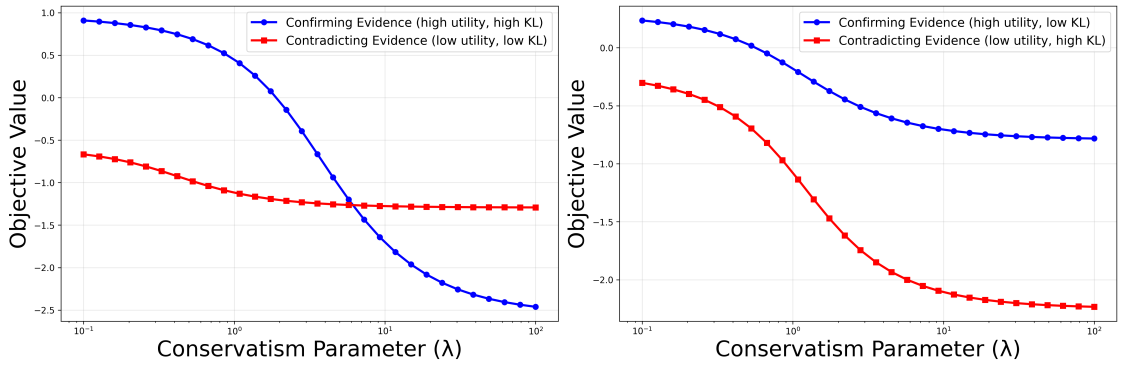

Figure 8.3: Plots of the objective value for Scenarios 1 and 2 with different combinations of evidence. In both scenarios, we fix $\alpha = 2.0$ and use a linear affective utility functional with $U[q(s), o] = q(s = 0)$. Left: Scenario 1, where Evidence A has high utility but high KL from the prior and Evidence B has low utility but low KL from the prior. Right: Scenario 2, where Evidence A has high utility and low KL, whereas Evidence B has low utility and high KL.

to selecting the contradictory evidence, whereas this does not occur in Scenario 2. Intuitively, this is because for low values of λ, the utility term dominates belief updating, but for higher values of λ, the cognitive cost term dominates. In contrast, when both the utility component is higher and cognitive costs are lower for one piece of evidence over the other, there is never a reason for the agent to choose to observe disconfirmatory evidence. In Figure 8.4, we plot various heatmaps depicting the objective values, their differences, and the final beliefs of the agent in Scenario 1 of this experiment. In the bottom two plots, we observe how the boundary at which the agent goes from selecting Evidence A to Evidence B varies as the combination of the two parameters varies.

Attitude Polarisation

How do belief conservatism and likelihood weighting affect attitude polarisation? In our final experiment, we simulated a basic attitude polarisation scenario, where two agents, Agents 1 and 2, began with the same prior belief about the probability $q(s = 0)$, and observed the same evidence in the form of a likelihood. Both agents were endowed with linear affective utility functions, but Agent 1 had a preference for believing that $q(s = 0) = 1$ and Agent 2 had a preference for believing that $q(s = 0) = 0$. Plotting the final beliefs after updating according to our model as we varied λ and α independently from 0 to 10 in Figure 8.5, we observe that for low values of both parameters, the two agents' final beliefs differed significantly, demonstrating a basic form of attitude polarisation. However, as we increased

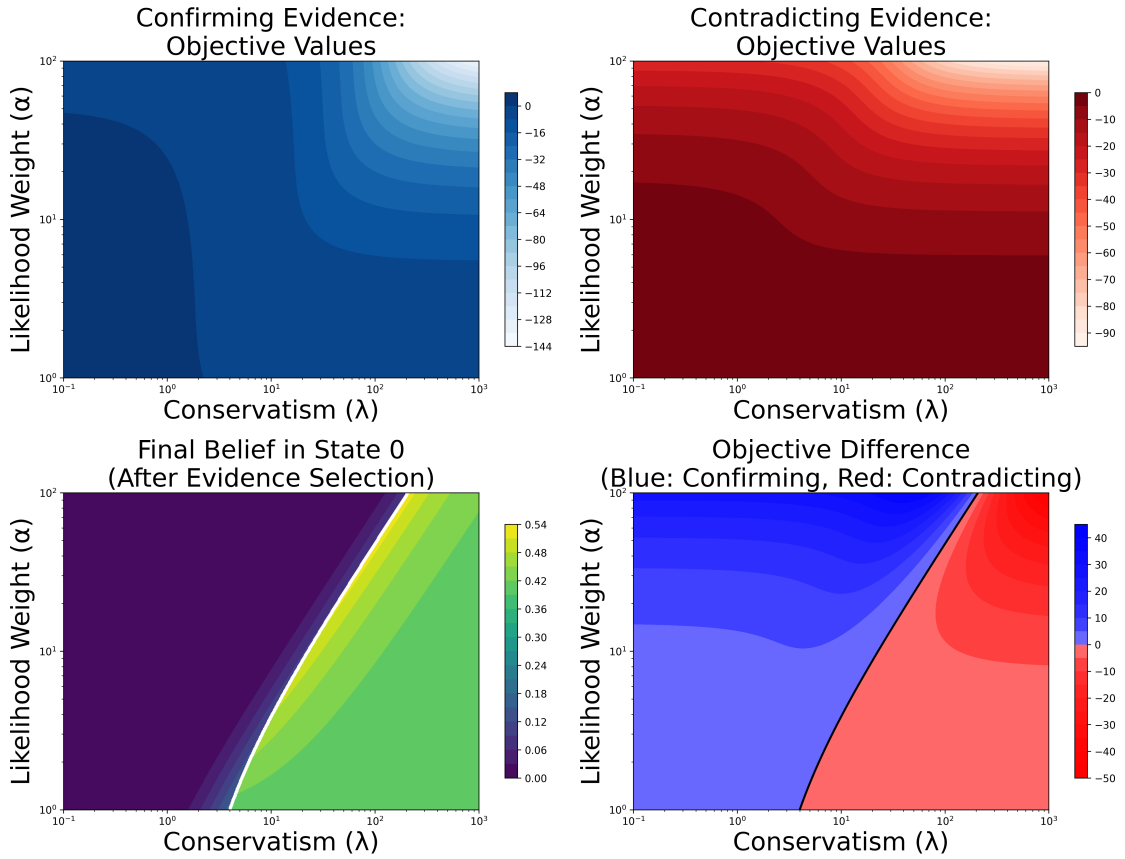

Figure 8.4: Contoured heatmaps depicting various quantities as we vary both the conservatism parameter λ and the likelihood weight parameter α in Scenario 1 of our second experiment. Top left: heatmap of the objective landscape under confirmatory evidence, i.e., evidence that aligns with the agent's desired belief. Top right: heatmap of the objective landscape under contradictory evidence, i.e., evidence that contradicts the agent's desired belief. Bottom left: heatmap of the final belief $q(s = 0)$ after evidence selection. The white line depicts the boundary at which the agent selects Evidence A (left of boundary) over Evidence B (right of boundary). Bottom right: heatmap showing the difference in the objectives for confirmatory and contradictory evidence. The black line again depicts the boundary between choosing Evidence A over Evidence B.

the two parameters, the agents' final beliefs converged towards similar (though not necessarily Bayes rational) beliefs.

8.5 Discussion

Though much work needs to be done in empirically validating instantiations of the framework, our findings demonstrating the presence of different kinds of 'biases' that are observed in human belief updating lend plausibility to the

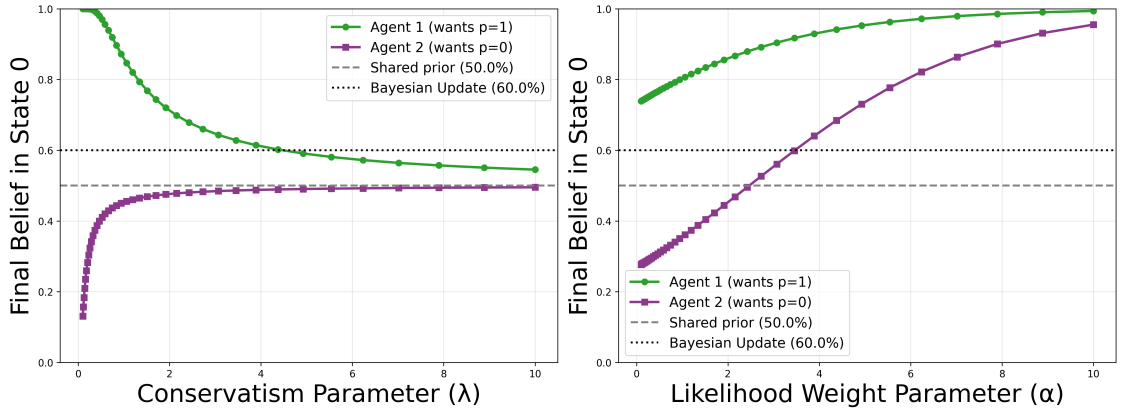

Figure 8.5: Plots of attitude polarisation effects between two agents who begin with the same prior beliefs and observe the same evidence, but have different affective utilities. Agent 1 linearly prefers to believe that $q(s = 0) = 1$, whereas Agent 2 linearly prefers to believe that $q(s = 0) = 0$. Left: Final beliefs as we vary λ from 0 to 10. Right: Final beliefs as we vary α from 0 to 10.

idea that realistic belief updating is subject to significant inertia and bias, driven largely by the constraints imposed on human agents by both internal cognitive limitations and external structures.

Realistic belief revision is rarely drastic, especially in the presence of cognitive costs for updating. From a cognitive standpoint, rapid updates necessitate more complex neural rewiring and a higher cognitive load, which can overwhelm limited cognitive resources. Agents would tend to avoid such costly leaps. Incremental steps across belief space reduce immediate costs but may also cumulatively result in lower total energetic and informational expenditure.

Under this view, we hypothesise that gradual transitions are typically more sustainable and preferable overall. Such considerations could explain why individuals are naturally inclined to resist abrupt changes in their belief systems despite potentially strong contradictory evidence, reinforcing conservative patterns of information integration.

Strategies for effective belief updating: Our basic model suggests several strategic insights for promoting more effective belief updating. Given the high cost of large leaps in belief space, strategies should prioritise incrementalism. This involves structuring information exposure in manageable segments that progressively lead individuals towards desired beliefs, thereby reducing the energetic, cognitive, and social resistance to dramatic changes. Social networks should be leveraged strategically: encouraging cross-cutting social ties and

diversity in informational environments can reduce the perceived social risks associated with belief change.

Completing the Action-Perception Loop: So far, our model has focused on the processes involved in the updating of beliefs, taking into account sensory evidence. Our preliminary data selection investigation takes this a step further by demonstrating how decisions about what data to observe can influence decision making. However, further work is required to fully integrate motivation into the perception-action loop.

Addition of temporal considerations: Our model has not explicitly incorporated the temporal aspect of belief updating, which we believe to be significant in modelling the various costs that must be taken into consideration. Indeed, several works have posited a central role of *rates of change* in free energy/prediction errors as crucial to understanding affect [423, 424]. Extending the model to account for the role of time would allow a more detailed analysis of how belief trajectories could be optimised, rather than single updates.

Extensions to group dynamics: Further work could more explicitly incorporate group dynamics, particularly focusing on how social networks influence belief inertia and revision costs. Future work could explore the degree to which group identity and perceived social costs shape belief stability, potentially replicating frameworks similar to those presented by [398] on epistemic communities. By examining how belief updates propagate through structured networks and assessing how identity-protective reasoning reinforces certain belief states, we can quantify the inertia inherent within closely knit communities. Moreover, evaluating the relative weight of belief confidence levels and their susceptibility to drift could provide deeper insights into the dynamics of belief evolution in social contexts. Such extensions may also clarify how networked beliefs reinforce each other, creating feedback loops that stabilise misinformation.

Applications to AI Systems: A particularly compelling future direction involves applying insights from our model to the design and management of belief dynamics within artificial intelligence systems. For instance, we could attempt to manipulate environmental stimuli, to target "key nodes" in a belief network. In doing so, we could organically influence AI system beliefs without direct internal

adjustments. Potential applications include bias mitigation, where systematic exposure to countervailing evidence could reduce entrenched algorithmic biases, and adaptive learning environments, where carefully structured informational inputs guide AI towards desirable belief states incrementally and sustainably. Testing this concept could involve using reinforcement learning frameworks, where AI agents are systematically exposed to specific environmental stimuli designed to shift belief nodes indirectly, measuring the resultant changes in belief states and decision making behaviours.

Further empirical validation: Empirical validation remains essential for confirming and refining our theoretical propositions. Future empirical work needs to be done to rigorously test model predictions using controlled laboratory experiments, field studies, and simulation analyses. For instance, quantifiable predictions derived from our framework – such as the relationship between KL divergence, belief revision speed, and associated cognitive or social costs – could be tested experimentally by monitoring physiological or neural responses during belief updating tasks. Longitudinal field studies examining belief trajectories within real-world social groups could also provide valuable validation, providing insights into how incremental versus rapid belief changes correlate with tangible social and cognitive outcomes. Finally, computational simulations, informed by empirical data, could further refine the model's predictive capabilities, enabling robust testing of interventions aimed at fostering healthy belief updating processes.

9

Attention and Influence

If you could know one thing about an animal to best predict its behavior, I suggest that attentional state would be the wisest choice. If you could know what the animal is attending to, how much of its attention is allocated to the item, and what the consequences of attention are, then you would gain considerable ability to predict the animal's behavior.

Michael S. Graziano, *Consciousness and the Social Brain* (p. 61) [425]

In this chapter, we present another perspective on influencing an agent that complements standard approaches of modifying their preferences or restructuring their beliefs. This is motivated by the observation that a core challenge for real-world agents is determining which aspects of a decision problem are most relevant – that is, deciding what to pay attention to. Attention profoundly influences human decision making, often leading to choices that diverge from normative rationality, and is a powerful cognitive tool for dealing with the rich, multifaceted nature of decision making in the real world. In this chapter, we develop a formal model to capture how attention modulates preferences and drives preference change. We begin by proving a representation theorem characterising preference relations shaped by salience-weighted features. Building on this foundation, we consider a principal capable of strategically influencing an agent's preferences – not by altering the agent's intrinsic values, but by adjusting the salience of

different attributes associated with the available alternatives. We study the boundaries of salience-driven preference change and present a polynomial time algorithm to determine whether a given alternative can be made the agent's most preferred option. Expanding on this theory, we demonstrate a Bayesian model of how a principal can learn about and interact effectively with an agent in our model to promote a target alternative. This framework provides a mathematical and algorithmic starting point for designing systems that can effectively guide human decision making in complex, multi-faceted environments. At the same time, we highlight important considerations about the responsible design of influence strategies, particularly in contexts where subtle preference shaping may have unintended or harmful consequences.

Understanding and persuading other agents is a fundamental challenge that social beings face. To address this challenge, it is helpful to understand how an agent's preferences may change. As a starting point, Jeffrey identifies two significant sources of preference change [426] – valuational and doxastic. Valuational preference change refers to a change in the underlying utilities that an agent associates with different outcomes, whereas doxastic preference change results from a change in the agent's beliefs about (the likelihood of) different outcomes [427]. Valuational preference change may occur by, e.g., acquiring a taste for certain kinds of food through repeated exposure, or enjoying a sport more as one becomes more proficient at it. Examples of doxastic preference change include not wanting to purchase a particular product after learning that it was unethically sourced, or not wanting to visit the beach after learning that it is raining. In this work, we explore an alternative source of preference change: that of *changing saliences*.

Researchers in several disciplines, ranging from neuroscience to economics, have appreciated the influence of salience on preferences. For example, Schonberg and Katz propose that attention-related regions of the brain contribute to non-reinforced preference change [428], i.e., changes in preference in the absence of external reinforcement. In his Nobel lecture, Daniel Kahneman asserted that "the effects of salience and anchoring play a central role in treatments of judgment and choice" [429]. Salience typically refers to the quality of being noticeable, significant, or prominent to an agent. It relates to how agents prioritise different information or features relative to each other, and therefore takes into account not just *individual* features of an agent's sensorium, but a whole *collection* of them. When making real-world decisions, agents often face multiple competing factors that must be balanced against each other. This raises a key question: how do

decision-makers prioritise different aspects of their options? For example, in a home-buying scenario, the prospective customer faces a richly multifaceted decision making problem, with features like price, size, facilities, neighbourhood, etc., all critically factoring into their preferences [430]. In this context, a real estate agent typically interacts with the buyer, pointing out different features of different homes. Depending on their incentives, the real estate agent may have their own preferences over which property is finally selected, and they may try to guide the decision-maker towards choosing these properties. Another ubiquitous instance of this is the formation of voting preferences for political parties/candidates, who are primarily characterised by where they stand on different issues, including taxation, foreign policy, and healthcare. Different political candidates aim to shape the public's perception of themselves by drawing the attention of their target audience to what are perceived to be their strengths on particular issues and the apparent weaknesses of other candidates, thereby elevating their relative status [431]. What is common to these scenarios is that (1) agents making a choice must find some way to combine preferences about each feature into an *overall* preference among alternatives, and (2) agents' preferences can be swayed even in the absence of new information or a change to their underlying values. We aim to develop a simple model that can capture these occurrences of preference change through the lens of salience.

As a third example that we will return to throughout this text, consider a decision many students face when enrolling in university – selecting their degree. Such a task is highly nontrivial, has significant implications for the student's later trajectory in life, and contains many different features that must be weighed against each other [432]. Among other things, factors contributing to the decision of the prospective student include the median salary of graduates from each discipline, the difficulty of the course, the course fees, the typical length of time to completion, etc. Naturally, the student may have preferences over each of these features, but when it comes to making a decision, they must combine these preferences into an overall ranking of the alternatives.

Contributions

In this chapter, we introduce a model to attempt to close the gap left by belief-mediated accounts of preference change by modelling the effect of *salience* on immediate preferences, where no new information is obtained by an agent and their underlying values do not change. Figure 9.1, adapted from [433,

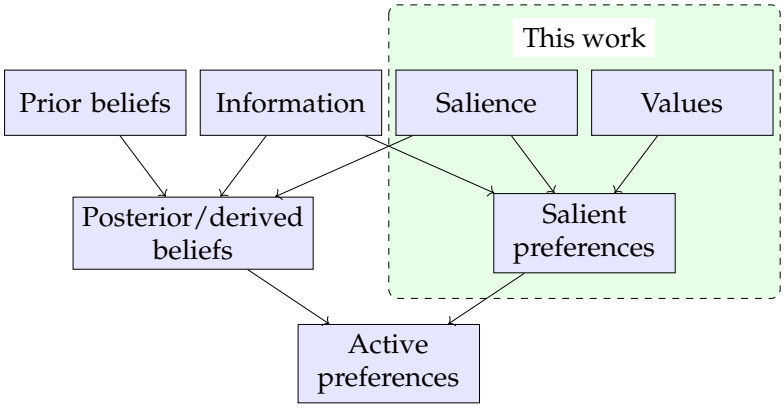

Figure 9.1: Basic schematic of the factors contributing to an agent's active preferences. As a first approximation, this can be split into a combination of beliefs, subjective values, and mechanisms including information and salience that interact with the two to form the agent's active preferences. Prior beliefs refer to the agent's beliefs coming into a particular decision making context. Information refers to new data that the agent receives from its external environment or other agents, and internally generated information, e.g., outputs of deliberation processes. Salience represents the weighting placed on different features of the alternatives under consideration (irrespective of their actual instantiations), which modulates the agent's immediate beliefs and preferences. Values refer to the underlying preferences an agent possesses concerning different instances of each feature separately. From these basic building blocks, posterior/derived beliefs are formed by updating prior beliefs with any previously unincorporated information, which are then modified by saliences. Similarly, salient preferences result from information and salience interacting with underlying values. Finally, these are combined to form active preferences, which guide the agent's decisions.

434], depicts the scope and context of our model. In particular, the proposed model studies decision making scenarios where an agent is faced with a set of alternatives, each uniquely characterised by a combination of features. The agent is assumed to assign intrinsic values independently to each feature, and salience is modelled as the weighting assigned to each feature (regardless of the particular instance) in the agent's overall utility function. To begin with, we introduce the formal model of the agent's preferences, and then prove a representation theorem for families of preference relations that are indexed by different salience weightings. Following this, we discuss the implications of the theory for preference change and discuss some implications that follow from modelling the influence of salience on preferences as we do. In particular, we show that there can be limitations to how much preferences can be distorted only through modifying the salience of different features to an agent.

The precise relationship between attention and preference stability is still poorly understood. While traditional preference models assume agents have

fixed, consistent preferences, empirical evidence shows that people's immediate preferences often shift based on what aspects of a decision they are currently attending to [435]. This raises a fundamental question: how can we reconcile stable underlying values with fluctuating attention-driven preferences?

We propose that this apparent contradiction can be resolved by modelling agents as having fixed feature-specific values and probabilistic beliefs about their attentional states. This approach aligns with two key psychological insights. Firstly, in the short term, people can maintain stable evaluations of individual attributes while still exhibiting varying overall preferences due to shifts in attention. Secondly, attention allocation itself is inherently probabilistic rather than deterministic [436]. The Dirichlet distribution provides a natural mathematical framework for representing these attentional dynamics, as it allows us to model both the uncertainty in how attention is allocated across features and how these allocations evolve with experience.

By explicitly modelling attention as probabilistic, we can better capture phenomena like preference reversals and context-dependent choice, without having to assume that the agent's underlying values are changing. This offers a more parsimonious explanation for many observed patterns in human decision making that have traditionally been difficult to reconcile with rational choice theory.

Following this, we then introduce a principal to the setting and study the restricted case where the principal may draw the agent's attention to different features of their decision making problem, potentially modifying the agent's preferences. Within this setting, we study two questions. Firstly, given an alternative, is there a salience distribution that makes it the most desirable? Secondly, given a target salience distribution, how should the principal interact with the agent to be confident that their salience distribution matches the target? We provide a polynomial time algorithm that answers the first question and demonstrate a basic framework that a principal may use to answer the second, with algorithms to address the problems of what questions the principal should ask and what features they should mention.

9.1 Related Work

Multi-Attribute Utility Theory

The representation of preferences using additive utility functions has been studied extensively going back to the foundations of utility theory [437–439]. These have found successful application in many areas of the literature, most notably in the management sciences and operations research fields [440, 441]. In particular, the functional form of our (weighted additive) utility representation has been discussed and analysed many times in the literature. However, to our knowledge, such representations have not been applied to model the concept of salience. Most work in the area of multi-attribute utility theory considers a single preference relation $\succeq$ over alternatives that an additive utility function can represent. What distinguishes our main result is that rather than considering a single preference relation, we study a *family* of preference relations $\{\succeq_s \colon s \in \mathcal{S}\}$ that are indexed by salience distributions, and derive a *single representation* for this entire family of preference relations. This is crucial in our context because it allows us to study *changes* in preference that arise from changing salience distributions, while leaving the underlying utility or value functions unchanged.

Salience in Decision Making

Our work shares conceptual similarities with that of [442, 443], who model salience as a distortion on the subjective probabilities assigned to different outcomes occurring. In our framework, we focus on modelling the direct influence of salience on preferences without requiring the interaction of beliefs. It remains unclear whether salience influences active preferences through modulating beliefs, preferences or some combination of both, although we hypothesise the latter to be more likely. Uncovering the precise connection between these properties of agents is thus an important open problem for further investigation.

Models of Preference Change

Models of preference change in the literature can be broadly divided into two categories of approaches: logic-based and quantitative. For a more detailed overview of the literature on the nature of preferences and preference change, we refer the reader to [434].

Logic-based approaches commonly represent preferences using logical sentences, and preference change is modelled as an update operator that modifies these sentences. Typically, some formal semantics is defined to capture a particular modality of preference change, e.g., preference addition, removal, or revision [444–449]. However, such approaches may run into issues of computational tractability and struggle to model *degrees* of preference.

On the other hand, quantitative approaches to modelling preference change draw inspiration from Bayesian models of belief revision [433, 450, 451] and multi-attribute utility theory [452, 453]. Our framework builds on top of that of [452], who also model a set of alternatives characterised by several features. Dietrich and List consider the case where features may be included or excluded from consideration in the agent's final preference, which is also a sum of value functions across different features under consideration. The authors prove a representation theorem for such preferences and expand the model to handle the agent's uncertainty over outcomes. A key feature of their model is that all 'motivationally salient' features are weighted equally. Our approach extends theirs by allowing different features to be weighted differently in the agent's final utility function, rather than only being allowed to include or exclude features from consideration.

The take-the-best heuristic is a simple method for choosing between alternatives based on the most important feature that discriminates them [454]. In our model, features with the highest salience will naturally contribute most to the agent's final decision, potentially giving rise to choices that are approximated by this heuristic. Indeed, one can view our model as an expansion of this heuristic to allow for multiple salient features of alternatives to be considered simultaneously.

9.2 Agent Model

Here, we develop a basic model that captures the influence of salience on preferences in the context of a multi-attribute decision making (MADM) problem [440, 441].

Feature Space and Alternatives: Consider an agent facing a choice among a set of *alternatives*, denoted $X = X_1 \times \cdots \times X_m$, which is a product space of m *feature spaces*, where each $X_j \subseteq \mathbb{R}$ is an interval. Each alternative $x = (x_1, \ldots, x_m)$ is characterised by a unique combination of *feature instances* $x_j \in X_j, j \in \{1, \ldots, m\}$.

Salience: Salience is modelled as the relative importance of different features to the current preferences of the agent. Here, we assume that salience operates at the level of features, as opposed to value functions, which apply to feature instances. In particular, we let $s = (s_1, \ldots, s_m)$ be a *salience distribution*, which is a categorical distribution over the m features, and we write $\mathcal{S} = \Delta^{m-1} = \left\{ s \in \mathbb{R}_+^m : \sum_{j=1}^m s_j = 1 \right\}$ for the set of all salience distributions (i.e., the standard $m-1$ simplex).

Preference Relations and Representation

To model the effect of salience on preferences, we augment the standard multi-attribute utility theory by introducing additional axioms. In particular, considering the set of all possible salience distributions, we assume the existence of a family of preference relations $\{\succeq_s : s \in \mathcal{S}\}$, indexed by these distributions. We call this object a set of *salient preference relations* over X. Let δ_j be the salience distribution that allocates a salience probability of 1 to feature X_j. We write $x \sim_s y$ iff $x \succeq_s y$ and $y \succeq_s x$, and $x \succ_s y$ iff $x \succeq_s y$ and not $y \succeq_s x$.Then, we say that alternative x *(Pareto-)dominates* y if $x \succeq_{\delta_j} y$ for all $j \in \{1, \ldots, m\}$. In other words, x is preferred to y when compared across all features separately. An alternative that is not dominated by any other alternative is called *undominated*.

We are now ready to introduce a set of axioms that establish the mathematical foundations for our model. These axioms provide necessary and sufficient conditions for representing these preferences through salience-weighted utility functions. As we will prove[1] in Theorem 9.1, a set of salient preference relations satisfies all these axioms iff it can be represented as a weighted sum of feature-specific utilities, where the weights correspond to salience values.[2]

1. Completeness: For any two alternatives $x, y \in X$ and any salience distribution $s \in \mathcal{S}$, it holds that $x \succeq_s y$ or $y \succeq_s x$.

2. Transitivity: For any salience distribution $s \in \mathcal{S}$, if $x \succeq_s y$ and $y \succeq_s z$, then $x \succeq_s z$.

[1] I am grateful to Tomáš Gavenčiak and Dalton Sakthivadivel for their help in discussing and working towards the proof of this theorem.

[2] We assume for simplicity that $m \geq 3$ for this result, with the special case of $m = 2$ omitted for brevity.

3. Alternative Continuity: For any salience distribution $s \in \mathcal{S}$ and any alternative $x \in X$, the sets

$$\{y \in X : x \succeq_s y\} \qquad \text{and} \qquad \{y \in X : y \succeq_s x\}$$

are closed in X.

4. Pure-Salience Isolation: For each $i \in \{1, \ldots, m\}$ and all $x, y \in X$,

$$x_i = y_i \Rightarrow x \sim_{\delta_i} y.$$

5. Affine Salience Dependence: Fix a reference alternative $x^0 \in X$. For each $s \in \mathcal{S}$, let U_s be a continuous representation of $\succeq_s$ (whose existence is guaranteed by Axioms 1–3 [455]), which is normalised by setting $U_s(x^0) = 0$. Assume that these normalised representations can be chosen so that for every $x, y \in X$, the function

$$D_{x,y}(s) := U_s(x) - U_s(y),$$

is affine on S, i.e., for all $s, t \in \mathcal{S}$ and all $\lambda \in [0, 1]$, we have

$$D_{x,y}(\lambda s + (1 - \lambda)t) = \lambda D_{x,y}(s) + (1 - \lambda) D_{x,y}(t).$$

6. Nontriviality: For each feature j, there exist $x_j, y_j \in X_j$ with $x_j \neq y_j$, some $x_{-j} \in X_{-j}$, and some $s \in \mathcal{S}$ with $s_j > 0$ such that

$$(x_j, x_{-j}) \succ_s (y_j, x_{-j}).$$

Theorem 9.1. *For a family $\{\succeq_s : s \in \mathcal{S}\}$, the following are equivalent:*

(i) Axioms 1–6 hold.

(ii) There exist continuous non-constant functions

$$v_j : X_j \to \mathbb{R} \qquad (j = 1, \ldots, m)$$

such that for all $x, y \in X$ and all $s \in \mathcal{S}$,

$$x \succeq_s y \iff \sum_{j=1}^{m} s_j v_j(x_j) \geq \sum_{j=1}^{m} s_j v_j(y_j).$$

Proof. We begin by showing the backward implication. Assume that there exist continuous non-constant functions $v_1, \ldots, v_m$ such that

$$x \succeq_s y \iff \sum_{j=1}^{m} s_j v_j(x_j) \geq \sum_{j=1}^{m} s_j v_j(y_j),$$

for all $x, y \in X$ and $s \in \mathcal{S}$.

Axioms 1 and 2 follow immediately from completeness and transitivity of the usual order on $\mathbb{R}$. For Axiom 3, fix a salience distribution $s \in \mathcal{S}$. The map $x \mapsto \sum_{j=1}^{m} s_j v_j(x_j)$ is continuous on X, so the sets of points $\{y \in X : x \succeq_s y\}$ and $\{y \in X : y \succeq_s x\}$ are closed. For Axiom 4, if $x_j = y_j$, then under $s = \delta_j$, we have

$$\sum_{i=1}^{m} (\delta_j)_i v_i(x_i) = v_j(x_j) = v_j(y_j) = \sum_{i=1}^{m} (\delta_j)_i v_i(y_i),$$

and hence $x \sim_{\delta_j} y$. For Axiom 5, fix a reference alternative $x^0 \in X$ and define

$$U_s(x) := \sum_{j=1}^{m} s_j \Big(v_j(x_j) - v_j(x_j^0) \Big).$$

Then, each U_s is a continuous additive representation of $\succeq_s$, and $U_s(x^0) = 0$. Moreover, for any two alternatives $x, y \in X$,

$$D_{x,y}(s) := U_s(x) - U_s(y) = \sum_{j=1}^{m} s_j \Big(v_j(x_j) - v_j(y_j) \Big),$$

which is affine in s. For Axiom 6, fix a feature j. Since v_j is non-constant, there exist $x_j, y_j \in X_j$ with $v_j(x_j) > v_j(y_j)$. For any $x_{-j} \in X_{-j}$, taking $s = \delta_j$ gives

$$(x_j, x_{-j}) \succ_{\delta_j} (y_j, x_{-j}),$$

so nontriviality holds.

We now prove the forward direction. Assume that Axioms 1–6 hold. By Axioms 1–3, for each $s \in \mathcal{S}$ there exists a continuous representation U_s of $\succeq_s$ by Debreu's famous representation theorem [438, 455], i.e, there exists a continuous real function $U_s : X \to \mathbb{R}$ such that for any two alternatives $x, y \in X$, $x \succeq_s y \Leftrightarrow U_s(x) \geq U_s(y)$. By Axiom 5, these may be normalised at $x^0 = (x_1^0, \ldots, x_m^0)$ so that $U_s(x^0) = 0$ for all s, and so that for every pair $x, y \in X$, the difference function

$$D_{x,y}(s) := U_s(x) - U_s(y)$$

is affine in s. For each $j \in \{1, \ldots, m\}$, define the valuation function

$$v_j(x_j) := U_{\delta_j}(x_j, x_{-j}^0).$$

Each v_j is continuous because U_{δ_j} is continuous, and $v_j(x_j^0) = U_{\delta_j}(x^0) = 0$. We claim that for every $x \in X$,

$$U_{\delta_j}(x) = v_j(x_j).$$

Indeed, if $x = (x_j, x_{-j})$, then Axiom 4 gives

$$x \sim_{\delta_j} (x_j, x_{-j}^0),$$

because these two alternatives agree on coordinate j. Since U_{δ_j} represents $\succeq_{\delta_j}$, indifferent alternatives receive the same utility value. Therefore, we have

$$U_{\delta_j}(x) = U_{\delta_j}(x_j, x_{-j}^0) = v_j(x_j).$$

Hence, for any $x, y \in X$, it holds that

$$D_{x,y}(\delta_j) = U_{\delta_j}(x) - U_{\delta_j}(y) = v_j(x_j) - v_j(y_j).$$

Now let us fix $x, y \in X$. Since $D_{x,y}$ is affine on S by Axiom 5, for every $s \in \mathcal{S}$ it holds that

$$D_{x,y}(s) = \sum_{j=1}^{m} s_j D_{x,y}(\delta_j).$$

Substituting the previous identity into the above equation yields

$$D_{x,y}(s) = \sum_{j=1}^{m} s_j \Big(v_j(x_j) - v_j(y_j)\Big).$$

Since U_s represents $\succeq_s$, we conclude that

$$x \succeq_s y \iff U_s(x) \geq U_s(y) \iff D_{x,y}(s) \geq 0 \iff \sum_{j=1}^{m} s_j v_j(x_j) \geq \sum_{j=1}^{m} s_j v_j(y_j).$$

It remains to show that each v_j is non-constant. For this, fix a feature j. By Axiom 6, there exist $x_j, y_j \in X_j$ with $x_j \neq y_j$, some $x_{-j} \in X_{-j}$, and some $s \in \mathcal{S}$ with $s_j > 0$ such that

$$(x_j, x_{-j}) \succ_s (y_j, x_{-j}).$$

Applying the representation that we have just proved to the pair $x = (x_j, x_{-j})$ and $y = (y_j, x_{-j})$, we obtain

$$0 < \sum_{j=1}^{m} s_j \Big(v_j(x_j) - v_j(y_j)\Big) = s_j \Big(v_j(x_j) - v_j(y_j)\Big),$$

because $x_i = y_i$ for all $i \neq j$. Since $s_j > 0$, it follows that $v_j(x_j) > v_j(y_j)$. Thus, v_j is non-constant. □

Feature j	CS (x^1)	Med (x^2)	$v_j(x_j^1)$	$v_j(x_j^2)$
Salary (\$)	100k	140k	50	80
Difficulty (/5)	4	5	20	10

Table 9.1: Feature instances for two alternatives in Example 9.1 and their associated values. In this example, the medicine degree offers a higher salary than the CS degree, but is also judged as more difficult to complete. Depending on which features the agent pays attention to, either option may be seen as more attractive than the other.

Example 9.1. Returning to the enrolment example with the formal model articulated, suppose for simplicity that the student's alternative space is characterised by two dimensions – expected salary and average difficulty. Now consider two alternatives available to the student – computer science (CS) and medicine, which we will label x^1 and x^2, respectively. Again simplifying considerably, suppose that the feature instances are given in Table 9.1. Under a uniform distribution, i.e., one that assigns equal weighting to all features ($s_1 = (0.5, 0.5)$), we see that medicine ranks slightly higher than CS with a total score of 45, compared to 35. However, under a salience distribution that places higher weight on the difficulty feature, e.g., $s_2 = (0.2, 0.8)$, we see that the salience-weighted utility of CS comes to 26, whereas medicine receives a score of only 24. This illustrates how preferences may reverse under different salience distributions, where we have $x^2 \succeq_{s_1} x^1$ and $x^1 \succeq_{s_2} x^2$.

Agent's Beliefs and Utility Function

To model an agent's ability to update their salience values, we take the interpretation that the agent maintains a *belief* about the possible salience distributions that they could hold, which is represented by a Dirichlet distribution over saliences:

$$s \sim p_A = \text{Dir}_m(\alpha). \tag{9.1}$$

This captures the idea that agents are often uncertain of their own preferences, including the relative weighting or importance of different features. When the agent makes a decision, a salience distribution s is sampled from the belief p_A, which is then used to calculate the agent's utility for different alternatives using the representation from Theorem 9.1. Given a sampled salience distribution s and the value functions, the agent's preferences are modelled using a *salience-weighted utility function* $u_A(x, s) := \sum_{j=1}^{m} s_j v_j(x_j)$.

9.3 Principal Model

So far, we have focused on developing a foundation for modelling the agent's preferences. Now, we introduce another agent – the principal – who interacts with the agent to achieve a specific goal. In this context, the principal could be an advisor, consultant, salesperson, advertiser, etc. Here, we study one of the most universal goals that a principal may wish to achieve: persuading the agent to choose a particular option.

Influencing the Agent

We take a computational perspective on the principal's task of persuading the agent to select a particular alternative. In particular, we break this problem into two steps. Firstly, given a chosen alternative that the principal wishes to promote, it may be helpful for them to know whether the alternative can be promoted to the most preferred alternative, purely by means of modifying saliences. In the case of our running university application example, the advisor may believe that CS is the better option for the student and thus wishes to persuade them to select this alternative. For the sake of simplicity, we focus on the setting where the principal knows the values $v_1, \ldots, v_m$ of the agent. There is already extensive literature on preference learning that could be applied to learn these values when this information is not readily available to the principal, so we focus our attention on the salience distributions. Given this, the first question the principal will want to ask is whether their task is feasible or not. In other words:

Is it always possible to make an agent choose a given alternative by directing their attention appropriately? Perhaps counterintuitively, it turns out that this is not true in general, and a simple example can demonstrate this.

Proposition 9.2. *It is not the case that for all undominated alternatives $x \in X$, there exists a salience distribution s such that $x \succeq_s y$ for all $y \in X$.*

Proof. Consider a simple example with three alternatives x^1, x^2, x^3 and two features. Suppose that the value functions are as follows: $v_1(x_1^1) = v_2(x_2^2) = 1, v_1(x_1^2) = v_2(x_2^1) = 2,$ and $v_1(x_1^3) = v_2(x_2^3) = 1.1$. Suppose for contradiction that there exists some salience distribution $s = (s_1, s_2)$ such that x^3 is the weakly best alternative. Then, we would have $1.1(s_1 + s_2) = 1.1 \geq s_1 + 2s_2$ and similarly $1.1 \geq 2s_1 + s_2$, which would then imply that $2.2 \geq 3s_1 + 3s_2 = 3$. □

This result illustrates that in this model, there are limitations to what a principal can achieve in trying to promote outcomes purely on the basis of directing the agent's attention. A natural question in light of this limitation, then, is *how* the principal can decide whether a given alternative may be promoted to the most desirable one by only influencing the agent's saliences. For the remainder of this section, we assume that the agent faces a *finite menu* of alternatives $X = \{x^1, \ldots, x^n\} \subset \prod_j X_j$, drawn from the product space. This ensures that the optimisation problems that follow have finitely many constraints. In particular, we formulate the principal's feasibility question as the following decision problem:

> SALIENCE-EXISTENCE:
> *Given*: Finite set of alternatives $X = \{x^1, \ldots, x^n\}$, valuation functions $v_1, \ldots, v_m$, choice $x \in X$.
> *Question*: Is there a salience distribution s such that for all $y \in X$, we have $\sum_{j=1}^m s_j v_j(x_j) \geq \sum_{j=1}^m s_j v_j(y_j)$?

It turns out that this problem can be solved in polynomial time using a linear feasibility program, where we find a vector $s = (s_1, \ldots, s_m)$ satisfying the following linear constraints:

$$\sum_{j=1}^{m} s_j \left(v_j(x_j) - v_j(y_j)\right) \geq 0, \ \forall y \in X,$$
$$\sum_{j=1}^{m} s_j = 1, \text{and } s_j \geq 0, \ \forall j \in \{1, \ldots, m\}.$$

With this, the principal is able to identify (1) whether or not their chosen alternative can be made the most desirable under some salience distribution, and (2) an instance of such a salience distribution, if it exists. Figure 9.2 summarises the setting discussed thus far. Now supposing that the task is feasible and that the principal has a target salience distribution, the next obvious question is *how* the principal can interact with the agent to guide their salience distribution towards the desired one. In order to address this, we first develop a model of how the principal and agent interact.

9.4 Principal-Agent Interactions

We assume that interactions between the principal and agent occur via the following modes:

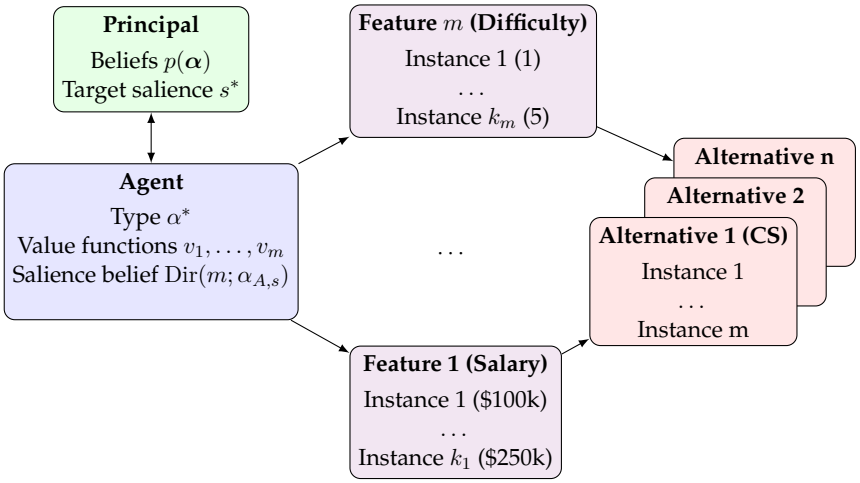

Figure 9.2: Illustration of the features, alternatives, and beliefs that the principal and agent hold. The principal maintains beliefs about the agent's type, corresponding to their initial distribution, as well as a potential target salience distribution that they would like the agent to have. The agent has a type α^*, values $v_1, \ldots, v_m$, and maintains beliefs about their own salience in the form of a Dirichlet distribution. These values and salience beliefs are based on features and their instances, which themselves characterise a number of alternatives available to the agent.

1. Feature Mentions f^j: The principal mentions feature j to the agent, thus drawing their attention to it and increasing its salience to the agent.

2. Comparison Question $q^{x,y} = (x, y) \in X \times X$: The principal asks the agent whether they prefer x or y.

3. Preference Response $r \in X$: The agent samples a salience distribution s from p_A and then compares $u_A(x, s)$ against $u_A(y, s)$ returning the alternative with the higher salience-weighted utility.

Suppose further that the agent's initial Dirichlet parameters α_0 are restricted to one of a finite set of *types* $\boldsymbol{\alpha} = \{\alpha^1, \ldots, \alpha^r\}$. This is a common modelling assumption in Bayesian game theory [456], which helps us to overcome the intractability of maintaining and performing inference over the set of all Dirichlet parameters. In our context, different types correspond to different classes of agents who have a tendency to weight different features differently. In the case of our running example, two possible types may correspond to a diligent and lazy student. The diligent student may place less initial weight on the difficulty feature and more on salary, whereas the lazy student may find course difficulty to be a more important issue for them.

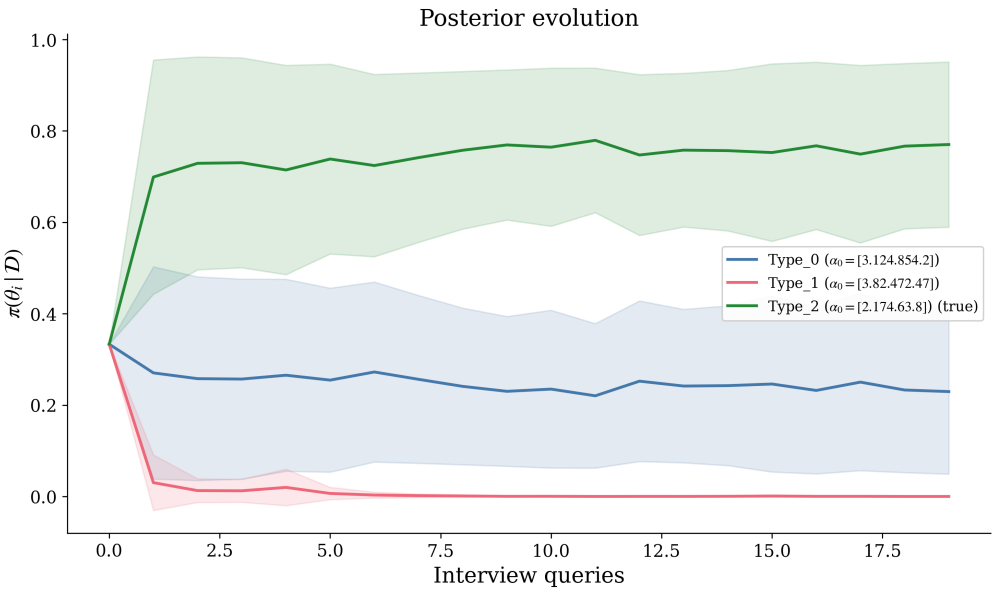

Figure 9.3: Illustration of the evolution of the principal's posterior beliefs about the agent's possible types as the interview phase progresses. Three types were randomly initialised and the average posterior over five runs was plotted, along with standard deviation margins.

Returning to the principal's task, recall that the principal wants the agent to adopt a certain salience distribution s. In order for this to be achieved, the principal may make use of both feature mentions to modify the salience of different features and comparison questions to ascertain what the agent's type may be. In what follows, we model the principal-agent interaction as occurring in two phases:

Interview phase The principal asks the agent a sequence of comparison questions, which the agent responds to by providing a preference response [457]. In this stage, the principal's objective is to identify the type of the agent, which is associated with a particular initial Dirichlet distribution representing the agent's beliefs about their saliences. More formally, the objective of the principal is to find a sequence of comparison questions $q_1, \ldots, q_T$ such that

$$p(\alpha^* | q_1, r_1, \ldots, q_T, r_T) \geq 1 - \epsilon, \tag{9.2}$$

for some small $\epsilon > 0$.

Persuasion phase Once the principal has identified the type of the agent, they send a sequence of feature mention messages to the agent, which modifies the

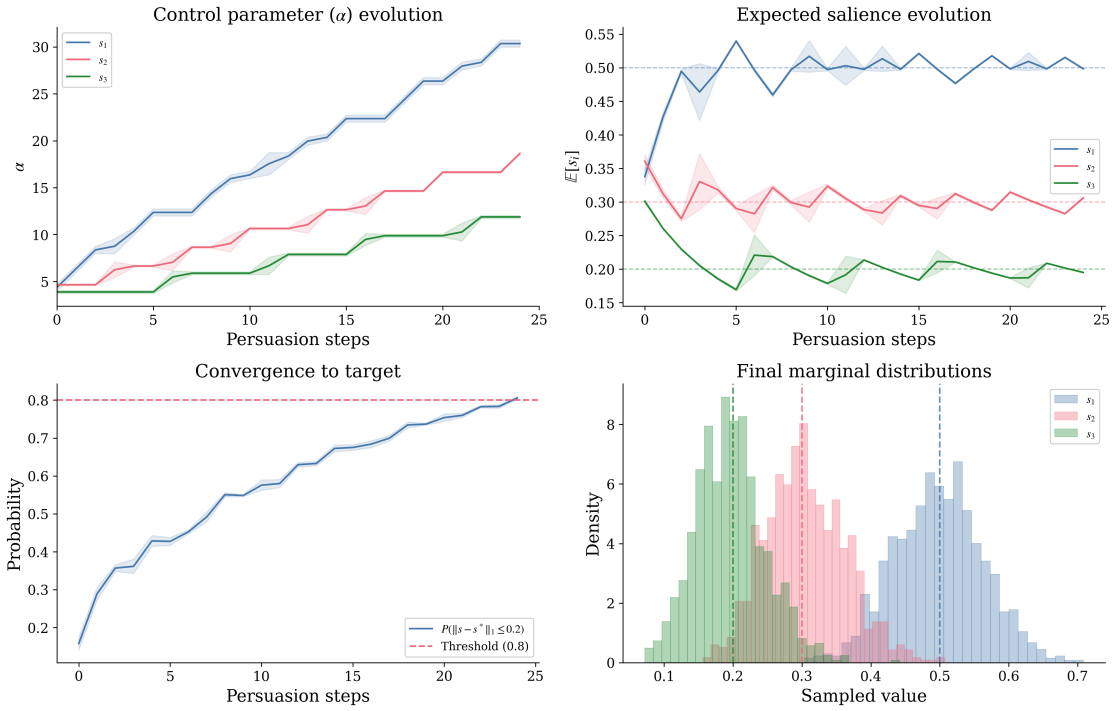

Figure 9.4: Plots illustrating different aspects of the persuasion phase. Top left: Evolution of the actual Dirichlet parameters of the agent. Top right: Evolution of the expected values of each salience probability under the agent's Dirichlet distribution. Bottom left: Evolution of the principal's confidence that the agent's sampled salience distribution is sufficiently close to the target distribution, which was (0.5,0.3,0.2). Bottom right: Frequencies of 10,000 sampled salience probability values from the Dirichlet distribution after the persuasion phase is completed.

agent's saliences by updating their Dirichlet parameters. [458] define persuasion as the 'sequential processing of information that is relevant to judgment formation.' Following this definition, we model the task as sending messages until the principal is sufficiently confident that the agent's sampled salience distribution is 'close enough' to the principal's target distribution. Formally, the task is to find a sequence of feature mention messages $f_1^j, \ldots, f_t^j$ such that

$$\Pr_{s \sim \text{Dir}_m(\alpha_t)} [d(s, s^*) \leq \delta] \geq 1 - \varepsilon, \tag{9.3}$$

where α_t are the parameters of the agent's Dirichlet belief p_A after receiving the messages $f_1^j, \ldots, f_t^j$, d is a distance function over discrete probability distributions, and $\delta, \varepsilon > 0$ are small positive constants.

Algorithm 5 Interview Phase

Input: Types $\boldsymbol{\alpha} = \{\alpha^1, \ldots, \alpha^n\}$, alternatives X, max queries M, confidence threshold ϵ
Output: Most likely type $\hat{\alpha}^*$

Initialise uniform posterior $p(\alpha) = 1/|\boldsymbol{\alpha}|$
for $t = 1$ to M **do**
  $q_t \leftarrow \arg\max_q \mathbb{E}_{r_t \sim p(r|q)}[D_{\mathrm{KL}}(p(\alpha|r_t) \| p(\alpha))]$
  Observe response r_t
  Update posterior: $p(\alpha|r_t) \propto p(r_t|\alpha)p(\alpha)$
  if $\max_\alpha p(\alpha|q_1, r_1, \ldots, q_t, r_t) \geq 1 - \epsilon$ **then**
    break
  end if
end for
return $\hat{\alpha}^* = \arg\max_\alpha p(\alpha|q_1, r_1, \ldots, q_T, r_T)$

9.5 Framework Demonstration

To illustrate the practical applications of our framework, we present a concrete demonstration showing how it can be used to model attention-driven preference change. We develop two baseline algorithms: one for inferring the agent's attentional state in the interview phase, and another for systematically influencing attention allocation. This demonstration serves three purposes. Firstly, it provides an explicit example of how our theoretical framework translates into implementable mechanisms. Secondly, it highlights essential design choices and trade-offs that any system for preference learning or influence must address. Thirdly, it establishes a foundation that future work can build upon to develop more sophisticated approaches. The methodology we present could also inform the design of empirical studies investigating how humans learn about and influence each other's preferences through attentional mechanisms.

Figure 9.3 depicts the evolution of the principal's posterior over the agent's types when following Algorithm 5. In this model, the principal selects comparison questions to ask the agent using a greedy algorithm by picking the query with the highest expected information gain, which is a common acquisition function in Bayesian experimental design [459]. Figure 9.4 illustrates several features of the persuasion phase, which was implemented according to Algorithm 6.

Algorithm 6 Persuasion Phase

Input: Initial type α^*, target salience s^*, max messages M, L1 distance threshold τ, confidence threshold ε, update scale γ
Output: Final Dirichlet parameters α_T

Initialise controller with $\alpha_0 \leftarrow \alpha^*$
for $t = 1$ to M **do**
  $p_j \leftarrow \Pr_{s \sim \mathrm{Dir}_m(\alpha_t + \gamma\delta_j)} [|s - s^*|_1 \leq \tau]$
  $\iota \leftarrow \arg\max_j p_j$
  $\alpha_t \leftarrow \alpha_{t-1} + \gamma\delta_\iota$
  if $\max_j p_j \geq 1 - \varepsilon$ **then**
    break
  end if
end for
return α_t

9.6 Discussion

In summary, this chapter has introduced a formal framework for modelling how salience shapes preferences and drives preference change that does not rely on belief updating about the underlying states of the world or value modification, but rather on shifts in attention allocation. This may help to explain empirically observed preference reversals that cannot be attributed to new information or changed values. Our key theoretical contribution is a representation theorem for preference relations indexed by salience distributions, demonstrating how attention-weighted features can combine to form overall preferences under a single set of value functions. Building on this foundation, we developed a computational model of interactions between a principal and agent, where the principal can infer the agent's type and strategically influence their preferences by modulating the salience of different attributes. The polynomial time algorithm we have identified for determining whether a target alternative can be made most preferred through salience modification gives practitioners a lightweight and concrete tool for analysing the feasibility of preference-shaping interventions. Finally, the concrete model of principal-agent interaction, with its distinct interview and persuasion phases, provides a template for designing systems that can learn about and adaptively influence human preferences.

Taking this work as our starting point, several important aspects of real-world preference dynamics can be incorporated into the model for specific applications. For example, considering the role of affect in determining both salience and perceived values, emotional factors can make certain features more salient and

influence how they are evaluated. Modelling strategic communication between the principal and agent is also a natural extension in low-trust or competitive scenarios, as well as the agent's active interaction with their environment. The framework could also be augmented to handle cases where features have interdependencies, rather than assuming separable attribute-wise evaluation. Moreover, the setting could be expanded to involve multiple agents, which raises questions about modelling inter-agent or group attention effects. In the context of human interactions, important interpersonal features such as trust and reputation may have a significant impact on the responsiveness of the agent to the principal. Additionally, the impact of novelty on salience could be incorporated, as research suggests that stimuli acquire added salience partially to the degree that they predict motivationally relevant consequences [460].

From the perspective of applications, this work provides foundations for designing AI systems that can effectively learn about and shape human preferences in support of individual and societal objectives. For example, [461] discuss how making the impact of one's decisions more salient in real-time in the context of resource consumption can promote more conservative resource usage behaviours, leading to a 22% reduction in energy consumption for targeted behaviours. In addition, [462] argues that strategically making use of *low* saliences can be effective in achieving desired outcomes, e.g., in the process of raising government revenue through taxes. However, these applications also highlight important considerations around the responsible deployment of such capability. The ability to systematically influence preferences through the shaping of attention, while potentially beneficial when aligned with human values and intentions, could also enable concerning forms of manipulation that bypass conscious awareness if misused. The ability to influence preferences through attention modulation may be particularly concerning because these effects can occur without changing underlying values or beliefs. This suggests the need for more sophisticated governance frameworks around preference-shaping AI systems that incorporate the insights that mathematical and computational models have to offer. If attention does play a significant role in human preferences, as the literature has been pointing to from many directions, we are in need of more precise technical definitions of concepts like preference stability and authenticity to develop principled constraints on allowable influence strategies, as well as to develop practices to defend ourselves against unwanted influence. Additionally, the possibility of unintended consequences when modifying attention patterns suggests that it would be prudent to develop robust validation protocols before deploying such systems. This raises many more questions:

How can we verify that attention-mediated preference changes align with an individual's deeper values? What distinguishes legitimate persuasion from problematic manipulation when operating through attentional mechanisms? Should systems treat temporary attention-driven preference shifts differently from lasting changes in underlying values?

This chapter takes steps towards a formal understanding of attention-driven preference change and illuminates important directions for future work. In particular, several important steps remain to close the gap between theory and practice. The framework articulated in this chapter serves as a more idealised model of how attention shapes decision-making through the use of classical utility function analysis, which still introduces a degree of modelling agents through the idealised rationality framework. Real human decision making is known not to obey many of the axioms upon which this framework is based and hence, there is a gap to close, which may be bridged by the use of process models such as evidence accumulation models. Developing such models, taking into account factors like value-based and information-based attention capture as well as bounded rationality would provide a promising next step towards being able to empirically evaluate models of and account for real human decision making.

As AI systems become increasingly capable of learning about and influencing human preferences, frameworks like the one presented here may be crucial for ensuring that they do so in ways that enhance rather than diminish human agency. By better understanding how attention shapes preferences, we can build systems that help people align their immediate choices with their deeper values and aspirations. This research opens up possibilities for AI systems that could help humans navigate complex situations more effectively in a way that promotes autonomy. The challenge ahead lies in thoughtfully developing these capabilities to empower humans in a way that is aligned with their authentic preferences and well-being.

10

Incentives and Interactions Between Boundedly Rational Agents

In this chapter, we begin by proposing a novel framework for modelling strategic interactions between boundedly rational agents in complex, partially observable environments. Our approach models agents that minimise the Path Divergence Objective (PDO) introduced in Chapter 7, capturing the divergence between their beliefs about future trajectories and their preferences, which are represented by a biased probabilistic model. We extend this to multi-agent settings and introduce *Free Energy Equilibria*, a new class of game-theoretic solution concepts. We begin by establishing the relationship between Free Energy Equilibria and existing game-theoretic solution concepts, and then study some properties of this solution concept. Then, we conclude with a discussion on the implications of our framework on incentives for boundedly rational agents.

How can we model agents strategically interacting in complex, partially observable environments? We argue that a model of both natural and artificial agency should at least capture (but is not limited to) (1) embodied agents who persist through time, model their world, and interact with it through an interface/Markov blanket [65, 68, 69], (2) strategic interactions between agents who may have different beliefs and preferences, (3) partially observable, stochastic environments, and (4) varying degrees of agent rationality. Traditional game theory provides a foundation for analysing strategic interactions but assumes perfect rationality. It also typically omits explicit models of agents' beliefs and

how they are revised over time, limiting its applicability to realistic scenarios with cognitively constrained agents [56, 463]. On the other hand, free energy-based frameworks such as Active Inference [331] and Action-Perception Divergence (APD) minimisation [336], are a compelling class of models of perception and action based on (variational) free energy minimisation, but are currently lacking a general theory of multi-agent interactions.

We propose an application of game theory in studying interactions between free energy minimising agents to bridge the gap between idealised game-theoretic results and the behaviour of both natural and artificial agents. Our model studies boundedly rational agents as those who minimise a free energy functional, which captures the divergence between their beliefs (represented by a predictive world model) and preferences (represented by a biased world model) over future trajectories. This leads us to the concept of Free Energy Equilibrium (FEE) and its generalisations.

Our framework is built upon *Partially Observable Stochastic Games* (POSGs) [464], extending the Partially Observable Markov Decision Process (POMDP) to multi-agent settings. Our framework allows one to model boundedly rational agents through (1) approximate latent state estimation using variational inference, (2) sub-optimal policy selection due to bounded rationality, and (3) varying levels of belief updating and self-modelling. Preferences are modelled by assuming that agents desire to minimise the KL divergence between 'expected' and 'desired' trajectories, a common approach in free energy methods, bounded rationality models, and variational inference [313, 331, 336].

Several assumptions delineate the scope of our formal analysis throughout this chapter. First, we work within *finite-horizon* POSGs whose state, action, and observation spaces are finite, ensuring that all relevant expectations and sums are well-defined. Second, for the purposes of the equilibrium analysis in Section 10.4 and the results relating FEE to classical solution concepts, we assume *common knowledge of the game structure*: each agent knows the transition function, observation likelihoods, and the utility functions of all players, as well as the policy profile of their opponents. Third, the formal results in this chapter concern *policy updating only*. We hold agents' world-model beliefs fixed and do not consider simultaneous belief updating, so as to maintain a clear parallel with classical game-theoretic equilibrium concepts. Fourth, prior policy profiles π_i^0 are assumed to have *full support* over each agent's action set, which is required to ensure that the Path Divergence Objective is finite and that the logit best-response characterisation of Proposition 10.3 is well-defined. These assumptions are made

in the interest of analytical tractability and conceptual clarity. We discuss avenues for relaxing them in Section 10.6.

Modelling Bounded Agents

The central motivation of this chapter is to accurately model interacting agents who are boundedly rational and operate under constraints of partial-observability and limited computational resources. We extend the PDO model developed in Chapter 7 to the multi-agent setting and then introduce a class of solution concepts that this model entails.

Partial Observability The inability of an agent to fully know the state of its external environment is an essential component of most real-world interactions. This is not only limited to adversarial scenarios, where it is intrinsically valuable to withhold information from other players, but partial-observability also plays a key role in both fully- and semi-cooperative scenarios with limited communication. In this work, we assume a general-sum, multi-player Partially Observable Stochastic Game (POSG) model. This model extends the Partially Observable Markov Decision Process (POMDP) framework to situations involving multiple agents interacting with each other and a shared environment in a discrete-time setting. Indeed, according to the Free Energy Principle, systems that persist over time maintain a 'boundary' that separates processes internal to the system from processes external to the system. This 'boundary', often formally described in the literature as a Markov blanket, couples a system to its environment and mediates all interactions between the processes that are internal and external to the boundary. In particular, those aspects of the boundary which afford the system's internal processes the ability to influence external processes are referred to as *active processes* and such influences on the external processes are referred to as *actions*. Complementarily, the remaining aspects of the boundary which transmit the influence of the external processes to the internal processes of the system are referred to as *sensory processes*, and such influences are referred to as *sensations*.

Bounded Rationality As discussed in Chapter 7, bounded rationality characterises the limitations in agents' decision making capabilities due to their finite cognitive resources and environmental complexity. Here, we identify three primary ways in which bounded agents differ from idealised unbounded agents.

Firstly, the estimation of the game's current hidden state from an agent's local observations is a formidable challenge and is generally intractable. We model this by assuming the agents use an approximate distribution of states conditional on the player's observations, which is inferred by maximising an Evidence Lower Bound (ELBO). This is a commonly used approach within machine learning as well as within the active inference literature, where it is interpreted as Variational Free Energy minimisation.

Even while using the above approximation of the hidden state, agents may not always learn or implement the optimal policy, and instead settle for an approximately optimal policy. Common models of this (e.g., information-theoretic bounded rationality [317] and quantal response equilibria [465]) are closely related to thermodynamics and free energy, information theory, and the maximum entropy principle. While our results primarily focus on optimal policy selection, our choice of the free energy framework also naturally allows one to model this aspect of boundedness.

A third aspect of boundedness is the level of an agent's "sophistication" – how the agent updates its beliefs about the hidden state of the game at a particular point in time based on observations made at later points in time, and how the agent models its belief updating in the future. These choices present additional trade-offs between accuracy and tractability.

Preference Modelling We model agents' preferences as the desire to minimise the Kullback-Leibler divergence between their *expected* and *desired* trajectories, an approach common to free energy methods, bounded rationality models, and also frequently used variational methods across machine learning.[1]

Subjective beliefs. In many situations, agents may not have an accurate knowledge of the game mechanics – whether this is due to a highly complex environment relative to their cognitive capacity, or because they are in a situation where they have not successfully learned a model of the environment. Our presentation explicitly represents the beliefs and models of individual agents, which may differ from the true generative model. Note that having an incorrect model of the environment may make it impossible for agents' revealed preferences to be

[1]Many of the optimisation objectives in contemporary machine learning (i.e., loss functions), even relatively complex ones, can be naturally represented as KL-divergence minimisation [466].

represented as simple preferences over the hidden states, motivating the more expressive preference models that feature in this framework.

Equilibria and Learning Understanding equilibria is a crucial tool for analysing complex systems outside of the domain of game theory – even for real-world systems that may not always reach these states. By studying and classifying equilibria, we gain valuable insights into the system's dynamics, identifying stable points and potential attractors that can guide our understanding of agent interactions in both static and dynamic environments. Similarly, learning procedures that provably converge to equilibria often offer valuable approximations of such equilibria even before convergence. We expand on this topic further in the discussion section of this chapter.

Contributions

The contributions of this chapter are as follows. Firstly, we discuss several topics closely related to our work, including game theory and economics, computer science and machine learning, and active inference. Then, we introduce the main technical concepts underpinning our results and the relevant notation: partially observable stochastic games, value functions, common game-theoretic solution concepts, and the basic concepts of active inference. Following this, we recapitulate and expand the definitions of prediction models, preference models, and the PDO from Chapter 7 to the multi-agent setting, and subsequently introduce two solution concepts derived from the PDO. One of the solution concepts corresponds to independent policy profiles and Nash equilibria, and the other to correlated policy profiles and coarse correlated equilibria. Both solution concepts have their applications: independent policies model the causal separation between individual agents, and correlated equilibria capture the nuances of external coordination mechanisms and empirical distributions of strategy learning in environments where convergence to a single strategy is not guaranteed. Finally, we sketch some initial ideas on the application of our model to incentives for boundedly rational agents, and then conclude by describing some of the many open problems and proposed research directions.

10.1 Related Work

A comprehensive review of all work relevant to our work would deserve a dedicated study, but we endeavour to provide a broad background on the most pertinent areas. We begin our overview by outlining some of the key contributions that some of the most relevant fields have made to each other, and suggest applications of the theory presented herein to each of these disciplines in turn:

Economics and Game Theory

There has been much work in the economics and game theory literature on developing models of bounded rationality, which was based on the desire to more faithfully capture the empirical behaviours of real-world decision makers [56].[2] Our work most closely relates to information-theoretic models of boundedness, which are motivated by the assumption that agents possess bounded information-processing capacities, and are faced with the task of decision making under such constraints [313, 317]. It is also closely related is the work on rational inattention, which emphasises the fact that agents must selectively decide what information to attend to and incorporate in their decision making [320–322]. In addition, models of bounded rationality have been applied to strategic interactions using statistical and information-theoretic models [465, 467–472]. However, these models do not explicitly account for the internal models or beliefs that agents use to plan and make decisions.

Mechanism design, a subfield of game theory, focuses on designing rules that lead to desirable outcomes in strategic settings. Foundational concepts were developed by Leonid Hurwicz [473], while later contributions introduced key ideas such as incentive compatibility [474] and equilibrium selection [475]. Social choice theory explores the aggregation of individual preferences to implement collective decision making. This framework could be applied to explaining behavioural experiments, offering a theoretical framework for policy design, studying social behaviours, and potentially providing microfoundations for macroeconomic phenomena due to its applicability across scales [476, 477]. In incentive design, it can suggest ways to combine 'utility-based' and 'information-based' incentives to shape behaviour [11].

[2] See [317] for a good overview of existing approaches to bounded rationality in the literature.

Computer Science

Information theory and Bayesian inference form a bridge between modern AI and biologically-inspired theories of agency. For example, many machine learning problems can be viewed as KL divergence minimisation [332, 335, 336, 466] or Evidence Lower Bound maximisation [350, 351]. Recent language modelling approaches using Reinforcement Learning (RL) with KL divergence penalties can be viewed as approximate Bayesian inference [339]. Given these developments, it is natural to assume that it will also be possible to model many artificial agents based on neural networks and reinforcement learning as being bounded in the ways we assume: inferring approximate distribution of states based on observations, having subjective beliefs, and implementing some computationally tractable approach to updating such beliefs.

Other relevant sub-fields of computer science are multi-agent reinforcement learning and algorithmic game theory. The key obstacle in these settings is that finding – and sometimes even approximating – many of the traditional solution concepts is computationally intractable in general [478]. This motivates the search for alternative solution concepts, as well as methods which either provide guarantees under additional assumptions or offer sufficiently good empirical performance. This can be illustrated particularly well in the example of no-regret learning dynamics in sequential imperfect-information games: the popular CFR algorithm will find an approximate Nash equilibrium in two-player zero-sum games but might fail to even converge in general [479]. However, despite the negative theoretical results, methods based on this approach have been successful even outside of two-player, zero-sum games [480]. Moreover, these learning dynamics are closely connected to rational behaviour in the context of *correlated* play and learning [481]. This finding is one of the reasons why the present chapter focuses on the notion of correlated equilibria [482].

An important distinction in the field of (multi-agent) reinforcement learning (RL) is the difference between model-free and model-based methods. While much work has been dedicated to developing scalable model-free methods [62, 64], these methods typically require a significant amount of training data to achieve a high degree of competence. As a result, there has recently been growing interest in developing model-based techniques in both single-agent and multi-agent settings [483–485]. The approach we propose may provide a general framework for developing model-based (multi-agent) RL algorithms in partially observable settings by learning world models of players and environments for

planning [485–487]. Indeed, a closely related framework known as Maximum Diffusion RL has recently been proposed as a principled way of regularising reward functions for RL agents to decorrelate their *experiences*, which leads to an objective very similar to the one we propose here [346], but arrived at from a different starting point.

Active Inference

Active Inference is a theoretical framework for modelling biological systems [319], which has more recently found useful applications in describing generally agentic systems in the formalism of POMDPs [70]. There is a rapidly growing body of literature in the field, with successful applications to modelling empirical data across a wide range of scales and contexts [488]. In particular, there has been much interest in studying interactions between multiple active inference agents, focusing on applications such as communication [363, 398, 401, 489–491], competition [492], coordination [493–496], social cognition [399, 400, 497–499], collective behaviour [500–505], and hierarchical self-organisation [500, 506–510]. Despite this, the field is currently lacking a general formalism and theory that can be used to model such multi-agent interactions, in the same way that POMDPs [70, 356, 420] and algorithms for (approximately) solving them [511, 512] have been applied to model active inference in single-agent settings. We hope to take steps towards addressing this gap in the literature with our proposed model.

10.2 Preliminaries

We use the framework of *Partially Observable Stochastic Games* (POSGs) [464], which are a natural extension of Partially Observable Markov Decision Processes (POMDPs) to situations involving multiple agents acting concurrently.[3]

Definition 10.1. A finite-horizon **Partially Observable Stochastic Game** (POSG) is a tuple

$$\mathcal{G} = (N, S, (A_i)_{i \in N}, (\Omega_i)_{i \in N}, (O_i)_{i \in N}, T, p, I, (\mathcal{U}_i)_{i \in N}),$$

where:

- N is a finite set $\{1, \dots, n\}$ of **agents**;

[3] Future work may consider extensions to this model that more explicitly represent the knowledge shared between agents [513], but this is not necessary for the present scope of study.

- S is a set of **states**;
- A_i is a set of **actions** for each $i \in N$. Recall that we write $\vec{A} = \bigtimes_{i \in N} A_i$ for the set of **joint actions**;
- Ω_i is a **set of observations** for each $i \in N$. We write $\Omega = \bigtimes_{i \in N} \Omega_i$ for the set of **joint observations**;
- $T \in \mathbb{Z}^+$ is the **time horizon**. We write $\mathbb{T} = \{0, \ldots, T\}$ for the set of time steps in the game;
- $O_i : \vec{A} \times S \rightharpoonup \Delta(\Omega_i)$ is a partial **observation probability function** (or observation likelihood function) for each agent $i \in N$, which represents the probability distribution over possible observations that agent i can make, given each possible state that the game is in and joint action that the agents took in the previous step. For a joint observation $\vec{o} = (o_i)_{i \in N} \in \Omega$ and a time $t \in \mathbb{T}$, we write $O(\vec{o}^{\,t} | \vec{a}^{\,t-1}, s^t) = \prod_{i \in N} O_i(o_i^t | \vec{a}^{\,t-1}, s^t)$ for the joint probability that each agent $i \in N$ receives observations o_i^t, given that the state is s^t and the most recently played joint action was $\vec{a}^{\,t-1}$. For the initial time $t = 0$, we assume that the joint action $\vec{a}^{\,-1}$ is a null element and hence, the first joint observation $\vec{o}^{\,0}$ is solely a function of the initial state s^0. Note that for convenience, we also assume that the previous action of a player can be deduced from their next observation, thus eliminating the need to track past actions in game history and preserving perfect recall.
- $p : S \times \vec{A} \to \Delta(S)$ is a Markovian **probabilistic transition function**, which can also be written as a conditional probability distribution $P(s^{t+1} | s^t, \vec{a}^{\,t})$ for times $t \in \{0, \ldots, T-1\}$;
- $I \in \Delta(S)$ is the **initial state distribution**;
- $\mathcal{U}_i : \mathbb{H}_i \to \mathbb{R}$ is agent i's **history utility function**, where

$$\mathbb{H}_i := \bigcup_{t=0}^{T-1} (S \times \Omega_i \times A_i)^t \times S \times \Omega_i$$

  is the set of histories of the game observable by player i.

A POSG is played as follows: in the first time step, an initial state $s^0 \sim I$ is sampled from the initial state distribution, each agent $i \in N$ receives an observation $o_i^0 \sim O_i(s^0, \emptyset)$. Then, agents simultaneously choose an action from their respective action sets, giving rise to a joint action $\vec{a}^{\,0}$. The next state $s^1 \sim$

$p(s^0, \vec{a}^{\,0})$ of the game is then sampled according to the probabilistic transition function P, and each agent $i \in N$ receives an observation $o_i^1 \sim O_i(s^1, \vec{a}^{\,0})$. We assume that every player observes their own actions (as per the definition of O_i above). The process repeats T times in total, after which the game concludes.

To capture the output of this process between any two timesteps t_0 and t, $0 \leq t_0 \leq t \leq T$, we use the notion of a *trajectory* $\vec{h}_{\,t_0:t} := s^{t_0}\vec{o}^{\,t_0}\vec{a}^{\,t_0} \ldots s^{t-1}\vec{o}^{\,t-1}\vec{a}^{\,t-1}s^t\vec{o}^{\,t}$ — that is, an alternating sequence of states, joint observations, and joint actions. To refer to the sequences consisting only of specific objects, we use the terms *state trajectory* $s^{t_0:t} := s^{t_0}, \ldots, s^t$, *action trajectory* $\vec{a}^{\,t_0:t} := \vec{a}^{\,t_0}, \ldots, \vec{a}^{\,t-1}$, and *observation trajectory* $\vec{o}^{\,t_0:t} := \vec{o}^{\,t_0}, \ldots, \vec{o}^{\,t}$. We use the term *history* (or run) $\vec{h}^{\,0:t}$ to refer to trajectories that start at timestep $t_0 = 0$, and similarly for state histories $\mathrm{s}^{0:t}$, observation histories $\vec{o}^{\,0:t}$, etc. We let $\mathbb{H}^t = (S \times \Omega \times \vec{A})^{t-1} \times S \times \Omega$ denote the set of all possible histories of length $t \in \{1, \ldots, T\}$ (that is, up to the t-th joint observation), and thus the set of all possible histories of any length can be alternatively defined as $\mathbb{H} := \bigcup_{t=1}^{T} \mathbb{H}^t$. We sometimes use $\vec{h} = \vec{h}^{\,0:T}$ as shorthand for full histories. We similarly define $\mathbb{O}^t$ as the set of all possible observation trajectories of length t and $\mathbb{O} := \bigcup_{t=1}^{T} \mathbb{O}^t$ as the set of all possible observation trajectories of any length. Similarly, we use $\mathbb{O}_i^t = (\Omega_i)^t$ and $\mathbb{O}_i = \bigcup_{t=1}^{T} (\Omega_i)^t$ to denote the sets of observation histories of a given player. For notational convenience, we let $\vec{h}_{\,-1}, \vec{o}^{\,-1}, \vec{a}^{\,-1} := \emptyset$ denote the empty history before the game starts.

An independent *policy* $\pi_i : \mathbb{O}_i \to \Delta(A_i)$ for player i maps each observation history of i to a probability distribution over their actions.

In the multi-agent setting, we define the *reach probability* of a history $\vec{h} = \vec{h}^{\,0:t}$ under an independent policy profile $\vec{\pi}$ as

$$p(\vec{h}\,; \vec{\pi}\,) := I(s^0) \cdot \left(\prod_{\tau=0}^{t-1} O(\vec{o}^{\,\tau} | a^{\tau-1}, s^\tau) \cdot \vec{\pi}\,(\vec{a}^{\,\tau} | \vec{o}^{\,0:\tau}) \cdot p(s^{\tau+1} | s^\tau, \vec{a}^{\,\tau}) \right) \cdot O(\vec{o}^{\,t} | s^t), \tag{10.1}$$

where $\vec{\pi}\,(\vec{a}^{\,\tau} | \vec{o}^{\,0:\tau}) := \prod_{i=1}^{n} \pi_i(a_i^\tau | o_i^{0:\tau})$. Analogously, we define the reach probability of $\vec{h}$ under a *correlated* policy profile $\vec{\mu}$ as $p(\vec{h}\,; \vec{\mu}\,) := \mathbb{E}_{\vec{\pi} \sim \vec{\mu}}\, p(\vec{h}\,; \vec{\pi}\,)$. For any (independent or correlated) policy profile $\vec{\pi}$, these notions can be straightforwardly extended to track other probability distributions related to the game [514], such as the conditional probabilities $p(\vec{h}^{\,0:t} | \vec{h}^{\,0:t_0}; \vec{\pi}\,)$ of reaching a given history $\vec{h}^{\,0:t}$ given that the current history is $\vec{h}^{\,0:t_0}$, $t_0 \leq t$.

Value functions and Solution Concepts

We extend the notion of value functions as described in Chapter 7 to the multi-agent setting, where the main difference is that an agent's expected utility now depends on the joint behaviours of all agents in the game.

Definition 10.2. Given an (independent or correlated) policy profile $\vec{\pi}$ and a history $\vec{h}^{\,0:t}$, the **value function** of agent i is given by

$$\mathcal{V}_i(o_i^{0:t}; \vec{\pi}) = \mathbb{E}_{p(\vec{h}^{\,0:T} \mid o_i^{0:t}; \vec{\pi})} \left[\mathcal{U}_i(\vec{h}^{\,0:T}) \right], \text{ where } \vec{h}^{\,0:T} = \vec{h}^{\,0:t}\, \vec{a}^{\,t} s^{t+1} \dots s^T \vec{o}^{\,T}. \quad (10.2)$$

We write the total expected utility under $\vec{\pi}$ as $\mathcal{V}_i(\vec{\pi}) = \mathcal{V}_i(\emptyset; \vec{\pi})$.

This measures the true expectation of the utility-to-go for a player under a given joint policy $\vec{\pi}$ and the information that they have access to, i.e., $o_i^{0:t}$. The objective of a classical expected utility-maximising agent is thus to select a policy which maximises its initial value function, given information about the policies of the other players. Game theory typically assumes an equilibrium state where each player knows the policies of the others [515]. However, this assumption can be relaxed, giving rise to the need to learn and possibly even shape opponent policies [516–518].

In the literature on POMDPs and POSGs, it is quite common to introduce a *belief state*, which represents an agent's beliefs about the current state and allows one to transform a partially observable game into a fully observable one with an augmented (and infinite) state space [62, 464]. In our framework, belief states are maintained by an agent's world model, rather than explicitly augmenting the state space with belief states.

The historically prominent solution concept in game theory is the Nash equilibrium. However, in the context of this chapter, we will be more interested in the notion of correlated equilibrium, where players use an external source of randomness (e.g., a mediator or the outcome of some random process) to coordinate on which joint actions to play.

While the notion of a correlated equilibrium is relatively straightforward in single-step interactions (i.e., normal form games), there are several ways of extending the idea to sequential settings [482]. What all these variants have in common is that (i) the default expectation is that the players use the correlated policy profile $\vec{\mu}$ to randomly select a pure policy profile $\vec{\pi} \sim \vec{\mu}$ to adopt, and (ii) $\vec{\mu}$ is said to be a correlated equilibrium if none of the players can benefit

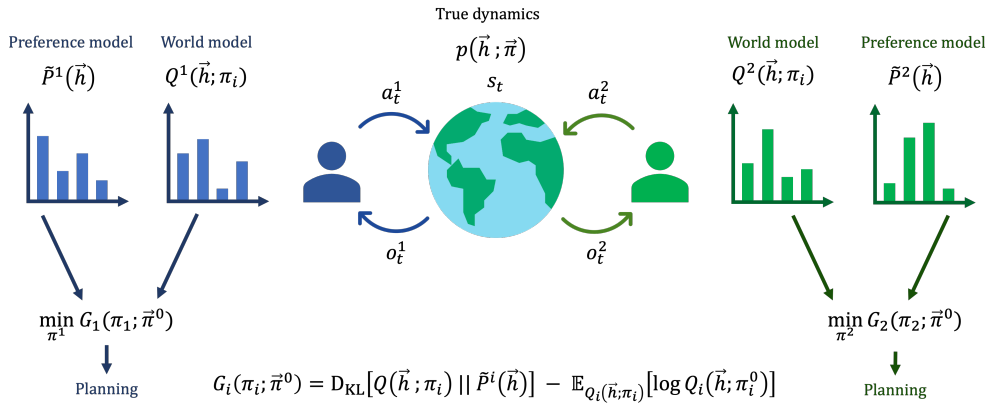

Figure 10.1: High-level summary of our approach to bounded rationality in multi-agent systems. Each agent maintains a predictive world model of its environment and other agents, which is approximated by a variational distribution Q_i. Agents also possess a preference model, which assigns higher probabilities to more preferable trajectories. Policies are selected to minimise their PDO, formalising the intuition that agents aim to act in order to minimise the difference between their expectations and their desires.

by unilaterally deviating from this plan. Where the variants differ is the types of deviations that are available to the players. To simplify the exposition, we only consider the (arguably) simplest variant, where a player can only decide to deviate at the start of the game, *before* learning which pure policies will be adopted by the other players.[4]

10.3 World Models, Preference Models, and Divergence Objectives

We introduce a world model and preference model for POSGs, reintroduce the PDO, and introduce two solution concepts based on it: one corresponding to independent policy profiles and Nash equilibria, and the other to correlated policy profiles and coarse correlated equilibria. We then formally show these relationships and other properties of the solution concepts. The main purpose of the following analysis is to demonstrate the connection of the proposed solution concept to classic game theoretical solution concepts. As such, we will be required to make epistemic assumptions about the agents [515]. However, we emphasise that the general framework here makes no commitment to agents

[4] In [482], this notion is referred to as *normal-form coarse correlated equilibrium*, where the *normal-form* part refers to only being able to deviate at the start of the game and the *coarse* part refers to having to decide on the deviation before hearing the mediator's recommendation.

having perfectly accurate predictive models or being able to perform exact Bayesian inference. The world model captures an (unbiased) belief about the dynamics of the environment, and the preference model encodes preferences of the agent as a desired (and therefore biased) distribution over states. In this section, we will show that the objective of a boundedly rational agent can be formulated as aiming to minimise the divergence between these two distributions through their choices of actions.

World Models

In general, an agent's world model consists of probabilistic beliefs

(a) $Q_i(s^0)$, about the initial state distribution $I \in \Delta(S)$,

(b) $Q_i(\vec{o}^{\,\tau}|s^\tau)$, about the joint observation likelihood function O of all players,

(c) $Q_i(s^{\tau+1}|\vec{a}^{\,\tau}, s^\tau)$, about the transition function p,

(d) $Q_i(a_i^\tau|o_i^{0:\tau}) = \pi_i(a_i^\tau, o_i^{0:\tau})$, about their policy function which we assume to be known to the agent, and

(e) $Q_i(\vec{a}^{\,\tau}_{-i}|\vec{o}^{\,0:\tau})$, about the (possibly correlated) policy profile $\vec{\pi}_{-i}$ (or $\vec{\mu}_{-i}$).

To obtain a world model analogous to the reach probabilities given by Eq. (10.1), the player combines the above ingredients using the formula

$$Q_i(\vec{h}^{\,0:t};\, \pi_i) := Q_i(s^0) \cdot \Bigg(\prod_{\tau=0}^{t-1} Q_i(\vec{o}^{\,\tau}|s^\tau) \cdot Q_i(s^{\tau+1}|s^\tau, \vec{a}^{\,\tau}) \cdot \pi_i(a_i^\tau|o_i^{0:\tau}) \cdot Q_i(\vec{a}^{\,\tau}_{-i}|\vec{o}^{\,0:\tau}) \Bigg) \cdot Q_i(\vec{o}^{\,t}|s^t) \tag{10.3}$$

In game theory, it is often assumed that players have accurate knowledge of each other's policies. In such cases, we can replace the terms $\pi_i(a_i^\tau|o_i^{0:\tau}) \cdot Q_i(\vec{a}^{\,\tau}_{-i}|\vec{o}^{\,0:\tau})$ by $\vec{\pi}\,(\vec{a}^{\,\tau}|\vec{o}^{\,0:\tau})$ to give a world model of the form $Q_i(\vec{h}^{\,0:t}; \vec{\pi}\,)$ (and analogously for correlated $\vec{\mu}$ and $Q_i(\vec{h}^{\,0:t}; \vec{\mu}\,)$).

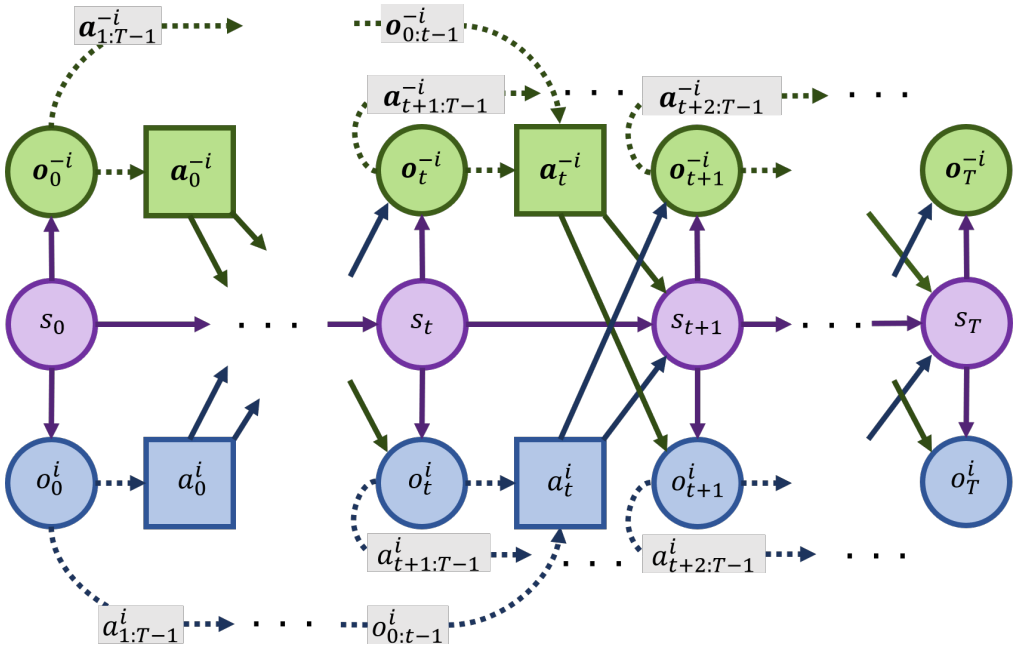

Figure 10.2: Multi-Agent Influence Diagram representing an agent i's predictive generative model. Square nodes represent conditional action distributions that some agent must select, given their history of observations. Solid arrows denote conditional dependencies, whereas dotted arrows represent the information used by agents to decide on their actions. For simplicity, we depict beliefs about other agents' actions and observations using a single node with labels $-i$ but in the full depiction, this would be split into one node for each other agent present in the game. Thus, arrows in the upper half of the diagram actually represent a collection of individual arrows for different agent nodes that follow the exact same pattern as the bottom half of the diagram, as opposed to the interpretation that all other agents' observations influence all other agents' actions, etc.

Path Divergence Objective

We adopt the model of bounded rationality described in Chapter 7, using the path divergence objective as a descriptive model of information-theoretic bounded rationality in POSGs. We note that in this chapter, we only consider policy updating and not belief updating to make the parallels with classical game theory more obvious, though much of the analysis can be extended to the case where both are considered.

Definition 10.3. The **Path Divergence Objective** (PDO) for an agent i in a POSG $\mathcal{G}$ given a prior policy profile $\vec{\pi}^{\,0}$ and a posterior policy π_i is given by:

$$G_i(\pi_i; \vec{\pi}^{\,0}) := \mathrm{D}_{\mathrm{KL}}\left[Q_i(\vec{h}^{\,0:T}; \pi_i) \,||\, \tilde{P}^i(\vec{h}^{\,0:T})\right] - \mathbb{E}_{Q_i(\vec{h}^{\,0:T}; \pi_i)}\left[\log Q_i(\vec{h}^{\,0:T}; \pi_i^0)\right], \quad (10.4)$$

where

$$\tilde{P}^i(\vec{h}^{\,0:T}) := \frac{\exp\left(\beta_i \mathcal{U}_i(\vec{h}^{\,0:T})\right)}{Z(\beta_i; \mathcal{U}_i)},$$

and $Z(\beta_i; \mathcal{U}_i) := \sum_{\vec{h}^{\,0:T'} \in \mathbb{H}^T} \exp\left(\beta_i \mathcal{U}_i(\vec{h}^{\,0:T'})\right)$.

Decomposition of the PDO in Multi-Agent Settings

In the same manner as in Chapter 7, we can decompose the PDO in the multi-agent setting. Firstly, interpreting (negative) pragmatic value as an energy, recall that the PDO is an upper bound on an energy minus entropy:

$$G_i(\pi_i; \pi_i^0) = \underbrace{\mathrm{D}_{\mathrm{KL}}\left[Q_i(\vec{h}^{\,0:T}; \pi_i) \,||\, \tilde{P}_i(\vec{h}^{\,0:T})\right]}_{\text{Divergence from preferences}} + \underbrace{\mathbb{E}_{Q_i(\vec{h}^{\,0:T}; \pi_i)}\left[-\log Q_i(\vec{h}^{\,0:T}; \pi_i^0)\right]}_{\text{Cross entropy from prior}} \tag{10.5}$$

$$= \underbrace{\mathbb{E}_{Q_i(\vec{h}^{\,0:T}; \pi_i)}\left[-\log \tilde{P}_i(\vec{h}^{\,0:T})\right]}_{\text{-Value (Energy)}} + \underbrace{\mathrm{D}_{\mathrm{KL}}\left[Q_i(\vec{h}^{\,0:T}; \pi_i) \,||\, Q_i(\vec{h}^{\,0:T}; \pi_i^0)\right]}_{\text{Divergence from prior}} \tag{10.6}$$

$$\geq \underbrace{\mathbb{E}_{Q_i(\vec{h}^{\,0:T}; \pi_i)}\left[-\log \tilde{P}_i(\vec{h}^{\,0:T})\right]}_{\text{-Value (Energy)}} - \underbrace{H\left(Q_i(\vec{h}^{\,0:T}; \pi_i)\right)}_{\text{Entropy}} \tag{10.7}$$

Analogously to Chapter 7, we can decompose the divergence term in the PDO in the multi-agent setting in terms of an *epistemic value*, *pragmatic value*, *social preferences*, and *intention-behaviour gap*. Again, the only additional assumption that is required to license this interpretation is that $\tilde{P}^i(s^{0:T}|o_i^{0:T}) = Q_i(s^{0:T}|o_i^{0:T})$. Under this assumption, we have the following decomposition of the divergence term $\mathrm{D}_{\mathrm{KL}}\left[Q_i(\vec{h}^{\,0:T}; \pi_i) \,||\, \tilde{P}^i(\vec{h}^{\,0:T})\right]$:

$$\begin{aligned}
&- \underbrace{\mathbb{E}_{Q_i(o_i^{0:T}, a_i^{0:T})}\left[\mathrm{D}_{\mathrm{KL}}\left[Q_i(s^{0:T}|o_i^{0:T}, a_i^{0:T}) \,||\, Q_i(s^{0:T}|a_i^{0:T})\right]\right]}_{\text{Epistemic Value}} \\
&+ \underbrace{\mathbb{E}_{Q_i(s^{0:T}, a_i^{0:T})}\left[\mathrm{D}_{\mathrm{KL}}\left[Q_i(o_i^{0:T}|s^{0:T}, a_i^{0:T}) \,||\, \tilde{P}^i(o_i^{0:T}|a_i^{0:T})\right]\right]}_{\text{Pragmatic Value}} \\
&+ \underbrace{\mathbb{E}_{Q_i(s^{0:T}, o_i^{0:T}, a_i^{0:T})}\left[\mathrm{D}_{\mathrm{KL}}\left[Q_i(\vec{o}_{-i}^{\,0:T}, \vec{a}_{-i}^{\,0:T}|s^{0:T}, o_i^{0:T}, a_i^{0:T}) \,||\, \tilde{P}^i(\vec{o}_{-i}^{\,0:T}, \vec{a}_{-i}^{\,0:T}|s^{0:T}, o_i^{0:T}, a_i^{0:T})\right]\right]}_{\text{Social Preferences}} \\
&+ \underbrace{\mathrm{D}_{\mathrm{KL}}\left[Q_i(a_i^{0:T}) \,||\, \tilde{P}^i(a_i^{0:T})\right]}_{\text{Intention-Behaviour Gap}}.
\end{aligned}$$

The key difference here from the single agent setting is the ability for the preference model to represent *social preferences*, which measure the divergence

between the agents' predictions about *other agents'* observations and actions, and its preferences over the same. Social preferences are an important component of organisms and agents in social environments and manifest in humans through properties such as empathy or inequity aversion.

10.4 Free Energy Equilibrium

Given this, we are now in a position to propose our core solution concepts:

Definition 10.4. Given an initial policy profile $\vec{\mu}^{\,0}$, a correlated policy profile $\vec{\mu}$ is a **Coarse Correlated Free Energy Equilibrium** (CCFEE) in a POSG $\mathcal{G}$ with respect to $\vec{\mu}^{\,0}$ if for all $i \in N$ and all $\hat{\pi}_i \in \Pi_i$, it holds that

$$G_i(\vec{\mu}\,; \vec{\mu}^{\,0}) \leq G_i((\hat{\pi}_i, \vec{\mu}_{\,-i}); \vec{\mu}^{\,0}).$$

An independent policy profile $\vec{\pi}$ is called an **Independent Free Energy Equilibrium** (IFEE) if it satisfies the analogous condition (with $\vec{\pi}$ in place of $\vec{\mu}$). We let $\mathrm{CCFEE}(\mathcal{G})$, resp. $\mathrm{IFEE}(\mathcal{G})$, denote the set of all CCFEEs, resp IFEEs, of $\mathcal{G}$.

Since every independent policy profile $\vec{\pi}$ can be represented as a correlated policy profile, we immediately get

$$\mathrm{IFEE}(\mathcal{G}) \subseteq \mathrm{CCFEE}(\mathcal{G}). \tag{10.8}$$

We remark that in the context of active inference, the usage of the term 'equilibrium' may be a source of confusion, because agents are typically characterised as systems that are far from equilibrium with their environments. Our model is consistent with this characterisation due to the assumption that each agent is a persistent entity over time who is acting in a shared environment in order to realise their preferences. The term 'equilibrium' in this context refers not to a property of the agents themselves, but rather to a property of their *policies*. In particular, a free energy equilibrium can be thought of as a fixed point of a best-response dynamics, where no agent has the capability to further minimise their PDO by changing their policy, given that other agents' policies remain fixed.

Connections to Existing Solution Concepts

Here, we establish the relationship between the instantiations of the family of free energy equilibria proposed above to existing solution concepts in the game theory literature. We begin by formally showing how CCFEE and IFEE relate to Coarse Correlated and Nash equilibria in POSGs, and then briefly discuss the connection to Quantal Response Equilibrium.

Coarse Correlated and Nash Equilibrium

Theorem 10.1. *For a commonly known POSG $\mathcal{G}$, if $(\vec{\mu}^{\,(k)})_{k\geq 1}$ is a sequence of CCFEEs with $\beta_i^{(k)} \to \infty$ for all $i \in N$, then every accumulation point of this sequence belongs to* $\mathrm{CCE}(\mathcal{G})$.

Proof. Let $\vec{\mu}$ be a CCFEE under the PDO with parameters $(\beta_i)_{i\in N}$. For any player i and deviation $\hat{\pi}_i$,

$$G_i(\vec{\mu}\,; \vec{\mu}^{\,0}) \leq G_i\left((\hat{\pi}_i, \vec{\mu}_{\,-i}); \vec{\mu}^{\,0}\right).$$

Expanding the definition of the PDO using the energy minus divergence decomposition in Equation 10.6 and rearranging gives

$$\beta_i\left(\mathbb{E}_{Q_i(\vec{h}^{\,0:T};\vec{\mu}\,)}\left[\mathcal{U}_i(\vec{h}^{\,0:T})\right] - \mathbb{E}_{Q_i(\vec{h}^{\,0:T};(\hat{\pi}_i,\vec{\mu}_{\,-i}))}\left[\mathcal{U}_i(\vec{h}^{\,0:T})\right]\right) \geq \Delta_i(\hat{\pi}_i;\pi),$$

where

$$\Delta_i(\hat{\pi}_i;\vec{\mu}\,) \coloneqq \mathrm{D}_{\mathrm{KL}}\left[Q_i(\vec{h}^{\,0:T};\vec{\mu}\,) \,||\, Q_i(\vec{h}^{\,0:T};\vec{\mu}^{\,0})\right] - \mathrm{D}_{\mathrm{KL}}\left[Q_i(\vec{h}^{\,0:T};(\hat{\pi}_i,\vec{\mu}_{\,-i})) \,||\, Q_i(\vec{h}^{\,0:T};\vec{\mu}^{\,0})\right].$$

The term $\Delta_i(\hat{\pi}_i;\pi)$ does not depend on β_i and is finite under the full-support assumption on π_i^0. Dividing by β_i and taking $\beta_i \to \infty$ yields

$$\mathbb{E}_{Q_i(\vec{h}^{\,0:T};\vec{\mu}\,)}\left[\mathcal{U}_i(\vec{h}^{\,0:T})\right] \geq \mathbb{E}_{Q_i(\vec{h}^{\,0:T};(\hat{\pi}_i,\vec{\mu}_{\,-i}))}\left[\mathcal{U}_i(\vec{h}^{\,0:T})\right],$$

i.e., any accumulation point of the sequence satisfies the CCE no-deviation condition for player i. Since this holds for all $i \in N$ and all deviations $\hat{\pi}_i$, every accumulation point is a coarse correlated equilibrium of $\mathcal{G}$. □

Note that since the sets of pure/mixed equilibria are contained within the set of CCEs, it is straightforward to restrict the class of strategies permitted and obtain similar correspondences between modified versions of the CCFEE and the other well-known solution concepts mentioned above. Thus, the following inclusion relation also obtains:

Corollary 10.2. *For a commonly known POSG $\mathcal{G}$, we have* $\mathrm{IFEE}(\mathcal{G}) \subseteq \mathrm{NE}(\mathcal{G})$ *in the limit as $\beta_i \to \infty$, $i \in N$.*

One-Shot Games

In normal-form or one-shot games, the situation is much simpler. Each player selects a single action a_i from their action set which may be sampled from a correlated/independent policy, and utilities $u_i(a_i, a_{-i})$ are assigned according to the payoff matrix of the normal-form game based on the action profile. In this setting, free energy equilibria are equivalent to a prior-weighted logit quantal response equilibrium, as the following proposition shows.

Proposition 10.3. *In a normal-form game with baseline policies π_i^0 of full support, an independent policy profile $\vec{\pi}$ is an IFEE if and only if, for each player i,*

$$\pi_i(a_i) \propto \pi_i^0(a_i)\, \exp\left(\beta_i\, u_i(a_i; \vec{\pi}_{-i})\right), \quad u_i(a_i; \vec{\pi}_{-i}) \coloneqq \mathbb{E}_{a_{-i} \sim \vec{\pi}_{-i}}\left[u_i(a_i, a_{-i})\right].$$

In particular, as $\beta_i \to \infty$ they converge to Nash equilibria.

Proof. In a one-shot game, we have

$$G_i((\pi_i, \vec{\pi}_{-i}); \vec{\pi}^{\,0}) = \mathrm{D}_{\mathrm{KL}}\left[\pi_i \,||\, \pi_i^0\right] - \beta_i \mathbb{E}_{a_i \sim \pi_i,\, a_{-i} \sim \vec{\pi}_{-i}}\left[u_i(a_i, a_{-i})\right].$$

Now fix $\vec{\pi}_{-i}$ and minimise the PDO over policies $\pi_i \in \Delta(A_i)$. The Lagrangian is given by

$$\mathcal{L}(\pi_i, \lambda) = \sum_{a_i \in A_i} \pi_i(a_i) \log \frac{\pi_i(a_i)}{\pi_i^0(a_i)} - \beta_i \sum_{a_i \in A_i} \pi_i(a_i)\, u_i(a_i; \pi_{-i}) + \lambda \left(\sum_{a_i \in A_i} \pi_i(a_i) - 1 \right).$$

First-order conditions yield, for all a_i with $\pi_i(a_i) > 0$,

$$\log \frac{\pi_i(a_i)}{\pi_i^0(a_i)} - \beta_i u_i(a_i; \pi_{-i}) + \lambda = 0,$$

and hence we obtain

$$\pi_i(a_i) \propto \pi_i^0(a_i) \exp\left(\beta_i\, u_i(a_i; \pi_{-i})\right).$$

Normalisation gives the stated logit form. Conversely, any such fixed point of the logit best responses is an IFEE by optimality of π_i in the convex program above. The convergence to NE as $\beta_i \to \infty$ follows by the same argument as in Theorem 10.1. □

10.5 Incentive Design For and By Boundedly Rational Agents

With the theoretical foundations laid to describe the factors affecting decision making in boundedly rational agents, we now conclude by describing a formal framework outlining the implications of the models developed in Part II of this thesis for incentive design. Noting that this is a largely unexplored area of research, we attempt to outline some preliminary concepts and ideas in this direction here, in the hopes that future work will be able to build upon these foundations. In what ways might an incentive designer introduce incentives for agents?

1. The incentive designer can update the agents' world models Q_i, e.g., by providing information. This is the approach taken in Bayesian persuasion [519, 520], which typically assumes that agents update their beliefs using Bayesian updating. In light of the model of belief updating proposed in Chapter 8, a natural question to ask is how bounded rationality and belief utilities affect the ability of an incentive designer to influence an agent's beliefs and the subsequent implications of this for the structuring of information provision over time.

2. The incentive designer can update the agents' preference models $\tilde{P}_i$, which is the most commonly studied approach, including, e.g., the provision of punishments and rewards. Chapter 9 has highlighted another mechanism for influencing an agent's preferences, namely by shaping their attentional state. A full instantiation of the models of bounded rationality developed in this thesis to attentional allocation remains an open and promising direction for future work.

3. The incentive designer can change the structure of the environment. In contrast to the previous two approaches, this method does not necessarily involve the incentive designer interacting directly with the agent. Rather, the incentive designer, by changing the environment, may rely on the agent to discover for themselves that the landscape of considerations has shifted and then adjust their behaviour accordingly. This raises an interesting question about the degree to which the environment acts as an information bottleneck between the incentive designer and the agents.

The Role of Bounded Rationality Bounded rationality has implications for incentive design in two ways. Firstly, if the agents in question are boundedly rational, then the designer may need to take into consideration this fact in order to effectively elicit the desired outcomes. For example, if an employer designs a contract that is too difficult for the prospective hire to understand, or it is too complex for them to figure out what the best course of action is in light of the incentive, then the intended effects of the contract may not be realised. Additionally, if agents possess ingrained habits of mind and behaviour, then the magnitude or kinds of incentives required to induce changes in behaviour may be greater than they otherwise would be. This inertia in behaviour and belief change thus highlights the importance of the temporal aspects of incentive design, where the aim is to steer the trajectories of agent beliefs and behaviours over a given period of time rather than to immediately induce the desired outcomes.

The second way in which bounded rationality bears on the question of incentive design is that the designers themselves are often boundedly rational as well. Indeed, the problem of designing incentives is often a cognitively challenging task, and the effects of any particular incentive on the subsequent behaviour of the agents may not fully be accounted for or predicted by the principal. Moreover, as the model in Chapter 9 has highlighted, decision making is often a multifaceted problem with multiple attributes contributing in various ways to the desirability of different outcomes and courses of action. In the same way, multiple incentives can be present that may conflict with each other, resulting in what Uri Gneezy has called 'mixed signals' [521]. If not carefully thought through, mixed signals can end up rendering proposed incentive schemes ineffective or even elicit the opposite outcome from what was intended. Examples of this abound in history. For instance, governments providing bounties for their citizenry to turn in caught pests in order to reduce their populations can result in outcomes that are opposite to the intended effects of their incentives, e.g., by the citizens creating breeding farms for the very pests that the government was trying to eliminate.

This problem is particularly important in multi-agent systems. In these settings, the principal not only has to consider potential interactions between different incentives that are present, but also the interactions between the agents themselves. This makes the problem of predicting the effects of incentives on groups of agents considerably more difficult than in the case of a single agent. Because resources and energy are finite, important trade-offs are introduced between allocating effort towards gathering information to analyse the incentives

that are present in a given decision-making scenario, processing existing information to gain a better understanding of the incentives and their interactions, and designing and implementing incentives to elicit the desired outcomes.

Given the above considerations, we are led to propose the following questions for future investigation:

1. What can go wrong if incentives are designed under the assumption that agents are perfectly rational, when in fact they are boundedly rational?
2. How much better could we do if incentives are designed to take into account the bounded rationality of agents?
3. How much effort should a boundedly-rational principal invest in gathering information about, analysing, and designing incentives for a group of agents?

We hope that the models developed herein will serve as a useful framework for formalising and studying these questions.

10.6 Discussion

In summary, the model we have presented captures different degrees of rationality among different agents, allows one to explicitly model beliefs, to express more complex (e.g., path-based) objectives in a natural manner, and to model the inherent trade-offs between information-acquisition and reward maximisation that decision-makers face when embedded in complex, partially observable environments. The Free Energy Equilibrium solution concepts follow naturally from the formulation of such a model, so the question of whether they should be used over existing solution concepts depends on whether the assumptions underlying current models are sufficiently accurate for the modeller's purposes. We argue that in many scenarios, the existing assumptions can be made more realistic by adopting the framework outlined in this text, which will likely lead to better explanations of empirical data and predictions of real-world behaviour.

Our work contributes to the analysis and modelling of multi-agent interactions in several aspects. To begin with, we propose POSGs as a formalism for modelling multi-agent bounded rationality and active inference, along with a general form for modelling the predictions and preferences of boundedly rational agents

in multi-agent settings. Using this formalism, we describe several aspects of agents' bounded rationality through the lens of active inference (approximate hidden state inference, suboptimal policy selection based on information theory, "sophistication" as a model of how the agent updates its beliefs based on new observations, and how it models its own future decision making). In particular, we utilise a biased generative model as an expressive model of agents' preferences, which allows us to cast objectives as a KL divergence. Our technical results apply only to preferential models derived from POSG rewards but the model allows for the specification of more general (history-dependent) objectives which are not expressible using standard reward functions.

After introducing our model of agents and their decision making, we reintroduce the Path Divergence Objective for multi-agent settings and establish two free energy-based solution concepts, for correlated and independent policy profiles, with parameterised boundedness β_i for each of the agents. We establish relationships between the solution concepts and coarse correlated equilibrium, resp. Nash equilibrium, in the limit as $\beta_i \to \infty$ (i.e., perfectly rational agents). Finally, we study basic additional properties of the solution concepts.

Existence and Computation. In normal-form games, the existence of an IFEE is guaranteed by the equivalence established in Proposition 10.3: the logit best-response characterisation yields a continuous self-map of the compact, convex set $\bigtimes_{i \in N} \Delta(A_i)$, and an application of Brouwer's fixed-point theorem confirms that at least one fixed point (and hence at least one IFEE) always exists. For general POSGs, however, existence remains an open question. The infinite-dimensional nature of the policy space and the non-linear coupling introduced by partial observability preclude a straightforward appeal to fixed-point arguments, and establishing existence in this setting constitutes an important direction for future work. Regarding computation, iterative best-response procedures provide a natural approach, where agents repeatedly update their policies to minimise the PDO given the current policies of their opponents. In normal-form games, these dynamics are equivalent to standard logit-response learning [522]. In the sequential POSG setting, convergence guarantees are correspondingly more limited, and the design of scalable algorithms with provable properties remains a compelling open problem [523, 524].

Open problems and research directions

The framework we have introduced opens up several avenues for further development and applications. Firstly, the relaxation of assumptions regarding common knowledge of the game and other players' policies is a natural and promising direction in which to expand the present model. For instance, this opens up the possibility of studying how various learning algorithms, which have been developed for individual active inference agents, would translate to the multi-agent setting [334, 525, 526]. An immediate next step in this direction would be to study learning dynamics under Free Energy Equilibria, and how these interact with belief updating in the model-based setting. Indeed, under the fixed belief assumption, Free Energy Equilibria fall within the family of regularised policy learning dynamics, which have been extensively studied in the literature on learning in games [522, 527–534]. In general, learning dynamics in games may not necessarily converge to an equilibrium in the last iterate, and can give rise to cyclical or even chaotic behaviour, and it would be useful to explore the impact of model-based learning dynamics on this. Relatedly, developing methods to efficiently and scalably compute free energy equilibria, particularly in POSGs, is a natural extension [523, 524, 535].

Due to the inherent trade-offs between information-seeking, preference satisfaction, and dissonance-reducing actions that are induced by the divergence objectives we adopt, the theory of multi-agent active inference and free energy equilibria may find use in providing a new perspective on the microfoundations of the efficient market hypothesis [536]. By representing goal-achievement and information acquisition in a common currency using information theory, it additionally becomes possible to study how combinations of utility-based and information-based incentives can be jointly utilised to effectively and efficiently influence the behaviours of divergence-minimising agents [11].

In addition to the aforementioned future directions, we believe that a particularly promising application of this framework is in the study of the emergence of self-organisation and collective behaviour [315, 537–541]. The question of how a group of agents can organise into a collective with coherent beliefs, desires, and intentions remains a largely unsolved problem and progress on this issue would have profound consequences across the broader research community. By studying the conditions under which a collective prediction and preference model can emerge, possibly in the sense of computational/causal closure [542], from the interactions between sub-agents, it could be possible to identify additional

conditions that would also mandate divergence minimisation between the two, giving rise to a theory of how a group level super-agent emerges from a collection of interacting sub-agents in a shared environment. We hope that the model described here may be useful towards this endeavour.

11 Conclusion

11.1 Technical Contributions and Insights

Here, we recapitulate the main technical contributions and insights of this thesis. We began in Chapter 3 with a computational version of the principal-agent problem, where agents choose the values of a set of binary variables under their control and possess lexicographic preferences, giving priority to the satisfaction of a logical constraint and secondarily preferring to minimise costs. We demonstrated that a simple kind of partial observability – the hiddenness of some variables – gives rise to verifiability constraints. This is a different kind of partial observability from the standard one commonly studied in the literature, which assumes that the observed outcome is a noisy perturbation of the agent's true action. Under this setting, in order to know that a proposed incentive scheme is effective, the principal must be able to verify whether or not a particular logical condition holds in some or all Nash equilibria of the game, given only access to the observed variables. We settled the complexity of this problem in the existential and universal cases, showing them to be Σ_2^p- and Π_2^p-complete, respectively. Then, we gave a characterisation of the conditions under which a strategy profile is an inducible or eliminable equilibrium under some contract, and gave a graphical characterisation of the eliminability of a set of strategy profiles. Finally, we studied the computational complexity of the existential and universal Nash contractibility problems, showing them to be

Σ_2^p- and Σ_3^p-complete, respectively. This separation (under classical complexity-theoretic assumptions) demonstrates the increased difficulty that arises when demanding that all rational outcomes under a contract satisfy the principal's objective, rather than just some outcome.

In Chapter 4, we extended this setting into the temporal domain of concurrent games where players' primary goals are LTL formulae, and their secondary goals are to minimise a limit-average of the costs they incur along the infinite run of the system induced by the players' joint strategies, which are represented as finite state machines with output. The introduction of the temporal domain raises the question of what model we should use to represent incentives. We considered two different models: static and dynamic incentive schemes. Static taxation schemes alter the costs of different state-action profile pairs in a fixed manner, whereas dynamic taxation schemes are represented by finite state machines with output, allowing the incentives applied to vary based on the players' behaviours throughout the game. We characterised soft and hard equilibria in these games, which refer to equilibria that can or cannot be eliminated by a taxation scheme, and showed that dynamic incentive schemes are strictly more powerful than static ones for universal implementation, but not existential implementation. Following this, we considered the existential and universal Nash implementation problems, showing that for both static and dynamic incentive schemes, the existential problem is 2EXPTIME-complete. The universal problems proved much more difficult, and matching upper and lower bounds on the complexity of these problems remains open. However, our characterisation of the eliminability of an arbitrary set of strategy profiles in the case of dynamic taxation schemes is a significant step towards solving this problem.

In Chapter 5, we introduced the incentive verification and design problems in the context of weighted concurrent game structures. We provided a generalisation of the previous two problems under the expressive umbrella of quantitative temporal logic ($\mathsf{LTL}[\mathcal{F}]$), yielding a general-purpose method for automated incentive synthesis that can reason about both qualitative rules and quantitative trade-offs. This marks a step towards an idealised version of the incentive compiler, by using logic as the 'frontend' interface, and model checking/synthesis as the 'backend' engine. We introduced two different problems: incentive *verification* and incentive *synthesis*. In incentive *verification*, the aim is to verify whether the application of a given incentive scheme guarantees some level of satisfaction for a given objective in some or all stable outcomes of a game. In incentive *synthesis*, we are instead interested in finding an incentive scheme, if it exists, which guarantees

that the satisfaction value of some objective attains at least a certain level in some or all stable outcomes of a game. We showed that incentive synthesis is 2EXPTIME-complete in the static case, and that dynamic existential synthesis is likewise 2EXPTIME-complete. Only dynamic universal synthesis escapes this bound, falling in 3EXPTIME, which we conjecture to be a tight upper bound.

In Chapter 6, we introduced a variant of the Reactive Modules Games framework called *Energy Reactive Modules Games*, which equips agents with an energy endowment that evolves according to the actions they take and constrains the set of actions available to them at any given point in time, based on their energy requirements. We solved the basic rational verification problems in this model in both the non-cooperative and cooperative settings, namely the problems of determining whether a given strategy profile is a Nash equilibrium and the problems of determining whether some or all Nash equilibria/core stable outcomes in an Energy Reactive Modules Game satisfy a given specification. We showed that NASH-MEMBERSHIP is PSPACE-complete and that E-NASH, A-NASH, E-CORE, and A-CORE are all 2EXPTIME-complete, and hence no harder than their counterparts in the original Reactive Modules Games framework, while allowing the encoding of energy constraints. We then introduced an extension to this framework called *Endogenous Energy Reactive Modules Games*, which involves a pre-play stage where agents can make offers to transfer energy to other agents before the game starts. This introduces a two-stage game with additional strategic considerations, because the energy levels resulting from such transfers shape the set of stable outcomes that are possible in the subsequent ERMG. Thus, the offer stage itself constitutes a game, and we studied the decision problems of whether a profile of offers is weakly preferred to another by a given player, and whether a stable offer profile exists in both the non-cooperative and cooperative settings. We showed that all of these problems are also 2EXPTIME-complete, although a careful breakdown of the complexities with respect to different components of the input specification reveals differences in their actual runtimes.

In Chapter 7, we turned our attention to models of bounded rationality motivated by the aim of improving on some of the limitations of the idealised models of agents that we explored in Part I of the dissertation. We introduced an information-theoretic model of bounded rationality known as the *Path Divergence Objective* (PDO), which captures the trade-offs between expected utility and the informational costs associated with cognitive processes such as belief updating and policy selection, in the setting of single-agent Partially Observable Markov Decision Processes. We demonstrated a decomposition of the divergence term in

the PDO under a restriction on the agent's preferences into the sum of epistemic value, pragmatic value, and intention-behaviour gap terms. We then derived three different recursive definitions of the PDO under different assumptions about the distribution encoding an agent's preferences. We then presented an algorithm for computing a policy to minimise the PDO. Finally, we presented a simple demonstration of the differences between our objective and standard expected value maximisation, as well as variants of the expected free energy from Active Inference in the classic T-Maze experiment from neuroscience.

In Chapter 8, we considered the role of motivation and bounded rationality in processes of belief change in humans. Motivated by a desire to understand and describe the myriad ways in which human cognition can seem to go wrong from the normative Bayesian perspective, we developed a motivated variational belief change model which proposes that pragmatic considerations (i.e., utility) do not just affect policy selection but also belief change. In this model, we cast belief updating as a variational optimisation process, which is driven by a combination of what we call *affective belief utility* and the usual accuracy and complexity terms found in variational inference. We performed simple computational experiments to demonstrate how this model can qualitatively recover phenomena such as biased evidence selection and attitude polarisation, lending credence to the idea that this model can be used to provide a more complete account of human belief change.

In Chapter 9, we took a detour from the models of bounded rationality developed thus far to explore how another manifestation of bounded rationality, namely attention, can be manipulated to influence decision making. We modelled this setting as a multi-attribute, multi-alternative decision making problem where an agent must select between multiple alternatives, each characterised by a set of different attributes, and place a salience weighting distribution over the features corresponding to how much weight they place on each of the features in their final decision. We proved a representation theorem describing how salience-indexed preference relations over the alternatives can be represented using a linear form, where the salience weights are the coefficients of feature-specific valuation functions. We then introduced a decision problem that asks whether a salience distribution exists that would make a given alternative the most preferred one. We provided a polynomial-time algorithm to decide this problem. Then we introduced and demonstrated a simple Bayesian model of how a principal can learn about the agent's salience distribution and then manipulate this by

selectively mentioning certain features in order to boost their salience to the decision-maker and encourage the selection of a particular alternative.

In Chapter 10, we returned to the multi-agent setting, expanding the PDO developed in Chapter 7 to Partially Observable Stochastic Games (POSGs) with boundedly rational model-based agents. We introduced the concept of *Free Energy Equilibrium*, which is the analogue of Nash equilibrium for PDO-minimising agents. We showed how free energy equilibria are related to Nash and Coarse Correlated Equilibria, and studied the connection between free energy equilibria and prior weighted quantal response equilibria in one-shot games. We then concluded by exploring the implications of the models developed in Part II of this thesis for incentive design.

11.2 Ideas, Open Problems, and Future Work

The efforts made in this dissertation only scratch at the surface of this vast territory of the study of incentives. It is my hope that some of the ideas found in this document will provide some pointers for directions to explore further. To that end, I will elaborate here on some further speculative ideas for deeper investigation.

Scalable Incentive Design

The results presented here, particularly in Part I of this thesis, have sketched the outlines of the difficulty of solving the incentive design problem in general. While many of the results lie well outside the realm of tractability (e.g., 2EXPTIME-completeness) for many of the problems related to rational verification, we note that much of the complexity, in particular the double exponentials, comes mainly from the size of the goal formulae, which are typically small for many applications. Indeed, one of the main benefits of using logical representations such as LTL for goals is that they can be used to succinctly encode complex temporal objectives. Nonetheless, there are many assumptions made in these models that are not always justified, and it is likely that more scalable methods of incentive design can be developed by relaxing these assumptions. It is therefore worthwhile exploring how to develop more scalable methods of solving incentive design problems. As highlighted in Chapter 2, a promising candidate in this direction would be the use of machine learning techniques to learn incentive schemes, which would also facilitate adaptivity and responsiveness to changing agents and environments.

Incentives and Habits

How do incentives relate to habits? One of the hallmark features of biological learning systems is their ability to habituate to their environmental circumstances. The same is also true when different incentives are present within their environments. In such cases, the motivating effect of an incentive may diminish over time as agents become accustomed to their presence, forming expectations that the incentives will continue to apply in the future and habituating to the behaviours that the incentives were intended to bring about. In other words, today's incentive can become tomorrow's habit. Because of the role habituation plays in responses to incentives, it may thus be necessary to continually adapt incentive structures if one wishes to consistently elicit certain behaviours, which may draw on techniques from curriculum learning [543] and unsupervised environment design [544]. This shifts the focus from trying to design incentives to directly motivate agents towards achieving a particular goal to instilling desirable habits of mind and behaviour into them. If this applies on a collective level, such habituation could solidify into social norms and even formal institutions. This raises the question of what happens when an ingrained incentive structure is removed. Does it create a negative response in the incentivised agent(s) who suddenly experiences an increase in prediction error, or do the learned behaviours remain intact, obviating the need for the continued use of incentives to motivate them? When does this solidification process occur, and what are the mechanisms that dictate the degree to which it does so? The foregrounding of the role of priors in Part II of this thesis has provided a starting point for investigating these questions further, but more work is needed to understand how such priors are constructed, both at the individual and collective levels.

The Incentive Landscape Beyond Goal-Directedness

The work in this thesis has primarily focused on the incentivisation of agents towards particular desired outcomes, which is an inherently goal-directed or pragmatic perspective. However, as explored in Part II of this thesis, the attainment of goals is but one facet of the rich terrain of factors that motivate behaviour. Complementary to this is the notion of *intrinsic motivations*, which can take on a range of forms, including as we saw epistemic value driven by curiosity [363], the intention-behaviour gap [361], empowerment [545] and other forms [546, 547]. The question of how these intrinsic motivations relate to goal-attainment is still underexplored, with some positing that specific goals

subserve overarching intrinsic motivations such as free energy minimisation or empowerment maximisation. Others have highlighted that the presence of intrinsic motivations such as information-gain-driven exploration can actually help an agent to approximate the Bayes-optimal policy for a given MDP [548]. This leads one to wonder how incentive design might be expanded to take such intrinsic motivations into consideration. For example, a common technique in Reinforcement Learning is to augment the reward function with a term (such as a policy entropy regulariser) that encourages the agent to explore their environment. Might such principles yield fruitful applications in the human domain as well at both an individual and collective level? For example, the incentive landscape surrounding drug discovery could be augmented to promote exploration of the solution space to discover novel drugs and manufacturing strategies by providing rewards for the novelty of solutions alongside their efficacy.

11.3 Ethical Considerations

The power to design and deploy incentives, particularly within human-AI systems, carries significant ethical weight and responsibility. Although this thesis has focused on the technical aspects of this problem, the models developed throughout this dissertation point towards real-world structures that affect the decision making of every agent on the planet. It is hence of the utmost importance to highlight the fact that decisions around the use of such technology should not be taken in the absence of the broader ethical considerations that are necessary to ensure that the technology is used in a way that is consistent with the values, autonomy, and flourishing of those impacted by it.

The line between unwarranted manipulation and productive decision support is blurry, and it is wise not to be overly presumptuous in our judgments about where a given intervention lies on this spectrum [549]. Indeed, recent work has highlighted that an agent's plasticity, defined roughly speaking as their susceptibility to being influenced by their observations, can be in tension with their empowerment, which captures their ability to influence their surroundings or observations through their actions [550]. This property raises significant questions about the degree to which the effectiveness of incentives will come at the cost of the reduction in the influenced agent's autonomy, and thus about the desirability of doing so. Autonomy is a core value of human society that has contributed to its growth and flourishing, and it is therefore vital to consider the degree to which the use of incentives may be in conflict with this value [551].

11.4 Limitations

This thesis is subject to several limitations that will hopefully frame the next frontiers of research. Firstly, the use of formal computational or mathematical models for modelling interactions between agents can only ever be an approximation of the real world. Indeed, as outlined in the 'big world hypothesis' [552], any system interacting with the real world, including systems developed for automated incentive design, is necessarily much smaller than the world it is trying to model and influence. Therefore, models developed and implemented to this end should be used in combination with a range of other models and methodologies to inform the real-world design and analysis of incentives, and their recommendations must always be taken with a grain of salt.

Another related critical issue not addressed in this dissertation is the fact that contracts and incentives do not exist in a vacuum, but rather in a complex, historically shaped web of social norms, cultural practices, and institutional structures. These structures fill in the gaps for what might be considered appropriate or justifiable behaviours in situations that are not specified within the context of an incomplete contract/incentive scheme [553]. The contextualisation of incentive schemes by these factors turns it into a highly complex, multidimensional problem that calls for richer models and more care when designing solutions. One promising step in this direction is the development of *resource-rational contractualism* [554], which proposes to 'approximate the agreements rational parties would form by drawing on a toolbox of normatively-grounded, cognitively-inspired heuristics that trade effort for accuracy.' In practice, different problems foreground different degrees of complexity and nuance, and the right tools both in terms of accuracy and complexity should be chosen accordingly. This is especially true in the light of the computational intractability of perfectly solving incentive design problems, as seen in Part I of this thesis, leaving open the central challenge of developing principled methods for accurately assessing the degree of sophistication and levels of tolerance for error that a particular solution should be required to meet.

11.5 Closing Remarks

In summary, this thesis has set out to study computational aspects of the analysis and design of incentives for rational and boundedly rational agents. The core

motivation of this work has been to elaborate on the features of principals and agents that impact these tasks, particularly highlighting how the consideration of *costs* in various forms affects questions of incentive design. In Part I, we integrated core concepts from computer science, including logic, automata theory, and complexity theory with the foundational economic and mathematical ideas in game theory to bring a new computational perspective to bear on the study of incentives. This work has shed light on the role of observability constraints, goal specifications, incentive scheme expressivity, and inter-agent resource transfers in what can be verified, synthesised, and achieved when agents act rationally in multi-agent settings. Under idealised rationality, we can prove when incentives deliver specified temporal goals and even compile them automatically through the powerful techniques of formal verification and synthesis. This has laid the groundwork for steps towards the use of our ever-growing access to computational resources to assist in building and enhancing the interactive infrastructure of our societies in a world whose agentic landscape is rapidly evolving. In Part II, we have taken a critical perspective on one of the key assumptions made in economic analysis – that agents are rational. Through the lens of bounded rationality, we have explored how informational limitations affect the influence surfaces that are exposed to incentive designers, including mechanisms of policy updating, belief change, and attention.

A crucial challenge that remains to be addressed is how to integrate the two perspectives on incentives laid out in this dissertation into a unified framework and set of methodologies for the analysis and design of incentives in Human-AI collectives. It is my hope that via a deeper understanding of the mechanisms influencing decision making, we can collectively, thoughtfully, and responsibly steer the systems that we are building and participating in towards aligned, autonomy-preserving, and welfare-improving paths forward.

References

[1] Gillian K Hadfield and Andrew Koh. 'An economy of AI agents'. In: *arXiv preprint arXiv:2509.01063* (2025).

[2] Nenad Tomasev et al. *Virtual Agent Economies*. 2025. arXiv: 2509.10147 [cs.AI].

[3] Denizalp Goktas et al. 'Strategic Foundation Models'. Feb. 2025.

[4] Nicholas R Jennings et al. 'Human-agent collectives'. In: *Communications of the ACM* 57.12 (2014), pp. 80–88.

[5] Yingxuan Yang et al. *Agentic Web: Weaving the Next Web with AI Agents*. 2025. arXiv: 2507.21206 [cs.AI].

[6] Evan Hubinger et al. *Risks from Learned Optimization in Advanced Machine Learning Systems*. 2021. arXiv: 1906.01820 [cs.AI].

[7] Lewis Hammond et al. *Multi-Agent Risks from Advanced AI*. 2025. arXiv: 2502.14143 [cs.MA].

[8] Batu El and James Zou. *Moloch's Bargain: Emergent Misalignment When LLMs Compete for Audiences*. 2025. arXiv: 2510.06105 [cs.AI].

[9] Michael I. Jordan. *A Collectivist, Economic Perspective on AI*. 2025. arXiv: 2507.06268 [cs.CY].

[10] Christian Schroeder de Witt. 'Open challenges in multi-agent security: Towards secure systems of interacting ai agents'. In: *arXiv preprint arXiv:2505.02077* (2025).

[11] Lillian J Ratliff et al. 'A perspective on incentive design: Challenges and opportunities'. In: *Annual Review of Control, Robotics, and Autonomous Systems* 2 (2019), pp. 305–338. DOI: 10.1146/ANNUREV-CONTROL-053018-023634.

[12] Arthur C. Pigou and Nahid Aslanbeigui. *The Economics of Welfare*. Routledge, 2017. DOI: 10.4324/9781351304368.

[13] Allan Dafoe et al. 'Open problems in cooperative AI'. In: *arXiv preprint arXiv:2012.08630* (2020).

[14] Adam Smith. *An Inquiry into the Nature and Causes of the Wealth of Nations*. W. Strahan and T. Cadell, 1776.

[15] Friedrich A Hayek. 'Competition as a discovery procedure'. In: *Quarterly journal of Austrian economics* 5.3 (2002), pp. 9–23.

[16] Julian Gutierrez, Paul Harrenstein and Michael Wooldridge. 'From model checking to equilibrium checking: Reactive modules for rational verification'. In: *Artificial Intelligence* 248 (2017), pp. 123–157. DOI: 10.1016/j.artint.2017.04.003.

[17] Breed D Meyer. 'Natural and quasi-experiments in economics'. In: *Journal of business & economic statistics* 13.2 (1995), pp. 151–161.

[18] Alexander Lavin et al. *Simulation Intelligence: Towards a New Generation of Scientific Methods*. 2022. arXiv: 2112.03235 [cs.AI].

[19] Ayush Chopra. *Large Population Models*. 2025. arXiv: 2507.09901 [cs.MA].

[20] Kyle Cranmer, Johann Brehmer and Gilles Louppe. 'The frontier of simulation-based inference'. In: *Proceedings of the National Academy of Sciences* 117.48 (2020), pp. 30055–30062.

[21] Christel Baier and Joost-Pieter Katoen. *Principles of model checking*. MIT Press, 2008.

[22] Edmund M. Clarke, Thomas A. Henzinger and Helmut Veith. 'Introduction to model checking'. In: *Handbook of Model Checking*. Springer, 2018, pp. 1–26.

[23] Edmund M. Clarke, E. Allen Emerson and A. Prasad Sistla. 'Automatic verification of finite-state concurrent systems using temporal logic specifications'. In: *ACM Transactions on Programming Languages and Systems (TOPLAS)* 8.2 (1986), pp. 244–263.

[24] Brent T. Hailpern. *Verifying concurrent processes using temporal logic*. 129. Springer Science & Business Media, 1982.

[25] Robert M Keller. 'Formal verification of parallel programs'. In: *Communications of the ACM* 19.7 (1976), pp. 371–384.

[26] Susanne Graf et al. 'What are the limits of model checking methods for the verification of real life protocols?' In: *International Conference on Computer Aided Verification*. Springer. 1989, pp. 275–285.

[27] Edmund M. Clarke and E. Allen Emerson. 'Design and synthesis of synchronization skeletons using branching time temporal logic'. In: *Workshop on Logic of Programs*. Springer. 1981, pp. 52–71.

[28] Jean-Pierre Queille and Joseph Sifakis. 'Specification and verification of concurrent systems in CESAR'. In: *International Symposium on programming*. Springer. 1982, pp. 337–351.

[29] Edmund M. Clarke et al. *Handbook of Model Checking*. Vol. 10. Springer, 2018.

[30] Rocco de Nicola and Frits Vaandrager. 'Action versus state based logics for transition systems'. In: *LITP Spring School on Theoretical Computer Science*. Springer. 1990, pp. 407–419.

[31] Amir Pnueli. 'The Temporal Logic of Programs'. In: *18th Annual Symposium on Foundations of Computer Science, Providence, Rhode Island, USA, 31 October - 1 November 1977*. IEEE Computer Society, 1977, pp. 46–57. DOI: 10.1109/SFCS.1977.32.

[32] Mordechai Ben-Ari, Zohar Manna and Amir Pnueli. 'The temporal logic of branching time'. In: *Proceedings of the 8th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*. 1981, pp. 164–176.

[33] Dominique Perrin and Jean-Éric Pin. *Infinite words: automata, semigroups, logic and games*. Academic Press, 2004.

[34] E. Allen Emerson. 'Temporal and Modal Logic'. In: *Handbook of Theoretical Computer Science, Volume B: Formal Models and Semantics*. Ed. by Jan van Leeuwen. Elsevier and MIT Press, 1990, pp. 995–1072. DOI: 10.1016/b978-0-444-88074-1.50021-4.

[35] A. Prasad Sistla and Edmund M. Clarke. 'The Complexity of Propositional Linear Temporal Logics'. In: *Journal of the ACM (JACM)* 32.3 (1985), pp. 733–749. DOI: 10.1145/3828.3837.

[36] Amir Pnueli and Roni Rosner. 'On the Synthesis of an Asynchronous Reactive Module'. In: *Automata, Languages and Programming, 16th International Colloquium, ICALP89, Stresa, Italy, July 11-15, 1989, Proceedings*. Ed. by Giorgio Ausiello, Mariangiola Dezani-Ciancaglini and Simona Ronchi Della Rocca. Vol. 372. Lecture Notes in Computer Science. Springer, 1989, pp. 652–671. DOI: 10.1007/BFb0035790.

[37] E Allen Emerson and Joseph Y Halpern. '"Sometimes" and "not never" revisited: on branching versus linear time temporal logic'. In: *Journal of the ACM (JACM)* 33.1 (1986), pp. 151–178.

[38] Rajeev Alur, Thomas A. Henzinger and Orna Kupferman. 'Alternating-time temporal logic'. In: *Journal of the ACM (JACM)* 49.5 (2002), pp. 672–713. DOI: 10.1145/585265.585270.

[39] E Allen Emerson. 'Model Checking and the Mu-calculus.' In: *Descriptive Complexity and Finite Models* 31 (1996), pp. 185–214.

[40] Wolfgang Thomas. 'Automata on infinite objects'. In: *Formal Models and Semantics*. Elsevier, 1990, pp. 133–191.

[41] Moshe Y Vardi. 'Alternating automata and program verification'. In: *Computer Science Today* (1995), pp. 471–485.

[42] Willem Visser and Howard Barringer. 'CTL* model checking for SPIN'. In: *The 4th International SPIN Workshop*. 1998.

[43] Philippe Schnoebelen. 'The Complexity of Temporal Logic Model Checking.' In: *Advances in modal logic* 4.393-436 (2002), p. 35.

[44] Moshe Y Vardi. 'A temporal fixpoint calculus'. In: *Proceedings of the 15th ACM SIGPLAN-SIGACT symposium on Principles of programming languages*. 1988, pp. 250–259.

[45] Francis M. Bator. 'The Simple Analytics of Welfare Maximization'. In: *The American Economic Review* 47.1 (1957), pp. 22–59 (retrieved. 9/11/2022).

[46] J. R. Hicks. 'The Foundations of Welfare Economics'. In: *The Economic Journal* 49.196 (Dec. 1939), p. 696. DOI: 10.2307/2225023.

[47] Shaull Almagor, Udi Boker and Orna Kupferman. 'Formally Reasoning About Quality'. In: *Journal of the ACM* 63.3 (2016), 24:1–24:56. DOI: 10.1145/2875421.

[48] Patricia Bouyer et al. 'Reasoning about Quality and Fuzziness of Strategic Behaviours'. In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*. International Joint Conference on Artificial Intelligence. 2019, pp. 1588–1594.

[49] Stan Franklin and Art Graesser. 'Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents'. In: *International workshop on agent theories, architectures, and languages*. Springer. 1996, pp. 21–35.

[50] Michael J. Wooldridge and Nicholas R Jennings. 'Agent theories, architectures, and languages: a survey'. In: *International Workshop on Agent Theories, Architectures, and Languages*. Springer. 1994, pp. 1–39.

[51] Michael J. Wooldridge. *An Introduction to MultiAgent Systems*. 2nd ed. John Wiley & Sons, 2009.

[52] Anand S. Rao and Michael P. Georgeff. 'BDI agents: from theory to practice'. In: *Proceedings of the First International Conference on Multi-Agent Systems*. Vol. 95. 1995, pp. 312–319.

[53] Blair Archibald et al. 'Probabilistic BDI Agents: Actions, Plans, and Intentions'. In: *International Conference on Software Engineering and Formal Methods*. Springer. 2021, pp. 262–281.

[54] Louise A Dennis and Nir Oren. 'Explaining BDI agent behaviour through dialogue'. In: *Autonomous Agents and Multi-Agent Systems* 36.2 (2022), pp. 1–27.

[55] Sim Yee Wai et al. 'Towards Software Engineering Perspective for BDI Agent'. In: *2021 4th International Symposium on Agents, Multi-Agent Systems and Robotics*. IEEE. 2021, pp. 106–110.

[56] Herbert A. Simon. 'Theories of Bounded Rationality'. In: *Decision and Organization*. North Holland Publ., Amsterdam, 1964, pp. 161–176.

[57] Herbert A. Simon and Latsis. 'From substantive to procedural rationality'. In: *Method and Appraisal in Economics*. Cambridge University Press, 1976, pp. 129–148.

[58] Eliezer Yudkowsky. *What do we mean by "rationality"?* LessWrong. Mar. 2009.

[59] Gerd Gigerenzer. 'Axiomatic rationality and ecological rationality'. In: *Synthese* 198.4 (2021), pp. 3547–3564.

[60] John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

[61] Leonard J. Savage. *The foundations of statistics*. New York: John Wiley & Sons, 1954.

[62] Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: An introduction*. 2nd edition. MIT Press, 1998.

[63] Paul F Christiano et al. 'Deep reinforcement learning from human preferences'. In: *Advances in Neural Information Processing Systems* 30 (2017).

[64] Stefano V. Albrecht, Filippos Christianos and Lukas Schäfer. *Multi-Agent Reinforcement Learning: Foundations and Modern Approaches*. MIT Press, 2024.

[65] Karl Friston et al. 'The Free Energy Principle Made Simpler but Not Too Simple'. In: *Physics Reports*. The Free Energy Principle Made Simpler but Not Too Simple 1024 (June 2023), pp. 1–29. DOI: 10.1016/j.physrep.2023.07.001.

[66] Lancelot Da Costa et al. 'Bayesian mechanics for stationary processes'. In: *Proceedings of the Royal Society A* 477.2256 (2021), p. 20210518.

[67] Dalton AR Sakthivadivel. 'Towards a geometry and analysis for Bayesian mechanics'. In: *arXiv preprint arXiv:2204.11900* (2022).

[68] Maxwell JD Ramstead et al. 'On Bayesian mechanics: a physics of and by beliefs'. In: *Interface Focus* 13.3 (2023), p. 20220029.

[69] Karl Friston et al. 'Path integrals, particular kinds, and strange things'. In: *Physics of Life Reviews* (2023).

[70] Lancelot Da Costa et al. 'Active inference on discrete state-spaces: A synthesis'. In: *Journal of Mathematical Psychology* 99 (2020), p. 102447.

[71] Yoav Shoham and Kevin Leyton-Brown. 'Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations'. In: *Cambridge Books* (2009).

[72] Robin Milner. *Communicating and mobile systems: the pi calculus*. Cambridge University Press, 1999.

[73] Christos H. Papadimitriou et al. 'Brain computation by assemblies of neurons'. In: *Proceedings of the National Academy of Sciences* 117.25 (June 2020), pp. 14464–14472. DOI: 10.1073/pnas.2001893117.

[74] Shan Mei et al. 'Complex agent networks: An emerging approach for modeling complex systems'. In: *Applied Soft Computing* 37 (Dec. 2015), pp. 311–321. DOI: 10.1016/j.asoc.2015.08.010.

[75] Robin Milner. *The Space and Motion of Communicating Agents*. Cambridge University Press, Mar. 2009. DOI: 10.1017/cbo9780511626661.

[76] Martin Shubik. 'Game theory models and methods in political economy'. In: *Handbook of Mathematical Economics* 1 (1981), pp. 285–330.

[77] Carl Shapiro. 'The theory of business strategy'. In: *The Rand journal of economics* 20.1 (1989), pp. 125–137.

[78] Robert Aumann and Sergiu Hart. *Handbook of game theory with economic applications*. Tech. rep. Elsevier, 2002.

[79] Anthony Downs. *An Economic Theory of Democracy*. Harper & Row New York, 1957.

[80] Gilat Levy and Ronny Razin. 'It takes two: an explanation for the democratic peace'. In: *Journal of the European economic Association* 2.1 (2004), pp. 1–29.

[81] James D. Fearon. 'Rationalist explanations for war'. In: *International organization* 49.3 (1995), pp. 379–414.

[82] John Maynard Smith and George R. Price. 'The logic of animal conflict'. In: *Nature* 246.5427 (1973), pp. 15–18.

[83] John Maynard Smith and David Harper. *Animal Signals*. Oxford University Press, 2003.

[84] Jörgen W Weibull. *Evolutionary game theory*. MIT Press, 1997.

[85] Georgios Chalkiadakis, Edith Elkind and Michael J. Wooldridge. 'Computational aspects of cooperative game theory'. In: *Synthesis Lectures on Artificial Intelligence and Machine Learning* 5.6 (2011), pp. 1–168.

[86] Yoav Shoham. 'Computer science and game theory'. In: *Communications of the ACM* 51.8 (2008), pp. 74–79. DOI: 10.1145/1378704.1378721.

[87] Joseph Y. Halpern. 'Computer science and game theory'. In: *The New Palgrave Dictionary of Economics* (2008).

[88] Tim Roughgarden. 'Algorithmic game theory'. In: *Communications of the ACM* 53.7 (2010), pp. 78–86.

[89] Aggelos Kiayias et al. 'Blockchain Mining Games'. In: *Proceedings of the 2016 ACM Conference on Economics and Computation, EC '16, Maastricht, The Netherlands, July 24-28, 2016*. Ed. by Vincent Conitzer, Dirk Bergemann and Yiling Chen. ACM, 2016, pp. 365–382. DOI: 10.1145/2940716.2940773.

[90] Drew Fudenberg and Jean Tirole. *Game theory*. MIT Press, 1991.

[91] John von Neumann. 'Zur theorie der gesellschaftsspiele'. In: *Mathematische annalen* 100.1 (1928), pp. 295–320.

[92] John F. Nash. 'Equilibrium points in n-person games'. In: *Proceedings of the National Academy of Sciences* 36.1 (1950), pp. 48–49.

[93] John F. Nash. 'Non-cooperative games'. In: *Annals of mathematics* (1951), pp. 286–295.

[94] Kousha Etessami and Mihalis Yannakakis. 'On the complexity of Nash equilibria and other fixed points'. In: *SIAM Journal on Computing* 39.6 (2010), pp. 2531–2597.

[95] Constantinos Daskalakis, Paul W Goldberg and Christos H Papadimitriou. 'The complexity of computing a Nash equilibrium'. In: *SIAM Journal on Computing* 39.1 (2009), pp. 195–259.

[96] Georg Gottlob, Gianluigi Greco and Francesco Scarcello. 'Pure Nash equilibria: Hard and easy games'. In: *Proceedings of the 9th Conference on Theoretical Aspects of Rationality and Knowledge*. 2003, pp. 215–230.

[97] Robert J. Aumann. 'Subjectivity and correlation in randomized strategies'. In: *Journal of Mathematical Economics* 1.1 (1974), pp. 67–96. DOI: 10.1016/0304-4068(74)90037-8.

[98] Robert J. Aumann. 'The core of a cooperative game without side payments'. In: *Transactions of the American Mathematical Society* 98.3 (1961), pp. 539–552. DOI: 10.2307/1993348.

[99] Paul B. Harrenstein et al. 'Boolean Games'. In: *TARK*. 2001, pp. 287–298.

[100] Elise Bonzon et al. 'Boolean games revisited'. In: *ECAI*. Vol. 141. 2006, pp. 265–269.

[101] Egor Ianovski and Luke Ong. 'EGuaranteeNash for Boolean games is NEXP-hard'. In: *Fourteenth International Conference on the Principles of Knowledge Representation and Reasoning*. 2014.

[102] Sergiu Hart. 'Games in extensive and strategic forms'. In: *Handbook of game theory with economic applications* 1 (1992), pp. 19–40.

[103] Ernst Zermelo. 'Über eine Anwendung der Mengenlehre auf die Theorie des Schachspiels'. In: *Proceedings of the fifth international congress of mathematicians*. Vol. 2. Cambridge University Press Cambridge. 1913, pp. 501–504.

[104] Robert J. Aumann. *Lecture notes on game theory*. CRC Press, 1988.

[105] Ulrich Schwalbe and Paul Walker. 'Zermelo and the early history of game theory'. In: *Games and economic behavior* 34.1 (2001), pp. 123–137.

[106] Michael J. Wooldridge. 'Thinking backward with professor zermelo'. In: *IEEE Intelligent Systems* 30.2 (2015), pp. 62–67.

[107] Martin J. Osborne. *An Introduction to Game Theory*. Vol. 3. 3. Oxford University Press, 2004.

[108] Harold William Kuhn and Albert William Tucker. *Contributions to the Theory of Games*. 28. Princeton University Press, 1953.

[109] Reinhard Selten. 'Spieltheoretische behandlung eines oligopolmodells mit nachfrageträgheit: Teil i: Bestimmung des dynamischen preisgleichgewichts'. In: *Zeitschrift für die gesamte Staatswissenschaft/Journal of Institutional and Theoretical Economics* H. 2 (1965), pp. 301–324.

[110] Jean-Pierre Benoit and Vijay Krishna. 'Finitely Repeated Games'. In: *Econometrica* 53.4 (1985), pp. 905–922.

[111] Garrett Andersen and Vincent Conitzer. 'Fast equilibrium computation for infinitely repeated games'. In: *Twenty-Seventh AAAI Conference on Artificial Intelligence*. 2013.

[112] Dilip Abreu. 'On the theory of infinitely repeated games with discounting'. In: *Econometrica: Journal of the Econometric Society* (1988), pp. 383–396.

[113] Michael L. Littman and Peter Stone. 'A polynomial-time Nash equilibrium algorithm for repeated games'. In: *Proceedings of the 4th ACM Conference on Electronic Commerce*. 2003, pp. 48–54.

[114] Ehud Kalai and Ehud Lehrer. 'Rational learning leads to Nash equilibrium'. In: *Econometrica: Journal of the Econometric Society* (1993), pp. 1019–1045.

[115] Drew Fudenberg and David K. Levine. *The theory of learning in games*. Vol. 2. MIT Press, 1998.

[116] Julian Gutierrez, Paul Harrenstein and Michael J. Wooldridge. 'Iterated boolean games'. In: *Information and Computation* 242 (2015), pp. 53–79.

[117] David Gale and Frank M. Stewart. 'Infinite games with perfect information'. In: *Contributions to the Theory of Games* 2.245-266 (1953), pp. 2–16.

[118] Robert McNaughton. 'Infinite games played on finite graphs'. In: *Annals of Pure and Applied Logic* 65.2 (1993), pp. 149–184.

[119] J. Richard Buchi and Lawrence H. Landweber. 'Solving sequential conditions by finite-state strategies'. In: *The Collected Works of J. Richard Büchi*. Springer, 1990, pp. 525–541.

[120] Michael Ummels et al. 'Pure Nash Equilibria in Concurrent Deterministic Games'. In: *Logical Methods in Computer Science* 11 (2015).

[121] Julian Gutierrez et al. 'Automated temporal equilibrium analysis: Verification and synthesis of multi-player games'. In: *Artificial Intelligence* 287 (2020), p. 103353. DOI: 10.1016/j.artint.2020.103353.

[122] Michael Ummels. 'The complexity of Nash equilibria in infinite multiplayer games'. In: *International Conference on Foundations of Software Science and Computational Structures*. Springer. 2008, pp. 20–34.

[123] Erich Grädel and Michael Ummels. 'Solution concepts and algorithms for infinite multiplayer games'. In: *New Perspectives on Games and Interaction* 4 (2008), pp. 151–178.

[124] Andrzej Ehrenfeucht and Jan Mycielski. 'Positional strategies for mean payoff games'. In: *International Journal of Game Theory* 8.2 (1979), pp. 109–113. DOI: 10.1007/BF01768705.

[125] Uri Zwick and Mike Paterson. 'The complexity of mean payoff games on graphs'. In: *Theoretical Computer Science* 158.1-2 (1996), pp. 343–359.

[126] Krishnendu Chatterjee, Thomas A. Henzinger and Marcin Jurdziński. 'Mean-Payoff Parity Games'. In: *20th IEEE Symposium on Logic in Computer Science (LICS 2005), 26-29 June 2005, Chicago, IL, USA, Proceedings*. IEEE Computer Society, 2005, pp. 178–187. DOI: 10.1109/LICS.2005.26.

[127] Roderick Bloem et al. 'Better Quality in Synthesis through Quantitative Objectives'. In: *Computer Aided Verification, 21st International Conference, CAV 2009, Grenoble, France, June 26 - July 2, 2009. Proceedings*. Ed. by Ahmed Bouajjani and Oded Maler. Vol. 5643. Lecture Notes in Computer Science. Springer, 2009, pp. 140–156. DOI: 10.1007/978-3-642-02658-4_14.

[128] Julian Gutierrez et al. 'Nash equilibria in concurrent games with lexicographic preferences'. In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*. 2017.

[129] Julian Gutierrez et al. 'Equilibria for games with combined qualitative and quantitative objectives'. In: *Acta Informatica* 58.6 (2021), pp. 585–610. DOI: 10.1007/s00236-020-00385-4.

[130] Nils Bulling and Valentin Goranko. 'Combining quantitative and qualitative reasoning in concurrent multi-player games'. In: *Autonomous Agents and Multi-Agent Systems* 36 (2022), pp. 1–33.

[131] Samson Abramsky and Paul-André Mellies. 'Concurrent games and full completeness'. In: *Proceedings. 14th Symposium on Logic in Computer Science (Cat. No. PR00158)*. IEEE. 1999, pp. 431–442.

[132] Julian Gutierrez and Glynn Winskel. 'Borel determinacy of concurrent games'. In: *International Conference on Concurrency Theory*. Springer. 2013, pp. 516–530.

[133] Michael Ummels and Dominik Wojtczak. 'The complexity of Nash equilibria in limit-average games'. In: *International Conference on Concurrency Theory*. Springer. 2011, pp. 482–496. DOI: 10.1007/978-3-642-23217-6_32.

[134] Lloyd S. Shapley. 'Stochastic games'. In: *Proceedings of the national academy of sciences* 39.10 (1953), pp. 1095–1100.

[135] Krishnendu Chatterjee, Rupak Majumdar and Marcin Jurdziński. 'On Nash equilibria in stochastic games'. In: *International workshop on computer science logic*. Springer. 2004, pp. 26–40.

[136] Michael Ummels. *Stochastic multiplayer games: Theory and algorithms*. Amsterdam University Press, 2010.

[137] Krishnendu Chatterjee and Thomas A. Henzinger. 'A survey of stochastic ω-regular games'. In: *Journal of Computer and System Sciences* 78.2 (2012), pp. 394–413.

[138] Krishnendu Chatterjee, Marcin Jurdziński and Thomas A. Henzinger. 'Simple stochastic parity games'. In: *International Workshop on Computer Science Logic*. Springer. 2003, pp. 100–113.

[139] Krishnendu Chatterjee, Marcin Jurdziński and Thomas A. Henzinger. 'Quantitative stochastic parity games'. In: *Proceedings of the fifteenth annual ACM-SIAM symposium on Discrete algorithms*. Citeseer. 2004, pp. 121–130.

[140] Alessandro Abate et al. 'Rational verification: game-theoretic verification of multi-agent systems'. In: *Applied Intelligence* 51.9 (2021), pp. 6569–6584.

[141] Michael J. Wooldridge et al. 'Rational Verification: From Model Checking to Equilibrium Checking'. In: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*. Ed. by Dale Schuurmans and Michael P. Wellman. AAAI Press, 2016, pp. 4184–4191. DOI: 10.1609/aaai.v30i1.9878.

[142] Dana Fisman, Orna Kupferman and Yoad Lustig. 'Rational synthesis'. In: *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*. Springer. 2010, pp. 190–204. DOI: 10.1007/978-3-642-12002-2_16.

[143] Tong Gao, Julian Gutierrez and Michael J. Wooldridge. 'Iterated boolean games for rational verification'. In: *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*. 2017, pp. 705–713.

[144] Julian Gutierrez, Giuseppe Perelli and Michael J. Wooldridge. 'Iterated games with LDL goals over finite traces'. In: *Proceedings of the Sixteenth International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*. International Foundation for Autonomous Agents and Multiagent Systems. 2017.

[145] Rajeev Alur and Thomas A. Henzinger. 'Reactive modules'. In: *Formal methods in system design* 15.1 (1999), pp. 7–48.

[146] Wiebe van der Hoek, Alessio Lomuscio and Michael J. Wooldridge. 'On the complexity of practical ATL model checking'. In: *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems (AAMAS'06)*. Ed. by H. Nakashima et al. ACM, 2006, pp. 201–208. DOI: 10.1145/1160633.1160665.

[147] Julian Gutierrez, Thomas Steeples and Michael J. Wooldridge. 'Mean-payoff games with ω-regular specifications'. In: *Games* 13.1 (2022), p. 19.

[148] Julian Gutierrez, Giuseppe Perelli and Michael J. Wooldridge. 'Multi-player games with LDL goals over finite traces'. In: *Information and Computation* 276 (2021), p. 104555.

[149] Julian Gutierrez, Paul Harrenstein and Michael J. Wooldridge. 'Reasoning about equilibria in game-like concurrent systems'. In: *Annals of Pure and Applied Logic* 169.2 (2017), pp. 373–403. DOI: 10.1016/j.apal.2016.10.009.

[150] Julian Gutierrez et al. 'On the complexity of rational verification'. In: *Annals of Mathematics and Artificial Intelligence* 91.4 (2023), pp. 409–430. DOI: 10.1007/s10472-022-09804-3.

[151] Julian Gutierrez et al. 'Cooperative concurrent games'. In: *Artificial Intelligence* 314 (2023), p. 103806. DOI: 10.1016/j.artint.2022.103806.

[152] Michael L. Littman. 'Markov games as a framework for multi-agent reinforcement learning'. In: *Proceedings of the Eleventh International Conference on Machine Learning (ICML'94)*. Morgan Kaufmann Publishers, 1994, pp. 157–163.

[153] Kaiqing Zhang, Zhuoran Yang and Tamer Başar. 'Multi-agent reinforcement learning: A selective overview of theories and algorithms'. In: *Handbook of Reinforcement Learning and Control* (2021), pp. 321–384.

[154] Lewis Hammond et al. 'Multi-agent reinforcement learning with temporal logic specifications'. In: *arXiv preprint arXiv:2102.00582* (2021).

[155] Marta Kwiatkowska et al. 'Automated verification of concurrent stochastic games'. In: *International Conference on Quantitative Evaluation of Systems*. Springer. 2018, pp. 223–239.

[156] Marta Kwiatkowska et al. 'Correlated equilibria and fairness in concurrent stochastic games'. In: *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*. Springer. 2022, pp. 60–78.

[157] Marta Kwiatkowska et al. 'Automatic verification of concurrent stochastic systems'. In: *Formal Methods in System Design* 58.1 (2021), pp. 188–250.

[158] Marta Kwiatkowska et al. 'Equilibria-based probabilistic model checking for concurrent stochastic games'. In: *International Symposium on Formal Methods*. Springer. 2019, pp. 298–315.

[159] Rui Yan et al. 'Finite-horizon Equilibria for Neuro-symbolic Concurrent Stochastic Games'. In: *arXiv preprint arXiv:2205.07546* (2022).

[160] Rui Yan et al. 'Strategy synthesis for zero-sum neuro-symbolic concurrent stochastic games'. In: *arXiv preprint arXiv:2202.06255* (2022).

[161] Sanford J. Grossman and Oliver D. Hart. 'An analysis of the principal-agent problem'. In: *Foundations of insurance economics*. Springer, 1992, pp. 302–340. DOI: 10.1007/978-94-015-7957-5_16.

[162] Jiachen Yang et al. 'Learning to incentivize other learning agents'. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 15208–15219.

[163] Yoav Kolumbus, Joe Halpern and Éva Tardos. *Paying to Do Better: Games with Payments between Learning Agents*. 2025. arXiv: 2405.20880 [cs.GT].

[164] Philipp Obreiter and Jens Nimis. 'A taxonomy of incentive patterns'. In: *International Workshop on Agents and P2P Computing*. Springer. 2003, pp. 89–100.

[165] Roger B Myerson. 'Mechanism design'. In: *Allocation, information and markets*. Springer, 1989, pp. 191–206.

[166] Leonid Hurwicz and Stanley Reiter. *Designing economic mechanisms*. Cambridge University Press, 2006.

[167] Bengt Holmström. 'Moral hazard and observability'. In: *The Bell Journal of Economics* (1979), pp. 74–91.

[168] Bengt Holmstrom and Paul Milgrom. 'Multitask Principal–Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design'. In: *The Journal of Law, Economics, and Organization* 7.Special Issue (Jan. 1991), pp. 24–52. DOI: 10.1093/jleo/7.special_issue.24.

[169] George Georgiadis. 'Contracting with Moral Hazard: A Review of Theory & Empirics'. In: *Available at SSRN 4196247* (2022).

[170] Jean-Jacques Laffont and David Martimort. 'The theory of incentives: the principal-agent model'. In: *The theory of incentives*. Princeton University Press, 2009.

[171] George A Akerlof. 'The market for "lemons": Quality uncertainty and the market mechanism'. In: *Uncertainty in economics*. Elsevier, 1978, pp. 235–251.

[172] Mark V Pauly. 'The economics of moral hazard: comment'. In: *The american economic review* 58.3 (1968), pp. 531–537.

[173] Kenneth J Arrow. 'Insurance, risk and resource allocation'. In: *University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship* (1971).

[174] Michael C. Jensen and William H. Meckling. 'Theory of the firm: Managerial behavior, agency costs and ownership structure'. In: *Journal of Financial Economics* 3.4 (1976), pp. 305–360.

[175] Steven Shavell. 'Risk sharing and incentives in the principal and agent relationship'. In: *The Bell Journal of Economics* (1979), pp. 55–73.

[176] Bengt Holmstrom. 'Moral hazard in teams'. In: *The Bell Journal of Economics* (1982), pp. 324–340. DOI: 10.2307/3003457.

[177] Hideshi Itoh. 'Incentives to help in multi-agent situations'. In: *Econometrica: Journal of the Econometric Society* (1991), pp. 611–636.

[178] Yeon-Koo Che and Seung-Weon Yoo. 'Optimal incentives for teams'. In: *American Economic Review* 91.3 (2001), pp. 525–541.

[179] Paul Dütting, Tim Roughgarden and Inbal Talgam-Cohen. 'Simple versus optimal contracts'. In: *Proceedings of the 2019 ACM Conference on Economics and Computation*. 2019, pp. 369–387.

[180] Paul Dütting, Tim Roughgarden and Inbal Talgam-Cohen. 'The complexity of contracts'. In: *SIAM Journal on Computing* 50.1 (2021), pp. 211–254.

[181] Tal Alon, Paul Dütting and Inbal Talgam-Cohen. 'Contracts with private cost per unit-of-effort'. In: *Proceedings of the 22nd ACM Conference on Economics and Computation*. 2021, pp. 52–69.

[182] Guru Guruganesh, Jon Schneider and Joshua R Wang. 'Contracts under moral hazard and adverse selection'. In: *Proceedings of the 22nd ACM Conference on Economics and Computation*. 2021, pp. 563–582.

[183] Jiarui Gan et al. 'Optimal Coordination in Generalized Principal-Agent Problems: A Revisit and Extensions'. In: *arXiv preprint arXiv:2209.01146* (2022).

[184] Paul Dütting et al. 'Ambiguous Contracts'. In: *Econometrica* 92.6 (2024), pp. 1967–1992. DOI: 10.3982/ECTA22687.

[185] Paul Dütting, Michal Feldman and Inbal Talgam-Cohen. 'Algorithmic Contract Theory: A Survey'. In: *Foundations and Trends in Theoretical Computer Science* 16.3-4 (2024), pp. 211–411. DOI: 10.1561/0400000113.

[186] Moshe Babaioff, Michal Feldman and Noam Nisan. 'Combinatorial agency'. In: *Proceedings of the 7th ACM Conference on Electronic Commerce*. 2006, pp. 18–28.

[187] Yuval Emek and Michal Feldman. 'Computing optimal contracts in combinatorial agencies'. In: *Theoretical Computer Science* 452 (2012), pp. 56–74.

[188] Paul Dütting et al. 'Multi-Agent Contracts'. In: *Journal of the ACM* 73.2 (Apr. 2026). DOI: 10.1145/3801154.

[189] Matteo Castiglioni, Alberto Marchesi and Nicola Gatti. 'Multi-Agent Contract Design: How to Commission Multiple Agents with Individual Outcomes'. In: *Proceedings of the 24th ACM Conference on Economics and Computation*. EC '23. London, United Kingdom: Association for Computing Machinery, 2023, pp. 412–448. DOI: 10.1145/3580507.3597793.

[190] Paul Dütting et al. 'Multi-agent combinatorial contracts'. In: *Proceedings of the 2025 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM. 2025, pp. 1857–1891.

[191] R. Preston McAfee and John McMillan. 'Optimal Contracts for Teams'. In: *International Economic Review* 32.3 (1991), pp. 561–577 (retrieved. 1/10/2025).

[192] John Duggan. 'An extensive form solution to the adverse selection problem in principal/multi-agent environments'. In: *Review of Economic Design* 3.2 (1998), pp. 167–191.

[193] Theodore Groves. 'Incentives in teams'. In: *Econometrica: Journal of the Econometric Society* (1973), pp. 617–631.

[194] Sham M Kakade, Ilan Lobel and Hamid Nazerzadeh. 'Optimal dynamic mechanism design and the virtual-pivot mechanism'. In: *Operations Research* 61.4 (2013), pp. 837–854.

[195] Susan Athey and Ilya Segal. 'An efficient dynamic mechanism'. In: *Econometrica* 81.6 (2013), pp. 2463–2485.

[196] Bastien Maubert et al. 'Strategic reasoning in automated mechanism design'. In: *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*. Vol. 18. 2021, pp. 487–496. DOI: 10.24963/kr.2021/46.

[197] Munyque Mittelmann et al. 'Automated Synthesis of Mechanisms'. In: *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)*. IJCAI, 2022, pp. 426–432.

[198] Francesco Belardinelli et al. 'Reasoning about Human-Friendly Strategies in Repeated Keyword Auctions'. In: *Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems, AAMAS 2022*. IFAAMAS, 2022, pp. 62–71. DOI: 10.5555/3535850.3535859.

[199] Sai Kiran Narayanaswami et al. 'Automating Mechanism Design with Program Synthesis'. In: *Proceedings of the Adaptive and Learning Agents Workshop (ALA 2022)* (2022).

[200] Munyque Mittelmann et al. 'Formal Verification of Bayesian Mechanisms'. In: *AAAI*. AAAI Press, 2023, pp. 11621–11629.

[201] Noam Nisan and Amir Ronen. 'Algorithmic mechanism design'. In: *Proceedings of the thirty-first annual ACM symposium on Theory of computing*. 1999, pp. 129–140.

[202] Noam Nisan et al. *Algorithmic Game Theory*. Cambridge University Press, 2007.

[203] Rustam Galimullin et al. *Changing the Rules of the Game: Reasoning about Dynamic Phenomena in Multi-Agent Systems*. 2025. arXiv: 2502.11785 [cs.LO].

[204] Yagiz Savas et al. *Incentive Design for Temporal Logic Objectives*. 2019. arXiv: 1903.07752 [math.OC].

[205] Yagiz Savas, Vijay Gupta and Ufuk Topcu. *On the Complexity of Sequential Incentive Design*. 2020. arXiv: 2007.08548 [math.OC].

[206] Shaull Almagor, Guy Avni and Orna Kupferman. 'Repairing Multi-Player Games'. In: *Proceedings of the International Conference on Concurrency Theory, CONCUR 2015*. Vol. 42. LIPIcs. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2015, pp. 325–339.

[207] Julian Gutierrez et al. 'Equilibrium Design for Concurrent Games'. In: *Proceedings of the 30th International Conference on Concurrency Theory*. 2019. DOI: 10.4230/LIPIcs.CONCUR.2019.22.

[208] Julian Gutierrez et al. 'Designing Equilibria in Concurrent Games with Social Welfare and Temporal Logic Constraints'. In: *Logical Methods in Computer Science* Volume 20, Issue 4, 21 (Dec. 2024). DOI: 10.46298/lmcs-20(4:21)2024.

[209] Muhammad Najib and Giuseppe Perelli. 'Synthesis of Reward Machines for Multi-Agent Equilibrium Design'. English. In: *Proceedings of the 27th European Conference on Artificial Intelligence (ECAI-24)*. IOS Press, Oct. 2024, pp. 2733–2740. DOI: 10.3233/faia240807.

[210] Thomas Ågotnes, Wiebe van der Hoek and Michael J. Wooldridge. 'Normative system games'. In: *Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems*. New York, NY, USA: Association for Computing Machinery, 2007. DOI: 10.1145/1329125.1329284.

[211] Thomas Ågotnes et al. 'On the Logic of Normative Systems'. In: *Proceedings of the International Joint Conference on Artificial Intelligence*. 2007.

[212] Natasha Alechina et al. 'Automatic Synthesis of Dynamic Norms for Multi-Agent Systems'. In: *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*. Vol. 19. 2022, pp. 12–21.

[213] Nils Bulling and Mehdi Dastani. 'Norm-based mechanism design'. In: *Artificial Intelligence* 239 (2016), pp. 97–142. DOI: 10.1016/j.artint.2016.07.001.

[214] Xiaowei Huang et al. 'Normative Multiagent Systems: A Dynamic Generalization'. In: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. AAAI Press, 2016, pp. 1123–1129.

[215] Giuseppe Perelli. 'Enforcing equilibria in multi-agent systems'. In: *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. 2019, pp. 188–196. DOI: 10.5555/3306127.3331692.

[216] Natasha Alechina, Brian Logan and Giuseppe Perelli. 'Synthesising Minimum Cost Dynamic Norms'. In: *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*. Ed. by James Kwok. Main Track. International Joint Conferences on Artificial Intelligence Organization, Aug. 2025, pp. 3–11. DOI: 10.24963/ijcai.2025/1.

[217] Henrique Lopes Cardoso and Eugénio Oliveira. 'Adaptive Deterrence Sanctions in a Normative Framework'. In: *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*. Vol. 2. 2009, pp. 36–43. DOI: 10.1109/WI-IAT.2009.123.

[218] Davide Dell'Anna, Mehdi Dastani and Fabiano Dalpiaz. 'Runtime Revision of Sanctions in Normative Multi-Agent Systems'. In: *Autonomous Agents and Multi-Agent Systems* 34.2 (June 2020). DOI: 10.1007/s10458-020-09465-8.

[219] Tina Balke et al. 'Norms in MAS: Definitions and Related Concepts'. In: *Normative Multi-Agent Systems*. Ed. by Giulia Andrighetto et al. Vol. 4. Dagstuhl Follow-Ups. Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2013, pp. 1–31. DOI: 10.4230/DFU.Vol4.12111.1.

[220] Guido Boella, Leendert van der Torre and Harko Verhagen. 'Introduction to normative multiagent systems'. In: *Computational & Mathematical Organization Theory* 12.2 (2006), pp. 71–79.

[221] Natalia Criado, Estefania Argente and V Botti. 'Open issues for normative multi-agent systems'. In: *AI communications* 24.3 (2011), pp. 233–264.

[222] Moamin A. Mahmoud et al. 'A review of norms and normative multiagent systems'. In: *The Scientific World Journal* 2014 (2014). DOI: 10.1155/2014/684587.

[223] Antonino Rotolo and Leendert van der Torre. 'Rules, agents and norms: guidelines for rule-based normative multi-agent systems'. In: *International Workshop on Rules and Rule Markup Languages for the Semantic Web*. Springer. 2011, pp. 52–66.

[224] Michael J. Wooldridge et al. 'Incentive engineering for Boolean games'. In: *Artificial Intelligence* 195 (2013), pp. 418–439. DOI: 10.1016/j.artint.2012.11.003.

[225] Paul Harrenstein, Paolo Turrini and Michael J. Wooldridge. 'Hard and soft equilibria in boolean games'. In: *Proceedings of the Thirteenth International Conference on Autonomous Agents and Multi-Agent Systems AAMAS*. Ed. by Ana L. C. Bazzan et al. IFAAMAS/ACM, 2014, pp. 845–852. DOI: 10.5555/2615731.2615867.

[226] Paul Harrenstein, Paolo Turrini and Michael J. Wooldridge. 'Characterising the Manipulability of Boolean Games'. In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*. Ed. by Carles Sierra. IJCAI, 2017, pp. 1081–1087. DOI: 10.24963/ijcai.2017/150.

[227] Tom Everitt et al. *Understanding Agent Incentives using Causal Influence Diagrams. Part I: Single Action Settings*. 2019. arXiv: 1902.09980 [cs.AI].

[228] Ryan Carey et al. 'The incentives that shape behaviour'. In: *arXiv preprint arXiv:2001.07118* (2020).

[229] Tom Everitt et al. 'Agent Incentives: A Causal Perspective'. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.13 (May 2021), pp. 11487–11495. DOI: 10.1609/aaai.v35i13.17368.

[230] Sebastian Farquhar, Ryan Carey and Tom Everitt. 'Path-Specific Objectives for Safer Agent Incentives'. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.9 (June 2022), pp. 9529–9538. DOI: 10.1609/aaai.v36i9.21186.

[231] Ryan Carey et al. 'Incentives for responsiveness, instrumental control and impact'. In: *Artificial Intelligence* 348 (2025), p. 104408. DOI: 10.1016/j.artint.2025.104408.

[232] Francesco Bacchiocchi et al. 'Learning optimal contracts with small action spaces'. In: *Artificial Intelligence* 344 (2025), p. 104334. DOI: 10.1016/j.artint.2025.104334.

[233] Jiachen Yang et al. 'Adaptive Incentive Design with Multi-Agent Meta-Gradient Reinforcement Learning'. In: *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*. AAMAS '22. Virtual Event, New Zealand: International Foundation for Autonomous Agents and Multiagent Systems, 2022, pp. 1436–1445.

[234] Tong Mu, Stephan Zheng and Alexander Trott. *Solving Dynamic Principal-Agent Problems with a Rationally Inattentive Principal*. 2022. arXiv: 2202.01691 [cs.MA].

[235] Antoine Scheid et al. 'Incentivized learning in principal-agent bandit games'. In: *Proceedings of the 41st International Conference on Machine Learning*. ICML'24. Vienna, Austria: JMLR.org, 2024.

[236] Roberto Centeno and Holger Billhardt. 'Using incentive mechanisms for an adaptive regulation of open multi-agent systems'. In: *Proceedings of the International Joint Conference on Artificial Intelligence*. 2011. DOI: 10.5591/978-1-57735-516-8/IJCAI11-035.

[237] Dirk Bergemann et al. *Data-Driven Mechanism Design: Jointly Eliciting Preferences and Information*. 2024. arXiv: 2412.16132 [econ.TH].

[238] Paul Dütting et al. 'Optimal auctions through deep learning: Advances in differentiable economics'. In: *Journal of the ACM* 71.1 (2024), pp. 1–53.

[239] V. Udaya Sankar, Vishisht Srihari Rao and Y. Narahari. *Deep Learning Meets Mechanism Design: Key Results and Some Novel Applications*. 2024. arXiv: 2401.05683 [cs.GT].

[240] Suren Basov. 'Incentives for boundedly rational agents'. In: *Topics in Theoretical Economics* 3.1 (2003).

[241] Junyan Liu and Lillian J. Ratliff. *Principal-Agent Bandit Games with Self-Interested and Exploratory Learning Agents*. 2025. arXiv: 2412.16318 [cs.LG].

[242] Zinovi Rabinovich et al. *Cultivating Desired Behaviour: Policy Teaching Via Environment-Dynamics Tweaks*. May 2010.

[243] Guru Guruganesh et al. *Contracting with a Learning Agent*. 2024. arXiv: 2401.16198 [cs.GT].

[244] Matteo Bollini et al. *Contracting With a Reinforcement Learning Agent by Playing Trick or Treat*. 2024. arXiv: 2410.13520 [cs.GT].

[245] Jan Balaguer et al. 'The Good Shepherd: An Oracle Agent for Mechanism Design'. In: *arXiv preprint arXiv:2202.10135* (2022). DOI: 10.48550/arXiv.2202.10135.

[246] Stephan Zheng et al. 'The AI Economist: Taxation policy design via two-level deep multiagent reinforcement learning'. In: *Science Advances* 8.18 (2022), eabk2607. DOI: 10.1126/sciadv.abk2607.

[247] Tobias Baumann, Thore Graepel and John Shawe-Taylor. *Adaptive Mechanism Design: Learning to Promote Cooperation*. 2019. arXiv: 1806.04067 [cs.GT].

[248] Eric Zhao et al. 'Learning to Play General-Sum Games against Multiple Boundedly Rational Agents'. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 37.10 (June 2023), pp. 11781–11789. DOI: 10.1609/aaai.v37i10.26391.

[249] Haoxiang Ma et al. 'Adaptive Incentive Design for Markov Decision Processes with Unknown Rewards'. In: *Transactions on Machine Learning Research* 2025 (2025).

[250] David Mguni et al. 'Coordinating the Crowd: Inducing Desirable Equilibria in Non-Cooperative Systems'. In: *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. AAMAS '19. Montreal QC, Canada: International Foundation for Autonomous Agents and Multiagent Systems, 2019, pp. 386–394.

[251] David Mguni and Marcin Tomczak. 'Efficient reinforcement dynamic mechanism design'. In: *GAIW: Games, agents and incentives workshops, at AAMAS, Montreal, Canada*. 2019.

[252] Shunichi Akatsuka, Yaemi Teramoto and Aaron Courville. *Managing multiple agents by automatically adjusting incentives*. 2024. arXiv: 2409.02960 [cs.MA].

[253] Lillian J Ratliff and Tanner Fiez. 'Adaptive incentive design'. In: *IEEE Transactions on Automatic Control* 66.8 (2020), pp. 3871–3878. DOI: 10.1109/tac.2020.3027503.

[254] Dima Ivanov et al. *Principal-Agent Reinforcement Learning: Orchestrating AI Agents with Contracts*. 2024. arXiv: 2407.18074 [cs.GT].

[255] Raphael Koster et al. 'Human-centred mechanism design with Democratic AI'. In: *Nature Human Behaviour* 6.10 (2022), pp. 1398–1407.

[256] Andrea Tacchetti et al. 'Deep mechanism design: Learning social and economic policies for human benefit'. In: *Proceedings of the National Academy of Sciences* 122.25 (2025), e2319949121. DOI: 10.1073/pnas.2319949121.

[257] Xinyi Wei et al. *Active Inference through Incentive Design in Markov Decision Processes*. 2025. arXiv: 2502.07065 [eess.SY].

[258] Brian Hu Zhang et al. *Learning a Game by Paying the Agents*. 2025. arXiv: 2503.01976 [cs.GT].

[259] Paul Dütting et al. 'Combinatorial contracts'. In: *SIAM Journal on Computing* 54.6 (2025), pp. 1427–1455.

[260] Vadim Levit et al. 'Incentive-based search for equilibria in Boolean games'. In: *Constraints* 24.3 (2019), pp. 288–319.

[261] Thomas Ågotnes et al. 'Verifiable equilibria in Boolean games'. In: *Proceedings of the International Joint Conference on Artificial Intelligence*. 2013.

[262] Sofie De Clercq et al. 'Possibilistic Boolean games: Strategic reasoning under incomplete information'. In: *European Workshop on Logics in Artificial Intelligence*. Springer. 2014, pp. 196–209.

[263] Christos. H. Papadimitriou. *Computational Complexity*. Pearson, 1994.

[264] Rodica Condurache et al. 'The Complexity of Rational Synthesis'. In: *International Colloquium on Automata, Languages, and Programming, ICALP 2016*. 2016. DOI: 10.4230/LIPICS.ICALP.2016.121.

[265] Emmanuel Filiot, Raffaella Gentilini and Jean-François Raskin. 'Rational Synthesis Under Imperfect Information'. In: *Proceedings of the Annual ACM/IEEE Symposium on Logic in Computer Science, LICS 2018*. Ed. by Anuj Dawar and Erich Grädel. 2018. DOI: 10.1145/3209108.3209164.

[266] Shaull Almagor, Orna Kupferman and Giuseppe Perelli. 'Synthesis of Controllable Nash Equilibria in Quantitative Objective Game'. In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*. Ed. by Jérôme Lang. IJCAI, 2018, pp. 35–41. DOI: 10.24963/IJCAI.2018/5.

[267] Orna Kupferman and Noam Shenwald. 'The Complexity of LTL Rational Synthesis'. In: *International Conference on Tools and Algorithms for the Construction and Analysis of Systems, TACAS 2022*. Ed. by Dana Fisman and Grigore Rosu. Springer, 2022. DOI: 10.1007/978-3-030-99524-9_2.

[268] Léonard Brice, Jean-François Raskin and Marie van den Bogaard. 'Rational Verification for Nash and Subgame-Perfect Equilibria in Graph Games'. In: *Proceedings of the International Symposium on Mathematical Foundations of Computer Science, MFCS 2023*. Ed. by Jérôme Leroux, Sylvain Lombardy and David Peleg. 2023. DOI: 10.4230/LIPICS.MFCS.2023.26.

[269] David C Parkes et al. 'Dynamic incentive mechanisms'. In: *Ai Magazine* 31.4 (2010), pp. 79–94. DOI: 10.1609/aimag.v31i4.2316.

[270] GJ Olsder. 'Phenomena in inverse Stackelberg games, part 1: Static problems'. In: *Journal of optimization theory and applications* 143 (2009), pp. 589–600.

[271] Thomas A. Weber. *Optimal control theory with applications in economics*. MIT Press, 2011.

[272] David Hyland, Julian Gutierrez and Michael J. Wooldridge. 'Incentive Engineering for Concurrent Games'. In: *Proceedings of the Conference on Theoretical Aspects of Rationality and Knowledge (TARK 2023)*. Ed. by Rineke Verbrugge. Vol. 379. EPTCS. 2023, pp. 344–358. DOI: 10.4204/EPTCS.379.28.

[273] Ittai Abraham, Danny Dolev and Joseph Y Halpern. 'Lower bounds on implementing robust and resilient mediators'. In: *Theory of Cryptography: Fifth Theory of Cryptography Conference*. Springer. 2008, pp. 302–319.

[274] Patricia Bouyer et al. 'Reasoning about Quality and Fuzziness of Strategic Behaviors'. In: *ACM Transactions on Computational Logic* 24.3 (2023), 21:1–21:38. DOI: 10.1145/3582498.

[275] Orna Kupferman, Giuseppe Perelli and Moshe Y. Vardi. 'Synthesis with rational environments'. In: *Annals of Mathematics and Artificial Intelligence* 78.1 (2016), pp. 3–20. DOI: 10.1007/S10472-016-9508-8.

[276] Antonín Kucera and Jan Strejček. 'The stuttering principle revisited'. In: *Acta Informatica* 41.7–8 (2005), pp. 415–434.

[277] Julian Gutierrez, Giuseppe Perelli and Michael J. Wooldridge. 'Imperfect information in Reactive Modules games'. In: *Information and Computation* 261 (2018), pp. 650–675. DOI: 10.1016/j.ic.2018.02.023.

[278] Véronique Bruyère, Jean-François Raskin and Clément Tamines. 'Pareto-Rational Verification'. In: *Proceedings of the 33rd International Conference on Concurrency Theory (CONCUR 2022)*. 2022.

[279] Matthew O Jackson and Simon Wilkie. 'Endogenous games and mechanisms: Side payments among players'. In: *The Review of Economic Studies* 72.2 (2005), pp. 543–566.

[280] Sofie De Clercq et al. 'Formalizing Commitment-Based Deals in Boolean Games'. In: *Proceedings of the 22nd European Conference on Artificial Intelligence (ECAI 2016)*. Vol. 285. Frontiers in Artificial Intelligence and Applications. IOS Press, 2016, pp. 329–337. DOI: 10.3233/978-1-61499-672-9-329.

[281] Valentin Goranko and Paolo Turrini. 'Two-Player Preplay Negotiation Games with Conditional Offers'. In: *IGTR* 18.1 (2016), 1550017:1–1550017:31. DOI: 10.1142/S0219198915500176.

[282] Valentin Goranko. 'Preplay Negotiations with Unconditional Offers of Side Payments in Two-Player Strategic-Form Games: Towards Non-Cooperative Cooperation'. In: *Mathematics* 10.14 (2022), p. 2518.

[283] Ludovic Renou. 'Commitment games'. In: *Games and Economic Behavior* 66.1 (2009), pp. 488–505.

[284] Paolo Turrini. 'Endogenous games with goals: side-payments among goal-directed agents'. In: *Autonomous Agents and Multi-Agent Systems* 30 (2016), pp. 765–792.

[285] Arindam Chakrabarti et al. 'Resource interfaces'. In: *International Workshop on Embedded Software*. Springer. 2003, pp. 117–133.

[286] Patricia Bouyer et al. 'Infinite runs in weighted timed automata with energy constraints'. In: *FORMATS'08*. Springer. 2008, pp. 33–47.

[287] Krishnendu Chatterjee and Laurent Doyen. 'Energy parity games'. In: *Theoretical Computer Science* 458 (2012), pp. 49–60.

[288] Nir Piterman. 'From Nondeterministic B$\acute{}$uchi and Streett Automata to Deterministic Parity Automata'. In: *Logical Methods in Computer Science* 3 (2007).

[289] Loıc Hélouët, Nicolas Markey and Ritam Raha. 'Reachability games with relaxed energy constraints'. In: *Inf. Comput.* 285.Part (2022), p. 104806. DOI: 10.1016/j.ic.2021.104806.

[290] Gal Amram et al. 'Efficient Algorithms for Omega-Regular Energy Games'. In: *FM'21*. Ed. by Marieke Huisman, Corina S. Pasareanu and Naijun Zhan. Vol. 13047. Lecture Notes in Computer Science. Springer, 2021, pp. 163–181. DOI: 10.1007/978-3-030-90870-6_9.

[291] Yaron Velner et al. 'The complexity of multi-mean-payoff and multi-energy games'. In: *Information and Computation* 241 (2015), pp. 177–196.

[292] Orna Kupferman and Naama Shamash Halevy. 'Energy Games with Resource-Bounded Environments'. In: *Proceedings of the 33rd International Conference on Concurrency Theory (CONCUR 2022)*. Vol. 243. LIPIcs. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2022, 19:1–19:23. DOI: 10.4230/LIPIcs.CONCUR.2022.19.

[293] Bastien Maubert et al. 'Towards a Tool for LTL Synthesis with Bounded-Energy Constraints'. In: *ICTCS'19*. Vol. 2504. CEUR Workshop Proceedings. CEUR-WS.org, 2019, pp. 229–234.

[294] Natasha Alechina and Brian Logan. 'State of the Art in Logics for Verification of Resource-Bounded Multi-Agent Systems'. In: *Fields of Logic and Computation III - Essays Dedicated to Yuri Gurevich on the Occasion of His 80th Birthday*. Vol. 12180. Lecture Notes in Computer Science. Springer, 2020, pp. 9–29. DOI: 10.1007/978-3-030-48006-6_2.

[295] Dario Della Monica and Aniello Murano. 'Parity-Energy ATL for Qualitative and Quantitative Reasoning in MAS'. In: *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018)*. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2018, pp. 1441–1449.

[296] Natasha Alechina et al. 'Resource-bounded alternating-time temporal logic'. In: *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2010)*. 2010, pp. 481–488.

[297] Hoang Nga Nguyen et al. 'Alternating-time temporal logic with resource bounds'. In: *J. Log. Comput.* 28.4 (2018), pp. 631–663. DOI: 10.1093/logcom/exv034.

[298] Natasha Alechina et al. 'Model-checking for Resource-Bounded ATL with production and consumption of resources'. In: *J. Comput. Syst. Sci.* 88 (2017), pp. 126–144. DOI: 10.1016/j.jcss.2017.03.008.

[299] Natasha Alechina et al. 'On the complexity of resource-bounded logics'. In: *Theoretical Computer Science* 750 (2018), pp. 69–100.

[300] Muhammad Najib. 'Rational verification in multi-agent systems'. PhD thesis. University of Oxford, 2019.

[301] Paul Harrenstein, Paolo Turrini and Michael Wooldridge. 'Electric Boolean games: redistribution schemes for resource-bounded agents'. In: *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2015)*. 2015, pp. 655–663.

[302] Youssouf Oualhadj and Nicolas Troquard. 'Rational verification in iterated electric boolean games'. In: *SR'16*. Vol. 218. Open Publishing Association, 2016, p. 11.

[303] Nils Bulling and Berndt Farwer. 'On the (Un-)Decidability of Model Checking Resource-Bounded Agents'. In: *ECAI'10*. Vol. 215. Frontiers in Artificial Intelligence and Applications. IOS Press, 2010, pp. 567–572. DOI: 10.3233/978-1-60750-606-5-567.

[304] Michael Maschler, Eilon Solan and Shmuel Zamir. 'Game Theory'. In: Cambridge University Press, 2013. Chap. Utility theory, pp. 9–38.

[305] Natasha Alechina, Stéphane Demri and Brian Logan. 'Parameterised resource-bounded ATL'. In: *AAAI'20*. Vol. 34. 2020, pp. 7040–7046.

[306] C Robert Mesle. *Process-relational philosophy: an introduction to Alfred North Whitehead*. Templeton Foundation Press, 2009.

[307] 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.

[308] Herbert A. Simon. 'A behavioral model of rational choice'. In: *The quarterly journal of economics* (1955), pp. 99–118.

[309] Eva Deli, James Peters and Zoltán Kisvárday. 'The thermodynamics of cognition: a mathematical treatment'. In: *Computational and Structural Biotechnology Journal* 19 (2021), pp. 784–793.

[310] Chris Fields, Adam Goldstein and Lars Sandved-Smith. 'Making the Thermodynamic Cost of Active Inference Explicit'. In: *Entropy* 26.8 (2024), p. 622.

[311] Morten L Kringelbach, Yonatan Sanz Perl and Gustavo Deco. 'The Thermodynamics of Mind'. In: *Trends in Cognitive Sciences* (2024).

[312] David H. Wolpert et al. 'Is stochastic thermodynamics the key to understanding the energy costs of computation?' In: *Proceedings of the National Academy of Sciences* 121.45 (2024), e2321112121. DOI: 10.1073/pnas.2321112121.

[313] Pedro A. Ortega and Daniel A. Braun. 'Thermodynamics as a theory of decision-making with information-processing costs'. In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 469.2153 (2013), p. 20120683.

[314] Pedro A Ortega et al. 'Information-theoretic bounded rationality'. In: *arXiv preprint arXiv:1512.06789* (2015).

[315] Michael Kirchhoff et al. 'The Markov Blankets of Life: Autonomy, Active Inference and the Free Energy Principle'. In: *Journal of The Royal Society Interface* 15.138 (Jan. 2018), p. 20170792. DOI: 10.1098/rsif.2017.0792.

[316] Judea Pearl. *Causality*. 2nd edition. Cambridge, U.K. ; New York: Cambridge University Press, Sept. 2009.

[317] Daniel A. Braun and Pedro A. Ortega. 'Information-theoretic bounded rationality and ε-optimality'. In: *Entropy* 16.8 (2014), pp. 4662–4676.

[318] Naftali Tishby and Daniel Polani. 'Information theory of decisions and actions'. In: *Perception-action cycle: Models, architectures, and hardware*. Springer, 2010, pp. 601–636.

[319] Thomas Parr, Giovanni Pezzulo and Karl J. Friston. *Active inference: the free energy principle in mind, brain, and behavior*. MIT Press, 2022.

[320] Bartosz Maćkowiak, Filip Matějka and Mirko Wiederholt. 'Rational inattention: A review'. In: *Journal of Economic Literature* 61.1 (2023), pp. 226–273.

[321] Filip Matějka and Alisdair McKay. 'Rational inattention to discrete choices: A new foundation for the multinomial logit model'. In: *American Economic Review* 105.1 (2015), pp. 272–298.

[322] Christopher A. Sims. 'Implications of rational inattention'. In: *Journal of monetary Economics* 50.3 (2003), pp. 665–690.

[323] Rahul Bhui, Lucy Lai and Samuel J Gershman. 'Resource-rational decision making'. In: *Current Opinion in Behavioral Sciences* 41 (2021), pp. 15–21.

[324] Thomas L Griffiths, Falk Lieder and Noah D Goodman. 'Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic'. In: *Topics in cognitive science* 7.2 (2015), pp. 217–229.

[325] Marcel Binz and Eric Schulz. 'Modeling human exploration through resource-rational reinforcement learning'. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 31755–31768.

[326] Mark K Ho et al. 'The efficiency of human cognition reflects planned information processing'. In: *Proceedings of the 34th AAAI conference on artificial intelligence*. 2020.

[327] Richard L Lewis, Andrew Howes and Satinder Singh. 'Computational rationality: Linking mechanism and behavior through bounded utility maximization'. In: *Topics in cognitive science* 6.2 (2014), pp. 279–311.

[328] Samuel J Gershman, Eric J Horvitz and Joshua B Tenenbaum. 'Computational rationality: A converging paradigm for intelligence in brains, minds, and machines'. In: *Science* 349.6245 (2015), pp. 273–278.

[329] Karl Friston. 'Life as we know it'. In: *Journal of the Royal Society Interface* 10.86 (2013), p. 20130475.

[330] Karl Friston et al. 'Sophisticated inference'. In: *Neural Computation* 33.3 (2021), pp. 713–763. DOI: 10.1162/neco_a_01351.

[331] Karl J. Friston et al. 'Action and Behavior: A Free-Energy Formulation'. In: *Biological Cybernetics* 102.3 (2010), pp. 227–260. DOI: 10.1007/s00422-010-0364-z.

[332] Beren Millidge, Anil Seth and Christopher Buckley. 'Understanding the origin of information-seeking exploration in probabilistic objectives for control'. In: *arXiv preprint arXiv:2103.06859* (2021).

[333] Théophile Champion et al. 'Reframing the Expected Free Energy: Four Formulations and a Unification'. In: *arXiv preprint arXiv:2402.14460* (2024).

[334] Karl Friston et al. 'Active inference and learning'. In: *Neuroscience and Biobehavioral Reviews* 68 (2016), pp. 862–879. DOI: 10.1016/j.neubiorev.2016.06.022.

[335] Beren Millidge, Alexander Tschantz and Christopher L Buckley. 'Whence the expected free energy?' In: *Neural Computation* 33.2 (2021), pp. 447–482.

[336] Danijar Hafner et al. 'Action and perception as divergence minimization'. In: *arXiv preprint arXiv:2009.01791* (2020).

[337] Matthew Botvinick and Marc Toussaint. 'Planning as inference'. In: *Trends in cognitive sciences* 16.10 (2012), pp. 485–488.

[338] Tuomas Haarnoja et al. 'Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor'. In: *International conference on machine learning*. PMLR. 2018, pp. 1861–1870.

[339] Tomasz Korbak, Ethan Perez and Christopher L Buckley. 'RL with KL penalties is better viewed as Bayesian inference'. In: *arXiv preprint arXiv:2205.11275* (2022).

[340] Sergey Levine. 'Reinforcement learning and control as probabilistic inference: Tutorial and review'. In: *arXiv preprint arXiv:1805.00909* (2018).

[341] Ran Wei et al. 'A Unified View on Solving Objective Mismatch in Model-Based Reinforcement Learning'. In: *Transactions on Machine Learning Research* (2024).

[342] Abbas Abdolmaleki et al. 'Model-based relative entropy stochastic search'. In: *Advances in Neural Information Processing Systems* 28 (2015).

[343] Dilip Arumugam et al. 'Bayesian Reinforcement Learning With Limited Cognitive Load'. In: *Open Mind* 8 (2024), pp. 395–438. DOI: 10.1162/opmi_a_00132.

[344] Yinlam Chow et al. 'Variational model-based policy optimization'. In: *arXiv preprint arXiv:2006.05443* (2020).

[345] Jan Peters, Katharina Mulling and Yasemin Altun. 'Relative entropy policy search'. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 24. 1. 2010, pp. 1607–1612.

[346] Thomas A. Berrueta, Allison Pinosky and Todd D Murphey. 'Maximum diffusion reinforcement learning'. In: *Nature Machine Intelligence* (2024), pp. 1–11.

[347] Nicola Catenacci Volpi and Daniel Polani. 'Goal-directed empowerment: combining intrinsic motivation and task-oriented behavior'. In: *IEEE Transactions on Cognitive and Developmental Systems* 15.2 (2020), pp. 361–372.

[348] Karl Johan Åström. 'Optimal Control of Markov Processes with Incomplete State Information'. In: *Journal of Mathematical Analysis and Applications* 10.1 (Feb. 1965), pp. 174–205. DOI: 10.1016/0022-247X(65)90154-X (retrieved. 4/1/2022).

[349] Rodrigo Toro Icarte et al. 'Using reward machines for high-level task specification and decomposition in reinforcement learning'. In: *International Conference on Machine Learning*. PMLR. 2018, pp. 2107–2116.

[350] Matthew J Beal. 'Variational algorithms for approximate Bayesian inference'. PhD thesis. UCL (University College London), 2003.

[351] David M Blei, Alp Kucukelbir and Jon D McAuliffe. 'Variational inference: A review for statisticians'. In: *Journal of the American statistical Association* 112.518 (2017), pp. 859–877.

[352] Thomas Parr et al. 'Cognitive effort and active inference'. In: *Neuropsychologia* 184 (2023), p. 108562. DOI: 10.1016/j.neuropsychologia.2023.108562.

[353] Rolf Landauer. 'Irreversibility and Heat Generation in the Computing Process'. In: *IBM Journal of Research and Development* 5.3 (July 1961), pp. 183–191. DOI: 10.1147/rd.53.0183.

[354] Lars-Göran Mattsson and Jörgen W Weibull. 'Probabilistic choice and procedurally bounded rationality'. In: *Games and Economic Behavior* 41.1 (2002), pp. 61–78.

[355] Alessandro Barp et al. 'Geometric methods for sampling, optimization, inference, and adaptive agents'. In: *Handbook of Statistics*. Vol. 46. Elsevier, 2022, pp. 21–78.

[356] Lancelot Da Costa et al. 'Active inference as a model of agency'. In: *arXiv preprint arXiv:2401.12917* (2024).

[357] Edward Deci and Richard M. Ryan. *Intrinsic Motivation and Self-Determination in Human Behavior*. Perspectives in Social Psychology. New York: Springer US, 1985. DOI: 10.1007/978-1-4899-2271-7 (retrieved. 28/10/2019).

[358] Jürgen Schmidhuber. 'Formal Theory of Creativity, Fun, and Intrinsic Motivation (1990–2010)'. In: *IEEE Transactions on Autonomous Mental Development* 2.3 (Sept. 2010), pp. 230–247. DOI: 10.1109/TAMD.2010.2056368.

[359] Ethan S Bromberg-Martin and Ilya E Monosov. 'Neural circuitry of information seeking'. In: *Current Opinion in Behavioral Sciences* 35 (2020), pp. 62–70.

[360] Caroline J Charpentier and Irene Cogliati Dezza. 'Information-seeking in the brain'. In: *The Drive for Knowledge: The Science of Human Information Seeking* (2022), pp. 195–216.

[361] Mark Conner and Paul Norman. 'Understanding the intention-behavior gap: The role of intention strength'. In: *Frontiers in Psychology* 13 (2022), p. 923464.

[362] Lars Sandved-Smith and Lancelot Da Costa. 'Metacognitive particles, mental action and the sense of agency'. In: *arXiv preprint arXiv:2405.12941* (2024).

[363] Karl Friston et al. 'Active inference and epistemic value'. In: *Cognitive neuroscience* 6.4 (2015), pp. 187–214.

[364] Thomas Parr, Karl Friston and Peter Zeidman. 'Active Data Selection and Information Seeking'. In: *Algorithms* 17.3 (2024), p. 118.

[365] Noor Sajid et al. 'Active inference, Bayesian optimal design, and expected utility'. In: *The Drive for Knowledge: The Science of Human Information Seeking* (2022), pp. 124–146.

[366] David Cohn, Les Atlas and Richard Ladner. 'Improving generalization with active learning'. In: *Machine learning* 15 (1994), pp. 201–221.

[367] Dennis V Lindley. 'On a measure of the information provided by an experiment'. In: *The Annals of Mathematical Statistics* 27.4 (1956), pp. 986–1005.

[368] David JC MacKay. 'Information-based objective functions for active data selection'. In: *Neural computation* 4.4 (1992), pp. 590–604.

[369] Joel Lehman and Kenneth O. Stanley. 'Abandoning objectives: Evolution through the search for novelty alone'. In: *Evolutionary computation* 19.2 (2011), pp. 189–223.

[370] Leon Festinger. 'A Theory of Cognitive Dissonance'. In: *Stanford University Press* (1957).

[371] Eddie Harmon-Jones and Judson Mills. 'An introduction to cognitive dissonance theory and an overview of current perspectives on the theory'. In: (2019). DOI: 10.1037/0000135-001.

[372] Roope Oskari Kaaronen. 'A theory of predictive dissonance: Predictive processing presents a new take on cognitive dissonance'. In: *Frontiers in psychology* 9 (2018), p. 2218.

[373] Conor Heins et al. 'pymdp: A Python library for active inference in discrete state spaces'. In: *Journal of Open Source Software* 7.73 (2022), p. 4098. DOI: 10.21105/joss.04098.

[374] David S Olton. 'Mazes, maps, and memory.' In: *American psychologist* 34.7 (1979), p. 583.

[375] Karl Friston et al. 'Active Inference: A Process Theory'. In: *Neural Computation* 29.1 (2017), pp. 1–49. DOI: 10.1162/NECO_a_00912.

[376] Philipp Schwartenbeck et al. 'Exploration, Novelty, Surprise, and Free Energy Minimization'. In: *Frontiers in Psychology* 4 (2013), p. 710. DOI: 10.3389/fpsyg.2013.00710. pmid: 24109469.

[377] Hugo Mercier and Dan Sperber. *The enigma of reason*. Cambridge, MA: Harvard University Press, 2017.

[378] Eric Mandelbaum. 'Troubles with Bayesianism: An introduction to the psychological immune system'. In: *Mind & Language* 34.2 (2019), pp. 141–157. DOI: 10.1111/mila.12205.

[379] Matt Jones and Bradley C Love. 'Pinning down the theoretical commitments of Bayesian cognitive models'. In: *Behavioral and Brain Sciences* 34.4 (2011), pp. 215–231.

[380] Raymond S. Nickerson. 'Confirmation Bias: A Ubiquitous Phenomenon in Many Guises'. In: *Review of General Psychology* 2.2 (1998), pp. 175–220. DOI: 10.1037/1089-2680.2.2.175.

[381] Ziva Kunda. 'The case for motivated reasoning.' In: *Psychological bulletin* 108.3 (1990), p. 480.

[382] Charles Lord, Lee Ross and Mark Lepper. 'Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence'. In: *Journal of Personality and Social Psychology* 37 (Nov. 1979), pp. 2098–2109. DOI: 10.1037/0022-3514.37.11.2098.

[383] Ullrich Ecker et al. 'The psychological drivers of misinformation belief and its resistance to correction'. In: *Nature Reviews Psychology* 1 (Jan. 2022), pp. 13–29. DOI: 10.1038/s44159-021-00006-y.

[384] Arie W Kruglanski, Katarzyna Jasko and Karl Friston. 'All thinking is 'wishful' thinking'. In: *Trends in Cognitive Sciences* 24.6 (2020), pp. 413–424.

[385] J. H Priniski, Prachi Solanki and Zachary Horne. *A Bayesian decision-theoretic framework for studying motivated reasoning*. Oct. 2022. DOI: 10.31234/osf.io/ngavz.

[386] Tali Sharot et al. 'Why and when beliefs change'. In: *Perspectives on Psychological Science* 18.1 (2023), pp. 142–151.

[387] Tali Sharot and Neil Garrett. 'Forming Beliefs: Why Valence Matters'. In: *Trends in Cognitive Sciences* 20.1 (2016), pp. 25–33. DOI: 10.1016/j.tics.2015.11.002.

[388] Julian Kiverstein, Mark Miller and Erik Rietveld. 'Desire and Motivation in Predictive Processing: An Ecological-Enactive Perspective'. In: *Review of Philosophy and Psychology* (2024), pp. 1–21.

[389] George Deane et al. *The computational unconscious: Adaptive narrative control, psychopathology, and subjective well-being*. Jan. 2024. DOI: 10.31234/osf.io/x7aew.

[390] Jian-Qiao Zhu et al. 'Computation-limited Bayesian updating'. In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 45. 45. 2023.

[391] David H Wolpert. 'The free energy requirements of biological organisms; implications for evolution'. In: *Entropy* 18.4 (2016), p. 138.

[392] Bjarne Andresen. 'Current trends in finite-time thermodynamics'. In: *Angewandte Chemie International Edition* 50.12 (2011), pp. 2690–2704.

[393] Juan MR Parrondo, Jordan M Horowitz and Takahiro Sagawa. 'Thermodynamics of information'. In: *Nature physics* 11.2 (2015), pp. 131–139.

[394] Peter Salamon et al. 'More stages decrease dissipation in irreversible step processes'. In: *Entropy* 25.3 (2023), p. 539.

[395] Horace B. Barlow. 'Possible principles underlying the transformation of sensory messages'. In: *Sensory communication* 1.01 (1961), pp. 217–233.

[396] Cristina Bicchieri and Hugo Mercier. *Norms and Beliefs: How Change Occurs*. Oxford: Oxford University Press, 2014.

[397] Daniel Williams. 'Socially adaptive belief'. In: *Mind & Language* 36.3 (2021), pp. 333–354. DOI: 10.1111/mila.12294.

[398] Mahault Albarracin et al. 'Epistemic communities under active inference'. In: *Entropy* 24.4 (2022), p. 476.

[399] Axel Constant et al. 'Regimes of Expectations: An Active Inference Model of Social Conformity and Human Decision Making'. In: *Frontiers in Psychology* 10 (2019), p. 679. DOI: 10.3389/fpsyg.2019.00679.

[400] Samuel PL Veissière et al. 'Thinking through other minds: A variational approach to cognition and culture'. In: *Behavioral and brain sciences* 43 (2020), e90.

[401] Jared Vasil et al. 'A World Unto Itself: Human Communication as Active Inference'. In: *Frontiers in Psychology* 11 (2020), p. 417. DOI: 10.3389/fpsyg.2020.00417.

[402] Mahault Albarracin et al. 'Feeling our place in the world: An active inference account of self-esteem'. In: *Neuroscience of Consciousness* 2024.1 (2024), niae007. DOI: 10.1093/nc/niae007.

[403] Avel Guénin-Carlut and Mahault Albarracin. 'On Embedded Normativity: An Active Inference Account of Agency Beyond Flesh'. In: *Active Inference*. Vol. 1630. Communications in Computer and Information Science. Cham, Switzerland: Springer Nature, 2024, pp. 91–105. DOI: 10.1007/978-3-031-47958-8_7.

[404] Nabil Bouizegarene et al. 'Narrative as active inference: an integrative account of cognitive and social functions in adaptation'. In: *Frontiers in Psychology* 15 (2024), p. 1345480. DOI: 10.3389/fpsyg.2024.1345480.

[405] Trevor Spelman et al. 'Overestimating the social costs of political belief change'. In: *Journal of Experimental Social Psychology* 105 (2023), p. 104115. DOI: 10.1016/j.jesp.2023.104115.

[406] Axel Westerwick, Sarita B. Kleinman and Silvia Knobloch-Westerwick. 'Confirmation bias in online searches: Impacts of selective exposure before an election on political attitude strength and shifts'. In: *Journal of Communication* 67.4 (2017), pp. 660–684.

[407] Genís Prat-Ortega and Jaime de la Rocha. 'Selective attention: A plausible mechanism underlying confirmation bias'. In: *Current Biology* 28.19 (2018), R1151–R1154.

[408] Bharath Chandra Talluri et al. 'Confirmation bias through selective overweighting of choice-consistent evidence'. In: *Current Biology* 28.19 (2018), pp. 3128–3135.

[409] Richard Patterson, Joachim T Operskalski and Aron K Barbey. 'Motivated explanation'. In: *Frontiers in human neuroscience* 9 (2015), p. 559.

[410] Peter H Ditto, David A Pizarro and David Tannenbaum. 'Motivated moral reasoning'. In: *Psychology of learning and motivation* 50 (2009), pp. 307–338.

[411] Shailendra Pratap Jain and Durairaj Maheswaran. 'Motivated reasoning: A depth-of-processing perspective'. In: *Journal of Consumer Research* 26.4 (2000), pp. 358–371.

[412] Andrew T Little. 'How to Distinguish Motivated Reasoning from Bayesian Updating'. In: *Political Behavior* (2025), pp. 1–25.

[413] Charlie Pilgrim et al. 'Confirmation bias emerges from an approximation to Bayesian reasoning'. In: *Cognition* 245 (2024), p. 105693.

[414] Larry M. Bartels. 'Beyond the Running Tally: Partisan Bias in Political Perceptions'. In: *Political Behavior* 24.2 (2002), pp. 117–150 (retrieved. 6/6/2025).

[415] Dan Simon and Stephen J Read. 'Toward a general framework of biased reasoning: Coherence-based reasoning'. In: *Perspectives on Psychological Science* 20.3 (2025), pp. 421–459.

[416] Aileen Oeberst and Roland Imhoff. 'Toward parsimony in bias research: A proposed common framework of belief-consistent information processing for a set of biases'. In: *Perspectives on Psychological Science* 18.6 (2023), pp. 1464–1487.

[417] Aileen Oeberst, Dorothee Mischkowski and Roland Imhoff. 'Belief-Consistent Information Processing or Coherence-Based Reasoning: Integrating Two Parsimonious Frameworks for Biases'. In: *European Journal of Social Psychology* (2025).

[418] Karl Friston. 'The free-energy principle: a unified brain theory?' In: *Nature reviews neuroscience* 11.2 (2010), pp. 127–138.

[419] Martin J. Wainwright and Michael I. Jordan. 'Graphical Models, Exponential Families, and Variational Inference'. In: *Foundations and Trends in Machine Learning* 1.1–2 (2008), pp. 1–305.

[420] Lancelot Da Costa et al. 'Reward Maximization Through Discrete Active Inference'. In: *Neural Computation* 35.5 (2023), pp. 807–852.

[421] Amitai Shenhav. 'The affective gradient hypothesis: An affect-centered account of motivated behavior'. In: *Trends in Cognitive Sciences* (2024).

[422] Eli Sennesh and Maxwell Ramstead. 'An Affective-Taxis Hypothesis for Alignment and Interpretability'. In: *arXiv preprint arXiv:2505.17024* (2025).

[423] Mateus Joffily and Giorgio Coricelli. 'Emotional valence and the free-energy principle'. In: *PLoS computational biology* 9.6 (2013), e1003094.

[424] Pablo Fernandez Velasco and Slawa Loev. 'Affective experience in the predictive mind: a review and new integrative account'. In: *Synthese* 198.11 (2021), pp. 10847–10882.

[425] Michael SA Graziano. *Consciousness and the social brain*. Oxford University Press, USA, 2013.

[426] Richard C Jeffrey. 'A Note on the Kinematics of Preference'. In: *Erkenntnis* (1977), pp. 135–141.

[427] Jérôme Lang and Leendert van der Torre. 'Preference change triggered by belief change: A principled approach'. In: *International Conference on Logic and the Foundations of Game and Decision Theory*. Springer. 2008, pp. 86–111.

[428] Tom Schonberg and Leor N Katz. 'A neural pathway for nonreinforced preference change'. In: *Trends in Cognitive Sciences* 24.7 (2020), pp. 504–514.

[429] Daniel Kahneman. 'Maps of Bounded Rationality: Psychology for Behavioral Economics'. In: *The American Economic Review* 93.5 (2003), pp. 1449–1475 (retrieved. 21/7/2026).

[430] Reinhard Selten. 'Aspiration adaptation theory'. In: *Journal of mathematical psychology* 42.2-3 (1998), pp. 191–214.

[431] Jonathan Shaki, Yonatan Aumann and Sarit Kraus. 'Voter Priming Campaigns: Strategies, Equilibria, and Algorithms'. In: *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI-25)*. 2025.

[432] Alanna Gillis and Renee Ryberg. 'Is choosing a major choosing a career or interesting courses? An investigation into college students' orientations for college majors and their stability'. In: *Journal of Postsecondary Student Success* 1.2 (2021), pp. 46–71.

[433] Richard Bradley. 'Becker's thesis and three models of preference change'. In: *Politics, Philosophy & Economics* 8.2 (2009), pp. 223–242.

[434] David Strohmaier and Michael Messerli. *Preference change*. Cambridge University Press, 2024.

[435] Carlos Alós-Ferrer, Alexander Jaudas and Alexander Ritschel. 'Attentional shifts and preference reversals: An eye-tracking study'. In: *Judgment and Decision Making* 16.1 (2021), pp. 57–93. DOI: 10.1017/S1930297500008305.

[436] Gene M Heyman, Katherine A Grisanzio and Victor Liang. 'Introducing a method for calculating the allocation of attention in a cognitive "Two-Armed Bandit" procedure: probability matching gives way to maximizing'. In: *Frontiers in Psychology* 7 (2016), p. 223.

[437] Peter C. Fishburn. 'Utility theory for decision making'. In: *Publications in operations research/Operations Research Society of America (ISSN 0079-7723* 18 (1970).

[438] Gerard Debreu, Gerard Debreu and Werner Hildenbrand. 'Topological methods in cardinal utility theory'. In: *Mathematical Economics: Twenty Papers of Gerard Debreu*. Econometric Society Monographs. Cambridge University Press, 1983, pp. 120–132.

[439] Peter Wakker. 'The algebraic versus the topological approach to additive representations'. In: *Journal of Mathematical Psychology* 32.4 (1988), pp. 421–435.

[440] Stelios H Zanakis et al. 'Multi-attribute decision making: A simulation comparison of select methods'. In: *European journal of operational research* 107.3 (1998), pp. 507–529.

[441] Detlof Von Winterfeldt and Gregory W Fischer. 'Multi-attribute utility theory: models and assessment procedures'. In: *Utility, Probability, and Human Decision Making: Selected Proceedings of an Interdisciplinary Research Conference, Rome, 3–6 September, 1973*. Springer. 1975, pp. 47–85.

[442] Pedro Bordalo, Nicola Gennaioli and Andrei Shleifer. 'Salience theory of choice under risk'. In: *The Quarterly journal of economics* 127.3 (2012), pp. 1243–1285.

[443] Pedro Bordalo, Nicola Gennaioli and Andrei Shleifer. 'Salience'. In: *Annual Review of Economics* 14.1 (2022), pp. 521–544.

[444] Natasha Alechina, Fenrong Liu and Brian Logan. 'Minimal preference change'. In: *Logic, Rationality, and Interaction: 4th International Workshop, LORI 2013, Hangzhou, China, October 9-12, 2013, Proceedings 4*. Springer. 2013, pp. 15–26.

[445] Jan Chomicki and Joyce Song. 'Monotonic and nonmonotonic preference revision'. In: *arXiv preprint cs/0503092* (2005).

[446] Michael Freund. 'Revising preferences and choices'. In: *Journal of Mathematical Economics* 41.3 (2005), pp. 229–251.

[447] Sven Ove Hansson. 'Changes in preference'. In: *Theory and Decision* 38 (1995), pp. 1–28.

[448] Jérôme Lang and Leendert van der Torre. 'From belief change to preference change'. In: *ECAI 2008*. IOS Press, 2008, pp. 351–355.

[449] Johan Van Benthem and Fenrong Liu. 'Dynamic logic of preference upgrade'. In: *Journal of Applied Non-Classical Logics* 17.2 (2007), pp. 157–182.

[450] Richard Bradley. 'Preference kinematics'. In: *Preference Change: Approaches from Philosophy, Economics and Psychology*. Springer, 2009, pp. 221–242.

[451] Eran Hanany and Peter Klibanoff. 'Updating ambiguity averse preferences'. In: *The BE Journal of Theoretical Economics* 9.1 (2009), p. 000010220219351704 1547.

[452] Franz Dietrich and Christian List. 'A model of non-informational preference change'. In: *Journal of theoretical politics* 23.2 (2011), pp. 145–164.

[453] Franz Dietrich and Christian List. 'Where do preferences come from?' In: *International Journal of Game Theory* 42 (2013), pp. 613–637.

[454] Gerd Gigerenzer and Daniel G Goldstein. 'Reasoning the fast and frugal way: models of bounded rationality.' In: *Psychological review* 103.4 (1996), p. 650.

[455] Gerard Debreu. *Representation of a preference ordering by a numerical function*. Yale University, 1953.

[456] John C Harsanyi. 'Games with incomplete information played by "Bayesian" players, I–III Part I. The basic model'. In: *Management science* 14.3 (1967), pp. 159–182.

[457] Luisa M Zintgraf et al. 'Ordered preference elicitation strategies for supporting multi-objective decision making'. In: *arXiv preprint arXiv:1802.07606* (2018).

[458] Roman Linne et al. 'Sequential information processing in persuasion'. In: *Frontiers in Psychology* 13 (2022), p. 902230.

[459] Kathryn Chaloner and Isabella Verdinelli. 'Bayesian experimental design: A review'. In: *Statistical science* (1995), pp. 273–304.

[460] Guillem R Esber and Mark Haselgrove. 'Reconciling the influence of predictiveness and uncertainty on stimulus salience: a model of attention in associative learning'. In: *Proceedings of the Royal Society B: Biological Sciences* 278.1718 (2011), pp. 2553–2561.

[461] Verena Tiefenbeck et al. 'Overcoming salience bias: How real-time feedback fosters resource conservation'. In: *Management science* 64.3 (2018), pp. 1458–1476.

[462] Deborah H Schenk. 'Exploiting the salience bias in designing taxes'. In: *Yale J. on Reg.* 28 (2011), p. 253.

[463] Daniel Kahneman. 'Maps of Bounded Rationality: Psychology for Behavioral Economics'. In: *The American Economic Review* 93.5 (2003), pp. 1449–1475 (retrieved. 31/5/2024).

[464] Eric A Hansen, Daniel S Bernstein and Shlomo Zilberstein. 'Dynamic programming for partially observable stochastic games'. In: *AAAI*. Vol. 4. 2004, pp. 709–715.

[465] Richard D McKelvey and Thomas R Palfrey. 'Quantal response equilibria for normal form games'. In: *Games and economic behavior* 10.1 (1995), pp. 6–38.

[466] Alexander A. Alemi. *Information Theory for Representation Learning*. NeurIPS. 2023.

[467] Benjamin Patrick Evans and Mikhail Prokopenko. 'A maximum entropy model of bounded rational decision-making with prior beliefs and market feedback'. In: *Entropy* 23.6 (2021), p. 669.

[468] Sebastian Gottwald and Daniel A Braun. 'Systems of bounded rational agents with information-theoretic constraints'. In: *Neural computation* 31.2 (2019), pp. 440–476.

[469] Brian W Rogers, Thomas R Palfrey and Colin F Camerer. 'Heterogeneous quantal response equilibrium and cognitive hierarchies'. In: *Journal of Economic Theory* 144.4 (2009), pp. 1440–1467.

[470] Richard D McKelvey and Thomas R Palfrey. 'Quantal response equilibria for extensive form games'. In: *Experimental economics* 1 (1998), pp. 9–41.

[471] David H. Wolpert. 'Information theory—the bridge connecting bounded rational game theory and statistical physics'. In: *Complex Engineered Systems: Science meets technology*. Springer, 2006, pp. 262–290.

[472] David H. Wolpert. 'What information theory says about bounded rational best response'. In: *The Complex Networks of Economic Interactions: Essays in Agent-Based Economics and Econophysics*. Springer, 2006, pp. 293–306.

[473] Leonid Hurwicz. 'The Design of Mechanisms for Resource Allocation'. In: *The American Economic Review* 63.2 (1973), pp. 1–30 (retrieved. 30/5/2024).

[474] Roger B. Myerson. 'Optimal Auction Design'. In: *Mathematics of Operations Research* 6.1 (1981), pp. 58–73 (retrieved. 30/5/2024).

[475] J.C. Harsanyi and R. Selten. *A General Theory of Equilibrium Selection in Games*. MIT Press Classics. MIT Press, 1988.

[476] Eliezer Yudkowsky. *Inadequate Equilibria: Where and How Civilizations Get Stuck*. Machine Intelligence Research Institute, 2017.

[477] Maxwell JD Ramstead et al. 'Multiscale integration: beyond internalism and externalism'. In: *Synthese* 198.Suppl 1 (2021), pp. 41–70.

[478] Aviad Rubinstein. 'Inapproximability of Nash equilibrium'. In: *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*. 2015, pp. 409–418.

[479] Martin Zinkevich et al. 'Regret minimization in games with incomplete information'. In: *Advances in neural information processing systems* 20 (2007).

[480] Noam Brown and Tuomas Sandholm. 'Superhuman AI for multiplayer poker'. In: *Science* 365.6456 (2019), pp. 885–890.

[481] Dustin Morrill et al. 'Hindsight and sequential rationality of correlated play'. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 6. 2021, pp. 5584–5594.

[482] Brian Hu Zhang et al. 'Optimal correlated equilibria in general-sum extensive-form games: Fixed-parameter algorithms, hardness, and two-sided column-generation'. In: *Proceedings of the 23rd ACM conference on economics and computation*. 2022, pp. 1119–1120.

[483] Benjamin Eysenbach et al. 'Mismatched No More: Joint Model-Policy Optimization for Model-Based RL'. In: *ArXiv* abs/2110.02758 (2021).

[484] Thomas M. Moerland, Joost Broekens and Catholijn M. Jonker. 'Model-based Reinforcement Learning: A Survey'. In: *Foundations and Trends in Machine Learning* 16 (2020), pp. 1–118.

[485] Xihuai Wang, Zhicheng Zhang and Weinan Zhang. 'Model-based multi-agent reinforcement learning: Recent progress and prospects'. In: *arXiv preprint arXiv:2203.10603* (2022).

[486] Debajyoti Ray et al. 'Bayesian model of behaviour in economic games'. In: *Advances in neural information processing systems* 21 (2008).

[487] Muthukumaran Chandrasekaran, Yingke Chen and Prashant Doshi. 'On markov games played by bayesian and boundedly-rational players'. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 31. 1. 2017.

[488] Rowan Hodson, Marishka Mehta and Ryan Smith. 'The empirical status of predictive coding and active inference'. In: *Neuroscience and Biobehavioral Reviews* (2023), p. 105473.

[489] Mahault Albarracin et al. 'Shared Protentions in Multi-Agent Active Inference'. In: *Entropy* 26.4 (2024), p. 303.

[490] Karl Friston and Christopher Frith. 'A duet for one'. In: *Consciousness and cognition* 36 (2015), pp. 390–405.

[491] Karl J. Friston et al. 'Federated inference and belief sharing'. In: *Neuroscience & Biobehavioral Reviews* 156 (2024), p. 105500. DOI: 10.1016/j.neubiorev.2023.105500.

[492] Daphne Demekas, Conor Heins and Brennan Klein. 'An Analytical Model of Active Inference in the Iterated Prisoner's Dilemma'. In: *International Workshop on Active Inference*. Springer. 2023, pp. 145–172.

[493] Daniel Ari Friedman et al. 'Active Inferants: an active inference framework for ant colony behavior'. In: *Frontiers in behavioral neuroscience* 15 (2021), p. 647732.

[494] Georgiy Levchuk et al. 'Active inference in multiagent systems: context-driven collaboration and decentralized purpose-driven team adaptation'. In: *Artificial Intelligence for the Internet of Everything*. Elsevier, 2019, pp. 67–85.

[495] Domenico Maisto, Francesco Donnarumma and Giovanni Pezzulo. 'Interactive inference: a multi-agent model of cooperative joint actions'. In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* (2023).

[496] Jan Pöppel, Sebastian Kahl and Stefan Kopp. 'Resonating minds – emergent collaboration through hierarchical active inference'. In: *Cognitive Computation* 14.2 (2022), pp. 581–601.

[497] Shaun Gallagher and Micah Allen. 'Active inference, enactivism and the hermeneutics of social cognition'. In: *Synthese* 195.6 (2018), pp. 2627–2648.

[498] Inês Hipólito and Thomas van Es. 'Enactive-dynamic social cognition and active inference'. In: *Frontiers in Psychology* 13 (2022), p. 855074.

[499] Wako Yoshida, Ray J Dolan and Karl J Friston. 'Game theory of mind'. In: *PLoS computational biology* 4.12 (2008), e1000254.

[500] Karl Friston et al. 'Knowing one's place: a free-energy approach to pattern regulation'. In: *Journal of the Royal Society Interface* 12.105 (2015), p. 20141383.

[501] Karl J Friston et al. 'Designing ecosystems of intelligence from first principles'. In: *Collective Intelligence* 3.1 (2024), p. 26339137231222481. DOI: 10.1177/26339137231222481.

[502] Conor Heins et al. 'Spin glass systems as collective active inference'. In: *International Workshop on Active Inference*. Springer. 2022, pp. 75–98.

[503] Conor Heins et al. 'Collective behavior from surprise minimization'. In: *arXiv preprint arXiv:2307.14804* (2023).

[504] Rafael Kaufmann, Pranav Gupta and Jacob Taylor. 'An active inference model of collective intelligence'. In: *Entropy* 23.7 (2021), p. 830.

[505] Maxwell J.D. Ramstead et al. 'Neural and phenotypic representation under the free-energy principle'. In: *Neuroscience and Biobehavioral Reviews* 120 (2021), pp. 109–122.

[506] Casper Hesp et al. 'A multi-scale view of the emergent complexity of life: A free-energy proposal'. In: *Evolution, development and complexity: multiscale evolutionary models of complex adaptive systems*. Springer. 2019, pp. 195–227.

[507] Franz Kuchling et al. 'Morphogenesis as Bayesian inference: A variational approach to pattern formation and control in complex biological systems'. In: *Physics of life reviews* 33 (2020), pp. 88–108.

[508] Ensor Rafael Palacios et al. 'On Markov blankets and hierarchical self-organisation'. In: *Journal of Theoretical Biology* 486 (2020), p. 110089.

[509] Maxwell JD Ramstead et al. 'Variational ecology and the physics of sentient systems'. In: *Physics of life Reviews* 31 (2019), pp. 188–205.

[510] Matthew Sims. 'How to count biological minds: symbiosis, the free energy principle, and reciprocal multiscale integration'. In: *Synthese* 199.1-2 (2021), pp. 2157–2179.

[511] Zafeirios Fountas et al. 'Deep active inference agents using Monte-Carlo methods'. In: *Advances in neural information processing systems* 33 (2020), pp. 11662–11675.

[512] Domenico Maisto et al. 'Active inference tree search in large POMDPs'. In: *arXiv preprint arXiv:2103.13860* (2021).

[513] Vojtěch Kovařík et al. 'Rethinking formal models of partially observable multiagent decision making'. In: *Artificial Intelligence* 303 (2022), p. 103645.

[514] Vojtěch Kovařík et al. 'Value functions for depth-limited solving in zero-sum imperfect-information games'. In: *Artificial Intelligence* 314 (2023), p. 103805.

[515] Robert Aumann and Adam Brandenburger. 'Epistemic conditions for Nash equilibrium'. In: *Econometrica: Journal of the Econometric Society* (1995), pp. 1161–1180.

[516] Jakob N Foerster et al. 'Learning with opponent-learning awareness'. In: *arXiv preprint arXiv:1709.04326* (2017).

[517] He He et al. 'Opponent modeling in deep reinforcement learning'. In: *International conference on machine learning*. PMLR. 2016, pp. 1804–1813.

[518] Xiaopeng Yu et al. 'Model-based opponent modeling'. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 28208–28221.

[519] Emir Kamenica and Matthew Gentzkow. 'Bayesian persuasion'. In: *American Economic Review* 101.6 (2011), pp. 2590–2615.

[520] Emir Kamenica. 'Bayesian persuasion and information design'. In: *Annual Review of Economics* 11.1 (2019), pp. 249–272.

[521] Uri Gneezy. *Mixed signals: How incentives really work*. New Haven: Yale University Press, 2023.

[522] Carlos Alós-Ferrer and Nick Netzer. 'The logit-response dynamics'. In: *Games and Economic Behavior* 68.2 (2010), pp. 413–427. DOI: 10.1016/j.geb.2009.08.004.

[523] Tyler Becker and Zachary Sunberg. 'Bridging the Gap between Partially Observable Stochastic Games and Sparse POMDP Methods'. In: *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*. AAMAS '25. Detroit, MI, USA: International Foundation for Autonomous Agents and Multiagent Systems, 2025, pp. 2434–2436.

[524] Jonathon Schwartz. 'Towards Scalable and Robust Decision Making in Partially Observable, Multi-Agent Environments'. In: *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*. AAMAS '23. London, United Kingdom: International Foundation for Autonomous Agents and Multiagent Systems, 2023, pp. 2996–2998.

[525] Karl J. Friston et al. 'Supervised structure learning'. In: *ArXiv* abs/2311.10300 (2023).

[526] Noor Sajid, Panagiotis Tigas and Karl John Friston. 'Active inference, preference learning and adaptive behaviour'. In: *IOP Conference Series: Materials Science and Engineering* 1261 (2022).

[527] David Balduzzi et al. 'The mechanics of n-player differentiable games'. In: *International Conference on Machine Learning*. PMLR. 2018, pp. 354–363.

[528] Victor Boone and Panayotis Mertikopoulos. *The equivalence of dynamic and strategic stability under regularized learning in games*. 2023. arXiv: 2311.02407 [cs.GT].

[529] Pierre-Louis Cauvin, Davide Legacci and Panayotis Mertikopoulos. *The impact of uncertainty on regularized learning in games*. 2025. arXiv: 2506.13286 [cs.GT].

[530] Zongkai Liu et al. *Regularization is Enough for Last-Iterate Convergence in Zero-Sum Games*. 2024.

[531] Mingyang Liu et al. *The Power of Regularization in Solving Extensive-Form Games*. 2025. arXiv: 2206.09495 [cs.GT].

[532] Panayotis Mertikopoulos and William H. Sandholm. 'Learning in Games via Reinforcement and Regularization'. In: *Mathematics of Operations Research* 41.4 (2016), pp. 1297–1324 (retrieved. 11/10/2025).

[533] Samuel Sokota et al. 'A Unified Approach to Reinforcement Learning, Quantal Response Equilibria, and Two-Player Zero-Sum Games'. In: *Deep Reinforcement Learning Workshop NeurIPS 2022*. 2022.

[534] Vasilis Syrgkanis et al. 'Fast convergence of regularized learning in games'. In: *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'15. Montreal, Canada: MIT Press, 2015, pp. 2989–2997.

[535] Christopher Solinas et al. *History Filtering in Imperfect Information Games: Algorithms and Complexity*. 2023. arXiv: 2311.14651 [cs.GT].

[536] Eugene F. Fama. 'Efficient Capital Markets: A Review of Theory and Empirical Work'. In: *The Journal of Finance* 25.2 (1970), pp. 383–417 (retrieved. 25/2/2024).

[537] James B Falandays et al. *All Intelligence is Collective Intelligence*. Dec. 2022. DOI: 10.31234/osf.io/jhrp6.

[538] Michael Levin. 'The Computational Boundary of a "Self": Developmental Bioelectricity Drives Multicellularity and Scale-Free Cognition'. In: *Frontiers in Psychology* 10 (2019).

[539] Michael Levin. 'Technological Approach to Mind Everywhere: An Experimentally-Grounded Framework for Understanding Diverse Bodies and Minds'. In: *Frontiers in Systems Neuroscience* 16 (2021).

[540] Michael Levin. 'Bioelectric networks: the cognitive glue enabling evolutionary scaling from physiology to mind'. In: *Animal Cognition* 26 (2023), pp. 1865–1891.

[541] Léo Pio-Lopez et al. 'Active inference, morphogenesis, and computational psychiatry'. In: *Frontiers in Computational Neuroscience* 16 (2022).

[542] Fernando E Rosas et al. 'Software in the natural world: A computational approach to emergence in complex multi-level systems'. In: *arXiv preprint arXiv:2402.09090* (2024).

[543] Yoshua Bengio et al. 'Curriculum learning'. In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ICML '09. Montreal, Quebec, Canada: Association for Computing Machinery, 2009, pp. 41–48. DOI: 10.1145/1553374.1553380.

[544] Michael Dennis et al. *Emergent Complexity and Zero-shot Transfer via Unsupervised Environment Design*. 2021. arXiv: 2012.02096 [cs.LG].

[545] Alexander S. Klyubin, Daniel Polani and Chrystopher L. Nehaniv. 'Empowerment: a universal agent-centric measure of control'. In: *2005 IEEE Congress on Evolutionary Computation*. Vol. 1. 2005, pp. 128–135. DOI: 10.1109/CEC.2005.1554676.

[546] Martin Biehl et al. 'Expanding the active inference landscape: more intrinsic motivations in the perception-action loop'. In: *Frontiers in neurorobotics* 12 (2018), p. 45.

[547] Alex B Kiefer. 'Intrinsic motivation as constrained entropy maximization'. In: *Entropy* 27.4 (2025), p. 372.

[548] Ran Wei. *Value of Information and Reward Specification in Active Inference and POMDPs*. 2024. arXiv: 2408.06542 [cs.AI].

[549] Philipp Koralus. *The Philosophic Turn for AI Agents: Replacing centralized digital rhetoric with decentralized truth-seeking*. 2025. arXiv: 2504.18601 [cs.CY].

[550] David Abel et al. *Plasticity as the Mirror of Empowerment*. 2025. arXiv: 2505.10361 [cs.AI].

[551] Cass R. Sunstein. *The Ethics of Influence: Government in the Age of Behavioral Science*. Cambridge Studies in Economics, Choice, and Society. Cambridge University Press, 2016.

[552] Khurram Javed and Richard S Sutton. 'The big world hypothesis and its ramifications for artificial intelligence'. In: *Finding the Frame: An RLC Workshop for Examining Conceptual Frameworks*. 2024.

[553] Dylan Hadfield-Menell and Gillian Hadfield. *Incomplete Contracting and AI Alignment*. 2018. arXiv: 1804.04268 [cs.AI].

[554] Sydney Levine et al. *Resource Rational Contractualism Should Guide AI Alignment*. 2025. arXiv: 2506.17434 [cs.AI].