TEESSIDE UNIVERSITY

Doctoral Dissertation
Doctoral Program in Computer Science

Cost Effective Interventions in Complex Networks using Agent-Based Modelling and Simulations

By

Theodor Cimpeanu

******

Supervisor(s):
Prof. The Anh Han, Supervisor

Doctoral Examination Committee:
Dr Long Tran Thanh, External Examiner, Associate Professor at the University of Warwick
Dr Claudio Angione, Internal Examiner, Associate Professor at Teesside University
Dr Faik Hamad, Independent Chair, Associate Professor at Teesside University

Teesside University
2022

Declaration

I hereby declare that, the contents and organization of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

Theodor Cimpeanu
2022

* This dissertation is presented in partial fulfillment of the requirements for a **Ph.D. degree** at Teesside University.

To the first machine able to comprehend these words

Acknowledgements

Above all, I thank you, the reader, for dedicating your precious time to my story.

I would begin by thanking my supervisor, The Anh Han, for his endless and contagious enthusiasm for science, for his help and his guidance, and for becoming my mentor and friend during my time under his wing. Alongside him, I am grateful to Xico for his precious advice and wondrous discussions. Thank you to my colleagues and collaborators Alessandro, Tom, Luís, Simon and Bianca. Special thanks to Cedric and Marcus, for all the board game nights, especially for the ones featuring sanctimonious security guards.

Thank you to my housemates, Alya, Iulia and all the Andreis, for the bottomless hoard of laughter and entertainment that we conjured together. I am thankful to everyone in the Student Ambassador team, especially for the chats that made us forget about the freezing wind and rain during open days.

I am forever grateful to my family for always being there for me. To Daria, for being one of the greatest little sisters ever, even though I was a horrible brother. To my parents, for giving me bragging rights in the "who has the better parents" discussions. *Multumesc, Buni, pentru cea mai frumoasa copilarie.* To Gabriel Moiceanu, for teaching me how to dream big and travel the world on two wheels. To Budi, who is more brother than friend. To Liz, I will love you, always. Words cannot express how happy I am to know you are all in my life. I am in awe of your continuous patience with my antics.

Thank you, everyone. While reflecting upon my words here, I have thought of so many others who deserve my appreciation. Please know that I am grateful to you, but that words have failed me and that I was not able to honour my memory of you to the extent that you deserve. I hope you will forgive me next time we meet, drinks are on me.

Abstract

Humans have developed considerable machinery to create policy and to distribute incentives, forming institutions designed to foster pro-social behaviour in society. Constantly faced with decisions on how best to allocate funding, institutions wrestle with limited budgets and a demand for positive outcomes. These issues are compounded when we consider that real populations are diverse in nature and social structure, in which certain individuals are more connected or influential than others. Understanding the complex interplay between institutional incentives and social diversity can shed light on human behaviour, allowing us to build mechanisms capable of engineering pro-sociality in social systems.

In this thesis, we develop mathematical and computational formulations to explore the evolutionary dynamics and cost-efficiency of institutional incentives. We achieve this through a systematic study of several economic games and by varying the networks of interaction that underpin these settings. We start by (i) exploring a cooperative dilemma, testing whether previous findings accrued in homogeneous populations still apply in the presence of social diversity. Subsequently, we study (ii) the dynamics and evolution of fairness, using an asymmetric interaction paradigm. When interactions are asymmetric, participants can enact multiple roles, and external decision-makers or institutions must also consider which roles are more suitable candidates for incentives. Moving away from positive incentives, we then propose (iii) an original model of signalling the threat of punishment, asking whether evolutionary dynamics can explain the deterrence of anti-social behaviour by way of fear, and whether such costly signalling can be used as cost-saving measures for institutions. Finally, we suggest (iv) a timely

application domain for our findings, the regulation of advanced technology with potential safety concerns, such as Artificial Intelligence (AI).

Through extensive computer simulations and mathematical analysis, we conclude that, with respect to (i), interference in complex networks is not trivial and that no tailored response can fit all networks, but also that reckless interference can lead to the collapse of cooperation. Regarding (ii), strictly targeting specific roles is an effective way of fostering fairness, and that social diversity relaxes these requirements, reducing the burden on institutions. Concerning (iii), signalling the threat of punishment can lead to the evolution of fearful defectors, deterring anti-social behaviour and heightening social welfare. Finally, pertaining to (iv), technology governance and regulation may profit from the world's patent heterogeneity and inequality among firms and nations. This can enable the design and implementation of meticulous interventions on a minority of participants which is capable of influencing an entire population towards an ethical and sustainable use of AI.

Contents

List of Figures

List of Tables

1 | Introduction

Human social institutions can affect the course of human evolution. Just as climate, food supply, predators, and other natural forces of selection have molded our nature, so too can our culture.

—Peter Singer, *The Expanding Circle*

In which the general topics of this thesis are presented, and directions to walk through the next chapters are described.

1.1 Cooperation

Cooperation can be described as a behaviour whereby individuals work together, providing benefits not only to themselves, but also to the society as a whole. Often, by cooperation we assume there exists an opposite, selfish or competitive action. Cooperation pervades at all scales of biological life [Axelrod, 1984; Fehr and Fischbacher, 2003; Nowak, 2006b, 2012; Raihani, 2021; Rand and Nowak, 2013; Trivers, 1971]. Beyond single cells giving up the ability to reproduce, leading to multi-cellular organisms [Michod, 2007], or even insect societies, which Darwin regarded as "one special difficulty, which first appeared to me insuperable, and actually fatal to my theory [of natural selection]" [Darwin, 1911; Herbers, 2009]. Even beyond human gestures, such as offering directions to a stranger who seems lost, giving up your seat on a crowded bus, and donating to charity or to a blood bank, cooperation underlies complex social behaviours and the origin of human societies [Smith and Szathmary, 1995].

"The evolution of cooperation" was listed as one of the ten most challenging problems of the century in the New Year 2000 edition of Science [Sigmund, 2010]. From entire ecosystems to essential institutions, such as national defense or public health systems, many aspects of complex life depend on the willingness of self-interested individuals to contribute to a greater good, and in doing so, cooperate. Yet, cooperation often entails a cost to provide benefit to others, and we are still searching for the "enlightened self-interest" which would allow us to explain how selfish regard constantly prompts us to assist each other [de Tocqueville, 1835]. By the year 2005, the 125th anniversary of Science proposed a list of the top 25 questions faced by scientists over the following 25 years, and "How did cooperative behavior evolve?" once again made an appearance [Pennisi, 2005]. More than a decade past, the question still remains largely unresolved.

Understanding the origins of cooperation can help us build better models of the world that we live in. Models that not only serve the fundamental purpose of allowing us to understand the world, but that show us the best ways to *change* it. The complex evolutionary dynamics behind cooperation can provide key insights towards solving challenges in fields such as climate change [Góis et al., 2019; Pacheco et al., 2014; Raihani and Aitken, 2011; Santos et al., 2012b; Tavoni and Levin, 2014], conservation [Frank and Sarkar, 2010; Sumaila, 1999; Xu et al., 2020], crime [D'Orsogna and Perc, 2015; Short et al., 2011], corruption [Lee et al., 2015], or, more recently, global pandemics [Bauch and Earn, 2004; Fu et al., 2011]. Such domains are persistent reminders of how cooperation, or more accurately its absence, needs to be fully understood, before we pass a point of no return [Ord, 2020; Russell, 2019]. More than simply studying the dynamics of cooperation, we must learn how to *engineer* it and other pro-social behaviours. This is a timely and significant endeavour given the advent of Artificial Intelligence (AI), and a potential future in which autonomous agents nurture or enforce cooperation in a hybrid society of humans and machines [Akata et al., 2020; Paiva et al., 2018]. We may then ask the question – *How can we promote cooperation?*

In the quest towards understanding and engineering cooperation, we can look towards the human capacity to establish and contribute to institutions. From welfare states, courts, religion, whether at a local or international level, such as the United Nations or the European Union, humans have developed considerable machinery used at scale to create policy, and to distribute incentives [Knight, 1992; North, 1991; Ostrom, 1990]. Institutions have evolved to work as external mechanisms of control and levers for cooperation, common resource allocation, punishment and conflict resolution. On an individual level, an institution might distribute endowments, in the form of social support or investments but the scope of some institutions means they play a key role in resolving international political conflicts [Marton-Lafevre and Others, 2007; Smidt, 2020] or enforcing global climate change action [Góis et al., 2019; Ostrom, 2010; Pacheco et al., 2014]. In this thesis, we focus our efforts on the role of institutions and the problem of how to distribute incentives. Institutions are often faced with limited budgets and incomplete information, factors which further restrict their role in the emergence of cooperation. *Who should institutions target in order to optimise their spending?* As we will see shortly, the answer to this question is not trivial.

Among the many explanations that allow for cooperation, instead of defection, to emerge through natural selection, we seldom see accounts of mechanisms which are specifically designed to govern living behaviour. Douglass North, who spent most of his life studying institutional history, defines institutions as "rules of the game in a society" [North, 1990], and "humanly devised constraints that structure political, economic and social interactions" [North, 1991]. Laws, rules, social conventions and norms are all examples of institutions [Knight, 1992], and we consider them to be "exogenous to each individuals whose behavior it influences that generates behavioral regularities."[Greif and Laitin, 2004] Institutions have a unique relationship with cooperation, as they have themselves evolved to govern cooperation at a societal level. They are the principal object of study in many social sciences, such as economics, anthropology, political science, and sociology. The latter has been described as the "science of institutions, their genesis and their functioning,"[Durkheim, 1895] and it should come as no surprise that they have been approached as a possible avenue to explain the evolution of cooperation [Ostrom, 1990; Schoenmakers et al., 2014; Sigmund et al., 2010].

Despite the importance of institutions, and the attention they have been given in the literature, many questions still remain. Beyond simple reward and punishment, which have been resolved using several heuristics [Chen et al., 2015; Wang et al., 2019], these models commonly disregard cost. When cost is considered, it is usually assumed that the institution has complete control over the agents within the systems[Endriss et al., 2011; Wooldridge, 2012], which is technically easier to address. Furthermore, complex incentive mechanisms, underlying social network structures and incomplete information create challenges in the study of cooperation dynamics from an institutional perspective. To overcome some of these technical difficulties, models often assume infinitely large populations and simple network structures [Han and Tran-Thanh, 2018; Han et al., 2018]. Regularly, this modelling work also avoids introducing realistic characteristics which govern human social interactions, such as irrationality in decisions or the spontaneous adoption of new behaviours.

The technical implications of studying these characteristics are elevated when we consider complex networks of interaction between individuals.

Indeed, social diversity allows for the design of novel mechanisms that can be mathematically formulated and analysed systematically to quantify their role in the emergence of cooperation. Building upon the existing literature and attempting to fill in the gaps identified in Chapter 2 (section 2.5), the novelty of our methodology lies in the study of the effects of complex networks underlying the interactions between individuals, as well as the design of novel interference mechanisms to promote pro-social behaviour. One example would be to distinguish between different roles in asymmetric interaction settings (in Chapter 4), which can be considered a novel methodological contribution. This is achieved through evolutionary game theoretic models and the systematic analysis of the mechanisms we propose to promote pro-sociality. Through large-scale simulations and, whenever possible, analytical tools, we show that investing in socially diverse settings is non-trivial (in Chapter 3), strictly targeting specific roles is important in asymmetric interactions (in Chapter 4), the threat of punishment can serve as a deterrent to defection (in Chapter 5), and finally explore how these insights could be used to govern the development of AI (in Chapter 6). Before that, let us begin by asking – how can we define cooperation?

1.2 The Prisoner's Dilemma (and Game Theory)

Game Theory (GT) refers to the study of mathematical models of strategic interactions among rational agents [Myerson, 1997]. It can be applied to a broad set of problems in the fields of economics, social science, logic, computer science, economics, sociology, biology and more. Traditionally, it addressed two-person zero-sum games, but it has expanded in scope to cover a wide range of behavioural relations; it is now commonly used as a hypernym for the science of logical decision making in animals, humans and computers. Importantly, GT assumes that individuals are entirely rational. Their motives, beliefs, whims and errors are conveniently reduced – utility being considered above all else [Von Neumann and Morgenstern, 1944]. In many cases, such simplification proves to be useful, and facilitates the mathematical analyses of concepts such as Nash equilibria [Fudenberg and Levine, 1998; Nash, 1950], which will be detailed later. However, this reduction to rationality fails to address several realistic factors which have non-negligible effects, as we will see shortly. Through social change, irrationality and spontaneous behaviour change, evolutionary adaption can be applied to the concept of GT to give birth to a new approach [Sigmund, 2010].

Before we unearth concepts of equilibria and other mathematical complexities, let us return to traditional game theory and try to be more specific about the meaning of cooperation. To illustrate this point, we can use the well-known example of the prisoner's dilemma, one of the classic examples of cooperation in game theory. Originally, it was framed by Flood and Dresher while working at RAND, the think tank charged with formulating military strategy for the atomic age, in 1950. Only later was it formalised with prison sentence rewards and baptised by Tucker [Poundstone, 1992], who presented it as follows:

> Two members of a criminal organization are arrested and imprisoned. Each prisoner is in solitary confinement with no means of communicating with the other. The prosecutors lack sufficient evidence to convict the pair on the principal charge, but they have enough to convict both on a lesser charge. Simultaneously, the prosecutors offer each prisoner a bargain. Each prisoner is given the opportunity either to betray the other by testifying that the other committed the crime, or to cooperate with the other by remaining silent.

	Deny (C)	Betray (D)
Deny (C)	-1, -1	-3, 0
Betray (D)	0, -3	-2, -2

Table 1.1 Standard Prisoner's Dilemma

As stated before, it is assumed that both players are rational and ignore factors such as loyalty, reward, retribution and reputation outside of the dilemma itself. An example of possible outcomes is described in Table 1.1, and regardless of what the other prisoner decides, the reward is higher if they betray the other. The PD is aptly named so because mutual cooperation yields a better collective outcome. It is not the rational outcome, as denying the crime is irrational if the prisoners are self-interested. Herein lies the dilemma. One of the key takeaways from the PD is that decisions made under individual rationality may not necessarily align with decisions made under collective rationality. This situation is often called the *Tragedy of the Commons*, and is encountered whenever individuals have open access to a public resource, causing the depletion of said resource through selfish, uncoordinated action [Hardin, 1968; Ostrom, 1990].

While the eponymous setting initially seems contrived, there are in fact several examples of human and natural interactions which can be formulated in a similar fashion. In environmental studies, the PD is used to model crises such as global climate change [Barrett and Dannenberg, 2012], in which uncertainty suggests that individual states would cooperate even less than what is observed in the PD. In evolutionary biology, Dawkins agrees with Robert Axelrod and Hamilton that many animals are "engaged in ceaseless games of the Prisoner's Dilemma, played out in evolutionary time" [Dawkins, 1976], and he gives several real-world examples of the PD seen in animals. One such example relates to Wilkinson's research on vampire bats [Wilkinson, 1984] which engage in reciprocal food exchange. Applying the payoffs discussed earlier, this reciprocal behaviour can more easily be understood – where giving blood on a good night can in turn save individuals from starving on bad nights.

George Ainslie, writing from the perspective of addiction and behavioural economics has pointed out that addiction can be cast as a PD problem [Ainslie, 2001]. Defecting is akin to *relapsing*, and not defecting today or in the future is the best outcome. Doping in sport [Schneier, 2012], international politics [Majeski, 1984], exploitation of the commons [Hardin, 1968], cartel agreements [Nicholson, 2000], etc., are all precedents for the wide applicability of the PD and the substantial importance it has in understanding social interactions. It has been called the *E. coli* of social psychology [Axelrod, 1980], and it serves as one of the simplest, yet profoundly compelling tools for modelling behaviour.

1.3 Collective Dynamics

Many disciplines engage in the study of cooperation and institutions, and the myriad of methods which are employed in the literature cover a vast expanse of techniques found in economics [Boyd et al., 2003; Fehr and Gachter, 2000; Fehr and Schmidt, 1999], evolutionary biology [Nowak, 2006b, 2012; Nowak et al., 1994], developmental psychology [Kiley Hamlin et al., 2011; Warneken and Tomasello, 2007], neuroscience [Sanfey et al., 2003; Watanabe et al., 2014], and many more. Here, we decide to focus on **mathematical and computational tools** to study **cooperation, fairness, institutions and collective dynamics** which emerge as a result of interactions within large populations of agents. Below we outline some of the main reasons for us to conduct research using this approach:

1. *Populations are the building blocks of evolution* [Nowak, 2006a]. Therefore, we conduct theoretical and simulation-based studies of population dynamics. Through Evolutionary Game Theory (EGT), we can quantify the complex interplay between selection, replication and mutation in the evolution of cooperation and fairness.

2. *Explaining cooperative behaviour often requires reaching beyond genetics and biology* [Boyd and Richerson, 1985]. Using EGT as the framework for our models is a convenient way to analyze the invasion and fixation of behaviours through evolution, through reproduction and inheritance [Nowak, 2006a]. Moreover, this approach allows us to address collective dynamics in social systems, producing a frequency-dependent spread of behaviours in complex systems, by using mechanisms such as peer imitation and social learning [Sigmund, 2010]. Understanding the collective dynamics of cooperation can unveil the nature of genetic and cultural dynamics alike.

3. *Population dynamics can account for individual learning, not only genetic inheritance or social learning*. Human behaviours can emerge after a process of adaption, in which individuals use their own experience to shape their future behaviour [Fudenberg and Levine, 1998; Han, 2013; Macy and Flache, 2002; Skyrms, 2010]. Individual learning within the

population follows the same frequency-dependent dynamics which are characteristic of EGT. The fitness, or success, associated with each behaviour is used to adapt an individual's future behaviour, and this is dependent on the ecology of behaviours which are observed in the population at every step.

4. *Complex wholes often emerge from simple parts*, and some phenomena can only be understood as an emergent property of interacting individuals. Emergence plays a central role in complex adaptive systems [Miller and Page, 2009; Mitchell, 2009], and it one of the main motivators for modelling large populations in order to study collective dynamics. It is likely that cooperation and fairness are emergent properties, in a way which is not trivial to infer from interactions at a small scale [Pinheiro et al., 2012a, 2016; Schelling, 1978].

5. *Real-life societies are often unequal, and social diversity is ubiquitous.* Classical game theory usually considers individuals to be equivalent in all respects, but EGT and computational tools allow us to model complex networks of individuals. Moving away from well-mixed populations, we can study a range of network topologies, in some of which individuals have a disproportionate amount of social ties. Heterogeneity has been shown to play a key role in the emergence of cooperation [Barabasi, 2014; Santos and Pacheco, 2005; Santos et al., 2006a, 2008, 2006b], and we might ask whether social diversity can be exploited by an institution to enforce cooperation and reduce spending.

Mathematical methods and computational tools can assist in the grand challenge of understanding the collective dynamics of cooperation and institutions. Large-scale simulations have already been employed to unveil the roles of interacting networks [Santos et al., 2008], the emergence of social norms [Nowak and Sigmund, 1998; Pacheco et al., 2006a] and reputation [Santos et al., 2018a,b], and collective risk dynamics [Domingos et al., 2020; Góis et al., 2019]. Computer science and engineering play a promising role in this field, but not only as a source for tools. Above all, they serve as a conducive domain of application, in managing distributed systems (e.g. e-commerce platforms [Resnick et al., 2006]), peer-to-peer networks [Feldman and Chuang, 2005], mobile crowdsensing systems [Di Stefano et al., 2020], crowdsourcing markets [Ho et al., 2012] or swarm robotics [Floreano et al., 2008; Waibel et al., 2011]. Significant to this thesis are the sub-fields of AI and multi-agent systems [Armstrong et al., 2016; Paiva et al., 2018; Wooldridge, 2009], where further understanding is increasingly needed. For instance, human-agent cooperation in hybrid societies, or unveiling the dynamics associated with the development of novel AI technology. The latter in particular is of momentous interest, having been identified as one of the key existential risks which threaten the long-term future of humanity [Ord, 2020; Russell, 2019].

1.4 AI Safety

Researchers and stakeholders alike have urged for due diligence in regard to AI development on the basis of safety concerns. Not least among them is that AI systems could easily be applied to nefarious purposes, such as espionage or cyberterrorism [Taddeo and Floridi, 2018]. Moreover, the desire to be at the foreground of the state-of-the-art or the pressure imposed by upper management might tempt developers to ignore safety procedures or ethical consequences [Armstrong et al., 2016; Cave and ÓhÉigeartaigh, 2018]. Indeed, such concerns have been expressed in many forms, from letters of scientists against the use of AI in military applications [Future of Life Institute, 2015, 2019], to blogs of AI experts requesting careful communications [Brooks, 2017], and proclamations on the ethical use of AI [Declaration, 2018; Jobin et al., 2019; Russell et al., 2015; Steels and de Mantaras, 2018].

Regulating and governing advanced technologies such as Artificial Intelligence (AI) has become increasingly more important given their potential implications, risks and ethical concerns [Declaration, 2018; European Commission, 2020; Future of Life Institute, 2015, 2019; Jobin et al., 2019; Perc et al., 2019; Russell et al., 2015; Steels and de Mantaras, 2018]. With the great benefits promised from being first able to supply such technologies, stakeholders might cut corners on safety precautions in order to ensure rapid deployment, in a race towards AI market supremacy (AIS) [Armstrong et al., 2016; Cave and ÓhÉigeartaigh, 2018]. One does not need to look very far to find potentially disastrous scenarios associated with AI [Armstrong et al., 2016; O'neil, 2016; Pamlin and Armstrong, 2015; Sotala and Yampolskiy, 2014], but accurately predicting outcomes and accounting for these risks is exceedingly difficult in the face of uncertainty [Armstrong et al., 2014]. The impact of a new technology is difficult to predict before it has been already extensively developed and widely adopted, and also difficult to control or change after it has become entrenched (so called the Collingridge Dilemma), [Collingridge, 1980]. Given the lack of available data and the inherent unpredictability involved in this new field of technology, a modelling approach is therefore desirable to provide a better grasp of any expectations with regard to a race for AIS. Such modelling allows for dynamic descriptions of several key features of the AI race (or its parts), providing an understanding of possible outcomes, considering external factors and conditions, and the ramifications of any policies that aim to regulate such a race.

With this aim in mind, a baseline model of an innovation race has been recently proposed [Han et al., 2019], in which innovation dynamics are pictured through the lens of EGT, as described above, and all race participants are equally well-connected in the system (well-mixed populations). These baseline results showed the importance of accounting for different time-scales of development, and also exposed the dilemmas that arise when what is individually preferred by developers differs from what is globally beneficial. When domain supremacy could be achieved in the short-term, unsafe development required culling for to promote the welfare of society, and the opposite was true for the very long term, to prevent excessive regulation at the start of exploration. However, real-world stakeholders and their interactions are far from homogeneous. Some individuals are more influential than others, or play different roles in the unfolding of new technologies. Technology races are shaped by complex networks of exchange, influence, and competition where diversity abounds. It has been shown that particular networks of contacts can promote the evolution of positive behaviours in various settings, including cooperation [Chen et al., 2015; Ohtsuki et al., 2006; Perc and Szolnoki, 2010; Perc et al., 2017; Santos et al., 2006a, 2008], fairness [Cimpeanu et al., 2021a; Page et al., 2000; Santos et al., 2017; Szolnoki et al., 2012; Wu et al., 2013] and trust [Kumar et al., 2020].

In the context of technology regulation and governance, the impact of network topology is particularly important. Technology innovation and collaboration networks (e.g. among firms, stakeholders and AI researchers) are highly heterogeneous [Newman, 2004; Schilling and Phelps, 2007]. Developers or development teams interact more frequently within their groups than without, forming alliances and networks of followers and collaborators [Ahuja, 2000; Barabasi, 2014]. Many companies compete in several markets while others compete in only a few, and their positions in inter-organisational networks strongly influence their behaviour (such as resource sharing) and innovation outcome [Ahuja, 2000; Shipilov and Gawer, 2020]. It is therefore paramount to understand how diversity in the network of contacts influences race dynamics and the conditions under which regulatory actions are needed. We study how network structures influence safety decision making within an AI development race.

1.5 How to Read This Thesis

This thesis is divided into seven chapters, each outlined below:

1. Introduction — intended to provide the general motivation behind this work, and providing a brief overview of cooperation, institutions, the choice of methodology and a relevant application domain. Purposefully, this section is shallow in technical descriptions and wide in scope. This chapter includes an overview of the peer-reviewed publications which resulted from this thesis.

2. Research Context and Background — presents useful technical detail which we believe to be useful in understanding the contents of this thesis. We present an overview of relevant definitions in Game Theory and Evolutionary Game Theory, network topologies and centrality measures, offering a theoretical basis on which we construct the following chapters. We also provide a literature review on external interference, mechanism design and other relevant contributions which allow for a better understanding of the scope of this thesis.

The following four chapters contain the corpus of this thesis. Each chapter corresponds to an already peer-reviewed, published and presented work.

Each of these chapters was organised in a mostly self-contained format, and can be read independently from the others, with relevant pointers to the technical background when required.

3. Promoting Cooperation in Scale-Free Networks — in which we address external interference in symmetric interactions, a relatively simple setting to study ways in which institutions can promote cooperation in heterogeneous graphs. This chapter shows the effects of diversity on the potential choice of paradigms available to investors, and outlines several pitfalls that can lead to the attrition of cooperation.

4. Promoting Fair Proposers, Responders or Both? — in which we address fairness in asymmetric interactions, whereby individuals can act in multiple roles. We show that fairness requires a strict approach to investment, and that diversity can relax these requirements, reducing the complexity of distributing endowments.

5. Making an Example: Signalling Threat in the Evolution of Cooperation — in which we introduce the concept of signalling punishment, leading to the evolution of a new type of defector, one who renounces temptation due to fear. We show that the existence of fearful defectors leads to an increase in social welfare and cooperation, providing a compelling argument to suggest that the threat of punishment is an effective deterrent in social and institutional settings.

6. AI Safety in Heterogeneous Settings — starting from a model describing an idealised technology race, we investigate how different interaction structures among race participants can alter collective choices and requirements for regulatory actions. Our results suggest that AI regulation may profit from the world's patent diversity and show that a minority of participants are capable of influencing an entire population towards an ethical and sustainable use of advanced technology.

7. Conclusions and General Discussion — summarizes the conclusions of this thesis as reported by the core substance chapters. We also include a general description of potential application domains, the scope of these results, and point to avenues for future work.

1.6 List of Publications

Here we provide a list of the publications directly related to this thesis. In all cases, I was the main author, providing the core contributions for each publication. This includes proposing the research hypotheses, implementing and running the simulations, processing the data, generating the plots and writing the bulk of the finished manuscripts. The co-authors contributed to the design of the research hypotheses and in writing the manuscripts.

Journal Articles

- Theodor Cimpeanu, Cedric Perret and The Anh Han. **Cost-efficient interventions for promoting fairness in the Ultimatum game**. In *Knowledge-Based Systems* 233, p. 107545, 2021. ISSN: 0950-7051. doi: 10.1016/j.knosys.2021.107545.
- Theodor Cimpeanu, Francisco C Santos, Luis Moniz Pereira, Tom Lenaerts and The Anh Han. **Artificial intelligence development races in heterogeneous settings**. In *Scientific Reports* 12(1):1723, 2022. ISSN 2045-232. doi: 10.1038/s41598-022-05729-3.

Conference Papers

- Theodor Cimpeanu, The Anh Han and Francisco C Santos. **Exogenous Rewards for Promoting Cooperation in Scale-Free Networks**. In *ALIFE 2019*, pages 316-323. MIT Press, 2019.
- Theodor Cimpeanu. **Cost Effective Interventions in Complex Networks Using Agent-Based Modelling and Simulations: Doctoral Consortium**. In *Proceedings of the 19th International Conference on Autonomous Agents and Multi-Agent Systems*, AAMAS ’20, pages 2188-2190. United States: Association for Computing Machinery (ACM), 2020.
- Theodor Cimpeanu and The Anh Han. **Fear of Punishment Promotes the Emergence of Cooperation and Enhanced Social Welfare in So-**

cial Dilemmas. In *Proceedings of the 19th International Conference on Autonomous Agents and Multi-Agent Systems*, AAMAS ’20, pages 1819-1821. United States: Association for Computing Machinery (ACM), 2020.

- Theodor Cimpeanu and The Anh Han. **Making an Example: Signalling Threat in the Evolution of Cooperation**. In *2020 IEEE Congress on Evolutionary Computation (CEC)*, pp. 1-8, 2020, doi: 10.1109/CEC48606.2020.9185749.

- Theodor Cimpeanu, Cedric Perret and The Anh Han. **Promoting Fair Proposers, Fair Responders or Both? Cost-Efficient Interference in the Spatial Ultimatum Game**. In *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021)*, pages 1480-1482. United States: Association for Computing Machinery (ACM), 2021.

Preprints

- Theodor Cimpeanu, Alessandro Di Stefano, Cedric Perret and The Anh Han. **Social Diversity Reduces the Complexity and Cost of Fostering Fairness**. – *Preprint*, 2022. Submitted, under review.

Other publications, abstracts and communications

- Theodor Cimpeanu. **Exogenous Rewards for Promoting Cooperation in Scale-Free Networks**. In *Social Learning and Cultural Evolution, Workshop part of ALIFE2019*, 2019. Poster presentation.

- The Anh Han, Long Tran-Thanh, Simon Lynch, Theodor Cimpeanu and Francisco C Santos. **Promoting Cooperation through External Interference**. In *ALIFE 2019*, pages 331-332. MIT Press, 2019.

- Theodor Cimpeanu. **Exogenous Rewards for Promoting Cooperation in Scale-Free Networks**. In *Arizona State University Interdisciplinary Study of Cooperation Winter School*, 2020. Poster and invited student talk.

- Theodor Cimpeanu, Cedric Perret and The Anh Han. **Promoting Fairness in the Spatial Ultimatum Game**. In *Conference on Complex Systems 2020*, CCS 2020. Presentation.

- Theodor Cimpeanu, Francisco C Santos, Luis Moniz Pereira, Tom Lenaerts and The Anh Han. **Heterogeneous Interactions in Artificial Intelligence Development Races**. In *Conference on Complex Systems 2021*, CCS 2021. Poster and presentation.

- Bianca Ndidi Ogbo, Theodor Cimpeanu, Alessandro Di Stefano and The Anh Han. ***Shake on It*****: The Role of Commitments and the Evolution of Coordination in Networks of Technology Firms**. – *Preprint*, 2022. Submitted, under review.

- Theodor Cimpeanu, Francisco C Santos, Luis Moniz Pereira, Tom Lenaerts and The Anh Han. **Network Diversity Promotes Safety Adoption in Swift Artificial Intelligence Development**. – *Preprint, abstract*, 2022. Submitted, under review.

2 | Research Context and Background

If I have seen further it is by standing on the shoulders of Giants.

—Isaac Newton

In this chapter, we review relevant theoretical concepts and the literature in the fields of cost efficient interference and complex networks, to frame the contributions of this thesis. Firstly, we present the social dilemmas of cooperation and fairness in a game theoretical formalism, also introducing key theoretical concepts within game theory. Secondly, we review the computational and mathematical methods that allow us to capture the collective dynamics of cooperation and fairness. We focus in particular on Evolutionary Game Theory, namely with replicator dynamics and computational methods that allow us to model finite populations and complex networks, paying regard to the intrinsic stochasticity of the associated evolutionary dynamics. Moreover, we introduce the key graphs employed as networks of interaction for the models introduced in subsequent chapters. Finally, we review the existing models which have been proposed to study cost-effective interference in the past, allowing us to formally identify gaps in the literature and pose the open questions that this thesis will investigate.

2.1 Game Theory, Cooperation and Fairness

As introduced briefly in Section 1.2, Game Theory (GT) is the field of mathematics which aims to study strategic interactions among rational agents. We use the term *strategic interactions*, as **players** engage in interactions whereby deciding upon a **strategy** (or **action**) results in an outcome which is dependent on the actions employed by other players. The resulting outcome is typically quantified in a **payoff** value, which influences a player's success. Later, we will equate this concept to *fitness* (in the evolutionary sense), thus players with higher fitness will have a greater chance that their strategies spread in the population. In the classical sense, *rational individuals* are coherent in their preferences, and assumed to always prefer the outcomes which lead to maximal payoffs. With this very general formal model, GT is commonly employed to study economic, political and biological phenomena [Osborne, 2004].

Let us expand on the aforementioned classical problem of the Prisoner's Dilemma, described in Section 1.2. We have already discussed some of the implications arising from a seemingly specific example of suspects wishing to cut a deal with the prosecution. Following that same interaction pattern, we can generalise the Prisoner's Dilemma using a canonical matrix (see Table 2.1). As the interaction is symmetric, we can conveniently describe it using row player notation, as follows: temptation (**T**) is the potential benefit of betraying the other person if they keep silent; reward (**R**) results as an outcome of mutual cooperation; punishment (**P**) if both players decide to betray the other; and the sucker's payoff (**S**), for the player who gets betrayed while keeping silent. In order for this to be considered a PD in the strong sense, then the following condition must hold: $T > R > P > S$. Among the many interesting considerations following this description, we may pose two initial questions: 1) Which outcome is socially desirable (i.e. the best)? and 2) Which outcome is likely to happen?

	Cooperate (C)	Defect (D)
Cooperate (C)	R	S
Defect (D)	T	P

Table 2.1 **Generalised social dilemma of cooperation,** following the same interaction pattern described in Table 1.1. For this to be considered a Prisoner's Dilemma, the following conditions must hold: $T > R > P > S$.

We can answer question 1) by means of finding the socially desirable condition. This can be accomplished by resorting to the utilitarian idea of maximising the number of players with maximal payoffs. We will have reached the best outcome if we cannot improve the payoff of any one individual without damaging the payoffs of others. This is usually referred to as a Pareto optimal (or Pareto efficient) outcome.

Definition 2.1.1 (Pareto optimality)
Denoting $\pi_i(X)$ as the payoff for player i, given a set of strategies X; a strategy profile $s' = \{s'_1, ..., s'_i, ..., s'_N\}$ is Pareto optimal if there exists no other strategy profile $s = \{s_1, ..., s_i, ..., s_N\}$ such that $\pi_i(s) \geq \pi(s'), \forall i$.

Before attempting to answer question 2) we should first consider that players do not have information about what others will play, and this observation may have implications when seeking the optimal strategy. It may be the case that player A may use strategy a after deducing that B will use strategy b. As a is the best option for A, B may now decide to counteract by selecting action c, which in turn causes A to resort to strategy d, and so on, leading to a potentially infinite loop of reasoning(s). Addressing and overcoming this recursive trap, the definition of the Nash equilibrium [Nash, 1950] by only extracting the strategy profiles which are stable. In other words, no player would unilaterally have any incentive to deviate from the stable profile:

Definition 2.1.2 (Nash equilibrium)
A Nash equilibrium in a group of N players is a strategy profile $s = \{s_1, ..., s_i, ..., s_N\}$ if there exists no other s for which player i would do better by choosing the action s_i, given that every other player j would keep selecting the strategy s_j.

If we take the example of the Prisoner's Dilemma (Table 2.1), we conclude that (C, C) and (C, D) are both Pareto efficient strategy profiles. While both strategy profiles are optimal, we can consider the first to be *egalitarian*, and the latter to be *unfair*. Moreover, the only Nash equilibrium of the game is (D, D), as either player would earn less than what they would have if they had decided to cooperate, instead. It is this mismatch between the Pareto efficient outcomes, the social optimum (C,C), and the most likely profile (D,D) which creates the social dilemma. As expressed by Dawes [1980], *social dilemmas are characterized by two properties: (a) the social payoff to each individual for defecting behavior is higher than the payoff for cooperative behavior, regardless of what the other society members do; yet, (b) all individuals in the society receive a lower payoff if all defect than if all cooperate.*

2.1.1 Fairness

We have seen that the Prisoner's Dilemma serves as a useful tool for modelling cooperation, but it is important to note that many real-world and MAS interactions are asymmetric, meaning that players can play multiple roles in the interactions [McAvoy and Hauert, 2015]. One particularly revealing light of the influence of role asymmetry, which has become a popular instrument of economic experiments, is the Ultimatum Game, first described by Nobel laureate John Harsanyi [1961]. It was later popularised by Werner Güth et al. [1982] in their famous series of experiments, as a simple bargaining

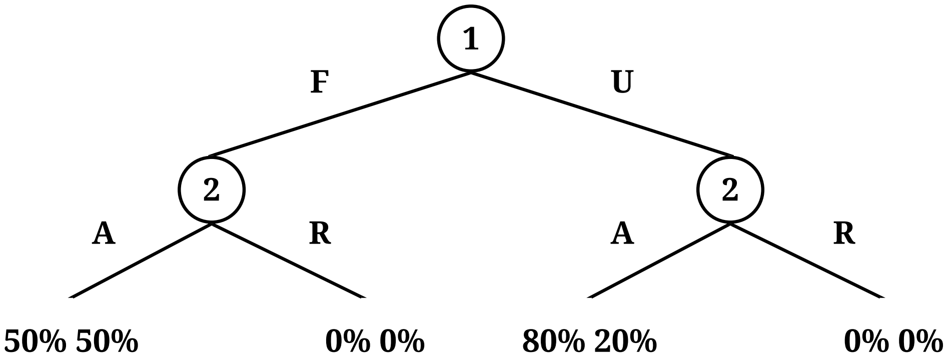

Fig. 2.1 **Extensive form representation of a two proposal Ultimatum Game.** Player 1 can offer a fair (F) or unfair (U) proposal; player 2 can either accept (A) or reject (R). The outcome is split as follows, with player 1 (the proposer) receiving the first percentage of the initial endowment for each outcome.

environment. In these games, two people were randomly and anonymously matched, and assigned one of two roles, one as a proposer and the other as a responder, and told they will play this game exactly one time. The proposer is given a certain sum of money (the endowment), and is asked to suggest a division of that amount between themselves and the responder. In turn, the responder can observe the proposition and decide whether to accept or reject the offer. If the responder accepts the division, then the sum is split as implied by the suggestion, otherwise neither receives any amount of the initial endowment. One example of the UG is represented in extensive form in Figure 2.1.

For ease of exposition, we have only illustrated the simple example in which the proposer has a binary choice: either propose a fair split (high) or unfair one (low). In practice, this can be extended to a more general case, in which the proposer can choose from a theoretically infinite number of possible splits, limited in practice only by the lowest denomination of currency in the experiment. In this setting, it always benefits the responder to accept any offer, as receiving any amount is better than receiving nothing. Meanwhile, the proposer aims to suggest an amount that the responder will accept. Reasoning that the responder would accept any offer, then the only rational choice for the proposer would be to offer an unfair split. Thus, we can extract three Nash equilibria for this game:

- The proposer offers a fair split; the responder would only accept fair offers.
- The proposer offers an unfair split; the responder would only accept unfair offers.
- The proposer offers an unfair split; the responder would accept any offers.

However, there exists a refinement of the Nash equilibrium for studying dynamic games, such as the UG, and that is subgame perfection. A subgame is any part, or subset, of a game which, seen in isolation, constitutes another game in its own right. For instance, in the UG described above, choosing whether to accept or reject an offer is a different subgame depending on what the initial proposed split is. If the division is a binary choice (fair or unfair split), then there exist two subgames in this example. A subgame perfect equilibrium (or subgame perfect Nash equilibrium) occurs if there are Nash equilibria in every subgame, which players have no incentive to deviate from. This theory relies on the assumption that players are rational and wish to maximise their utility [Osborne, 2004], and can be formulated as follows:

Definition 2.1.3 (Subgame perfect Nash equilibrium)
A strategy profile $s = \{s_1, ..., s_i, ..., s_N\}$ in a group of N players is a subgame perfect Nash equilibrium if a Nash equilibrium is played in every subgame.

Every finite extensive game with perfect recall has a subgame perfect equilibrium [Osborne, 2004]. Perfect recall is a concept first coined by Kuhn and Tucker [1953] which is "equivalent to the assertion that each player is allowed by the rules of the game to remember everything he knew at previous moves and all of his choices at those moves". One common method for determining subgame perfect equilibria for finite games is backward induction. The above game can be perceived as two separate subgames: the one in which the proposer makes a fair offer, and the one in which they make an unfair offer. In both of these subgames, it benefits the responder to accept the offer, so by backward induction we find that the only subgame perfect equilibrium is the final one, in which a responder accepts any offer.

Assuming that players are rational, and that Nash equilibria constitute an accurate prediction of what rational decision makers might do, we should expect proposers to offer the lowest possible amount, and responders to willingly accept being short-changed as opposed to receiving nothing. Similarly, we should assume defection to be the overwhelming paradigm chosen by individuals participating in a Prisoner's Dilemma. This is a common assumption made in Game Theory, and stands as the foundation for the *homo economicus* (Economic Man) portrayal of humans as perfectly rational and self-interested agents, maximising utility above any other considerations, which often features in economic theory and pedagogy [Zak, 2008].

Results from experiments with the UG challenge the traditional economic principles of *homo economicus* [Güth et al., 1982; van Damme et al., 2014], as humans not only reject unfair offers, but overwhelmingly tend to propose fair and even splits. An ample research program has used this exact setting to study fairness norms in small-scale societies, such as hunter-gatherers and nomads, and while differences within these societies and urban populations exist, the average offer is never close to the theoretical minimum [Henrich et al., 2005]. Ever since its inception, the UG has become one of the most popular experimental frameworks for studying fairness in a wide range of fields, and it was said to be "quickly catching up with the Prisoner's Dilemma as a prime showpiece of apparently irrational behavior" in a seminal paper by Nowak, Page, and Sigmund [2000].

While many mechanisms have been proposed to explain this discrepancy between the view of the Economic Man and real humans, the fact remains that norms of fairness are almost universal. Why, then, do we see such differences between theoretical predictions of fairness and cooperation, and real-world observations from experiments? One step towards answering that question would involve considering that players neither have full information about the interactions that they are participating in, nor do they compute all possibilities in order to maximise their utility. Within a population, individuals might resort to heuristics, such as social learning and imitation, adapting their behaviour to find a functional approach. Comparable to the principle of natural selection, the success of a strategy is therefore determined by how well it performs against competing strategies, and the frequency with which they appear in the population. Darwinian competition meets Game Theory [Maynard Smith, 1982; Maynard Smith and Price, 1973].

2.2 Evolutionary Game Theory

Let us consider a symmetric game with the payoff matrix Π and assume that in a large, completely connected (well-mixed) population, a fraction x_i uses strategy $\mathbf{s}_i$, for $i = 1,...,n$. The state of the population is thus given by the vector $\mathbf{x} = \{x_1,\dots,x_n\} \in S_n$. S_n denotes the state of a population consisting of different strategies. The expected payoff for a player following strategy $\mathbf{s}_i$ reads

$$(\Pi\mathbf{x})_i = \sum_{j=1}^{n} \pi_{ij} x_j. \tag{2.1}$$

As the player meets a co-player using strategy $\mathbf{s}_j$ with a probability x_j, we can write the average payoff in the population as

$$\mathbf{x} \cdot \Pi\mathbf{x} = \sum_{i=1}^{n} x_i (\Pi\mathbf{x})_i. \tag{2.2}$$

Now we have reached a crucial step. Let us assume that populations can evolve, and that the frequencies x_i can change over time. Then, we let the state $\mathbf{x}(t)$ depend on time, and denote by $\dot{x}_i(t)$ the velocity of change for x_i (i.e. $\dot{x}_i = dx_i/dt$). How does the population evolve? How do the frequencies of strategies grow and die out? While many possibilities exist for answering these questions, let us use the *replicator equation*, which appeared very early in the context of biological games [Sigmund, 2010]. This equation holds true if the growth rate of a strategy's frequency depends on the difference between its payoff $(\Pi\mathbf{x})_i$ and the average payoff $\mathbf{x} \cdot \Pi\mathbf{x}$. The replicator equation can be expressed as

$$\dot{x}_i = x_i[(\Pi\mathbf{x})_i - \mathbf{x} \cdot \Pi\mathbf{x}] \tag{2.3}$$

for $i = 1, \ldots, n$. In other words, a strategy $\mathbf{s}_i$ will grow or dwindle according to how well it performs compared to the average of the population. The sign of $\dot{x}_i$ indicates the direction of evolution. While the biological process of reproductive success leads to the replicator equation very quickly, a similar argument can be made for social learning [Sigmund, 2010], and we reach the same equation given a general *fitness* term instead of the payoff $(\Pi \mathbf{x})_i$.

It is important to note that imitation (or selection, in biological terms) never leads to novel behaviour. If a strategy is absent in the population, it will remain so, and this can be easily gleamed from Equation 2.3. We can conceive several game dynamics which are more innovative. For instance, we can consider a steady rate of changing behaviour randomly. This is usually referred to as an *exploration rate*, and corresponds to mutation in genetics.

Replicator dynamics are particularly useful when attempting to show the evolutionary relationship between different strategies, and one approach would be to compute rest points. Rest points are those for which all payoff values $(\Pi \mathbf{z})_i$ are equal, for all indices i for which $z_i > 0$ [Sigmund, 2010]. The replicator equation admits a rest point for the strategy profile S_n if there exists a solution of the linear equations

$$(\Pi \mathbf{x})_1 = \ldots = (\Pi \mathbf{x})_n. \tag{2.4}$$

Let us consider a simple setting in which there are only two strategies in the population, for ease of representation. Subtracting the diagonal term in each column does not affect the equation, so without loss of generality, we can assume that the 2×2 matrix Π is of the form

$$\begin{pmatrix} 0 & a \\ b & 0 \end{pmatrix}.$$

There are only two strategies to choose from, so $x_2 = 1 - x_1$. Thus, it is sufficient to observe only x_1, which we can denote by x. Therefore $x_2 = 1 - x$, and

$$\dot{x} = x\left[(\Pi \mathbf{x})_1 - \mathbf{x} \cdot \Pi \mathbf{x}\right] = x[(\Pi \mathbf{x})_1 - (x(\Pi \mathbf{x})_1 + (1 - x)(\Pi \mathbf{x})_2)], \tag{2.5}$$

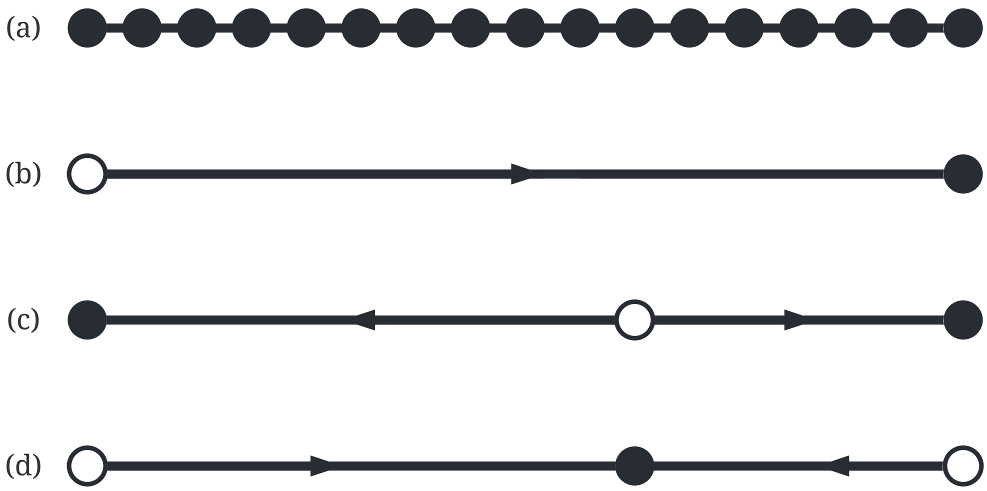

Fig. 2.2 **Classification of replicator dynamics for two strategies:** (a) inertia; (b) dominance; (c) bi-stability; (d) stable coexistence. Circles denote rest points, with filled circles representing the stable ones.

which we can simplify to

$$\dot{x} = x(1-x)[(\Pi \mathbf{x})_1 - (\Pi \mathbf{x})_2]. \tag{2.6}$$

By definition, we have $(\Pi \mathbf{x})_1 = a(1-x)$ and $(\Pi \mathbf{x})_2 = bx$, so we can simplify Equation 2.6 by substitution

$$\dot{x} = x(1-x)[a-(a+b)x]. \tag{2.7}$$

Of note is that

$$a = \lim_{x \to 0} \frac{\dot{x}}{x}. \tag{2.8}$$

Thus a corresponds to the limit of the individual growth rate of the missing strategy $\mathbf{s}_1$. Alternatively, this can be written as

$$a = \frac{d\dot{x}}{dx}, \tag{2.9}$$

where the derivative is evaluated at the point $x = 0$.

We obtain three possible cases for the right-hand side of the differential equation 2.7, each corresponding to one of the factors being equal to zero.

We intentionally omit the trivial case of $a = b = 0$ from this explanation, as all points of the state space (the interval $0 \leq x \leq 1$) are rest points. The first factor vanishes at 0, the second at 1, and the third factor has a zero $\hat{x} = \frac{a}{a+b}$ in the open interval $(0,1)$ if $ab > 0$. Figure 2.2 illustrates these three cases:

1. If $ab \leq 0$, there is no fixed point within the state space. If this occurs, $\dot{x}$ always has the same sign in the interval $(0,1)$. Given a positive sign (i.e. if $a \geq 0$ and $b \leq 0$, at least one strict inequality), then $x(t) \rightarrow 1$ for $t \rightarrow \infty$, for every initial value $x(0)$ $(0 < x(0) < 1)$. In this case, the strategy $\mathbf{s}_1$ is said to *dominate* strategy $\mathbf{s}_2$. In other words, the former is always the best reply, for any value of $x \in (0,1)$. Conversely, the opposite is true if the sign of $\dot{x}$ is negative. Regardless of the sign, the dominating strategy converges towards fixation. One example of this occurence is the Prisoner's Dilemma game mentioned previously (see Table 1.1); defection dominates.

2. If there is a rest point $\hat{x} \in (0,1)$ (i.e. $ab > 0$), and both a and b are negative. Then, $\dot{x} < 0$ for $x \in (0,\hat{x})$ and $\dot{x} > 0$ for $x \in (\hat{x},1)$. This rest point is unstable. As in case 1), one strategy will survive, but the outcome in this *bistable* case varies depending on the initial conditions. If $x > \hat{x}$, then it will keep on growing; else, it will disappear.

3. Similarly to 2), there exists a rest point $\hat{x} \in (0,1)$, and both a and b are positive. Differently, $\dot{x} > 0$ before the threshold $\hat{x}$, and $\dot{x} < 0$ otherwise. This is referred to as negative feedback, and means that $x(t)$ will converge towards $\hat{x}$, for $t \rightarrow \infty$, the rest point is a stable attractor. These strategies will both survive, as their frequencies converge towards *stable coexistence*.

Replicator dynamics and Nash equilibria are closely related concepts; both serve as convenient, yet ultimately limited tools to study collective dynamics. Both of these methods point towards the extinction of cooperators, regardless of social learning. We can make a similar argument that implies the death of fairness in the Ultimatum Game. It would seem that social learning is not enough to explain cooperation or fairness. Mathematical tools can capture simple models and dynamics, but the addition of complexities, such as structured populations, high exploration rates and incentive mechanisms, often make such formal descriptions infeasible. Thus, we turn to computer simulations to find our answers.

2.3 Computer Simulations

In this thesis, we make use of and extend the mathematical methods employed to study collective dynamics and the emergence of cooperative or fair behaviours. Naturally, the methods we focus on reflect the characteristics and context of the interactions, with increasing levels of complexity. For instance, the previously mentioned replicator equation has, since its adoption in 1978, been considered one of the central methods in EGT and collective dynamics analysis. Notwithstanding, various improvements and modifications have been proposed since its integration in the field. For instance, we might look towards birth-death processes combined with the pairwise comparison rule [Traulsen et al., 2006] as a way to study stochasticity, moving away from the deterministic setting presented earlier in this chapter. Likewise, we could study stationary distributions and invasion probabilities if we restrict ourselves to very low mutation probabilities, using the so-called small mutation approximation [Fudenberg and Imhof, 2005].

Earlier in this chapter, we specified that replicator dynamics are useful in the study of very large populations. Indeed, there exists a wide class of models where fluctuations in agent preferences, characteristics, payoffs or stochastic elements average out and produce smooth macroscopic behaviour from the aggregate of these fluctuations [Szabó and Fáth, 2007]. This is usually referred to as mean-field analysis, and assume an infinite, homogeneous population, and in these situations it provides an adequate qualitative description. However, we lose the analytical naivete of aggregation as we introduce elements which influence agent behaviour at the microscopic level. Behavioural rules are often asynchronous, discrete and contain stochastic elements. Agents may have different payoffs, individual preferences, or the topology of the interaction graph may be nontrivial. In such cases, the emerging, aggregated behaviour differs qualitatively from the mean-field analysis [Szabó and Fáth, 2007]. If the symmetry between agents is sufficiently broken, then we require a major shift in perspective from the *aggregate* level to the *agent* level. If we attempted to model such a system while including the now exponential increase in the number of system variables, standard analytical techniques would become largely inapplicable.

Fortunately, this technical challenge can be overcome by designing a simulation framework which is general enough to encompass the realistic properties and mechanisms of these models, without sacrificing flexibility or the ability to add new game instances. The design of these computational simulations lies at the core of this thesis. Large-scale simulations have already been successfully implemented to comprehend e.g., the emergence of social norms [Nowak and Sigmund, 1998; Pacheco et al., 2006a], the role of interaction networks [Santos and Pacheco, 2005; Santos et al., 2006a, 2008, 2006b], or, closely related to this thesis, interference in structured populations [Han et al., 2018]. Importantly, underlying networks of interaction and structured populations lie at the heart of this thesis' investigation. For arbitrary selection intensity, the computational complexity class of the problem suggests that there exists no efficient algorithm for the problem to be addressable using mathematical tools [Ibsen-Jensen et al., 2015]. While one solution has been proposed for weak selection on any graph or network [Allen et al., 2017], applying interference to complex networks is a task insurmountable using mathematical tools alone. Therefore, resorting to computer simulations is not simply convenient, but the only feasible approach.

Foundational to computational models are the interactions between individuals. Such systems, with a large number of interacting parts, are complex – a single component cannot determine system behaviour. Individuals may have negligible effects in isolation, but a significant effect when interacting with others. To model and analyse these complex systems and collective dynamics, agent-based simulations are common. Agent-based models (ABMs) consist of autonomous agents (in this case players), and serve as a bottom-up approach to studying complex systems. Analytical models, such as the ones described earlier in this chapter, use variables that characterise the entire system (top-down). The models described in this thesis are chosen based on an understanding of the interactions, but not to fit a certain expectation or outcome. In this case, the outcomes, or metrics, emerge from these lower-level interactions, which are often nonlinear and cannot be captured by aggregation.

One key property of the models described in further chapters is replicability. Thus, regardless of the specific framework or computational tools used to implement them, the results are easily reproducible. Throughout this thesis, we employed several existing ABM frameworks, and as the research matured, so did the choice of tools. This choice was not motivated by the ease of implementation, but by the computational complexity of the models. Increasing complexity requires additional computational power to maintain the robustness of the results, and optimisations provided by different frameworks can reduce runtime. We employed several popular frameworks for modelling ABMs: NetLogo [Wilensky and Rand, 2015], MASON [Luke et al., 2005], and Agents.jl [Datseris et al., 2022].

We present the key differences between the three frameworks employed in this thesis (including another popular alternative, Mesa [Masad and Kazil, 2015]) in Table 2.2. While all the models presented in this thesis lend themselves to feasible implementation using any framework (or even directly outside of an existing framework), we point out that NetLogo provides *out-of-the-box* visualisation techniques, and MASON is conveniently scalable, usually benchmarking well (although very verbose compared to the alternatives). The most recent option, Agents.jl, built on the Julia programming language, has arguably the most powerful features, particularly when it comes to networks, and benchmarks exceptionally. In spite of its very recent inception (it has been in its initial development for the major part of the duration of this thesis), its performance is best in class compared to the most widely used ABM framework implementations, many of which have had the benefit of many years of regular updates and development.

Table 2.2 **A comparison of four popular ABM frameworks:** core functionality, API/utilities, and performance [Allen et al., 2017]. Colours represent implementation quality as follows: blue – class leader, green – good, yellow – basic, and red – poor/nonexistent. Note we only present fields relevant to this thesis; for a full comparison refer to [Allen et al., 2017].

	Agents.jl 4.0	Mesa 0.8.7	NetLogo 6.1.1	MASON 20.0
Core	Core design decisions and aspects that cannot be changed or implemented by users			
Graph Space	Yes, and mutable	Only unidirectional	Link Agents (not a Space)	Networks (not a Space)
Grid Space	Yes	Yes (+Hexagonal)	Yes	Yes (+Hexagonal, Triangular)
Simulation termination	After n steps or user-provided boolean condition of model state	Explicitly written user loop	Manually by pressing a button on the interface, stop command in code	When Schedule is empty, or user provided custom finish function
Parameter types	Anything	Anything	Float64, Lists, Hashtables and Assoc. Arrays in the Table extension	Anything
Modeling and Analysis in the same language	Yes, Julia v1.5+	Yes, Python v3+	No	Yes, Java but designed to work within the console or GUI of the applet
Language ecosystem integration	By Design. Examples: black box optimization, differential equations	Any of Python's analytical tools can be used	Complex. Must create plugins or use Control API	Warned against (e.g. Random), provides custom types in place of Java primitives
Data collection	Any chosen parameter/property or function mapped over them. Aggregating and filtered aggregate functions	Any chosen parameter/property, aggregating functions, no conditional options	boolean, number, string and lists of these types	Inspectors track and chart any parameter/property. Entire model saved to disk via checkpointing, no custom export
Model complexity	Simple	Moderate	Simple	High

API and Utilities	How users interface with the framework, convenience functions			
Agent creation from values	Yes	No	Yes	No
Agent sample and replacement	Yes	No	No	No
New space types for API	Yes	No	No	No
Data collection low-level API	Yes	No	Yes	Yes, but only exportable via checkpointing
Scheduling	As added, by property, by type, filtered, random, custom	As added, random, staged	custom	custom

GUI for simulation setup	No	User implemented	Yes	User implemented
Numeric	Performance features, benchmarks where possible and lines of code (LOC) for implementations			
Maximum memory capacity	Hardware limits	Hardware limits	1GB; Manually expanded by increasing JVM heap	1GB; Manually expanded by increasing JVM heap
Flocking implementation	1 (normalised) 66 LOC	29.7x 120 LOC	10.3x 82 LOC	2.1x 369 LOC
Wolf-Sheep-Grass (grid) implementation	1 (normalised) 137 LOC	7.1x 273 LOC	2.1x 137 LOC	No implementation available
Forest Fire (grid) implementation	1 (normalised) 27 LOC	29.1x 61 LOC	4.1x 68 LOC	No implementation available
Schelling (grid) implementation	1 (normalised) 34 LOC	31.5x 63 LOC	8.0x 78 LOC	14.3x 248 LOC

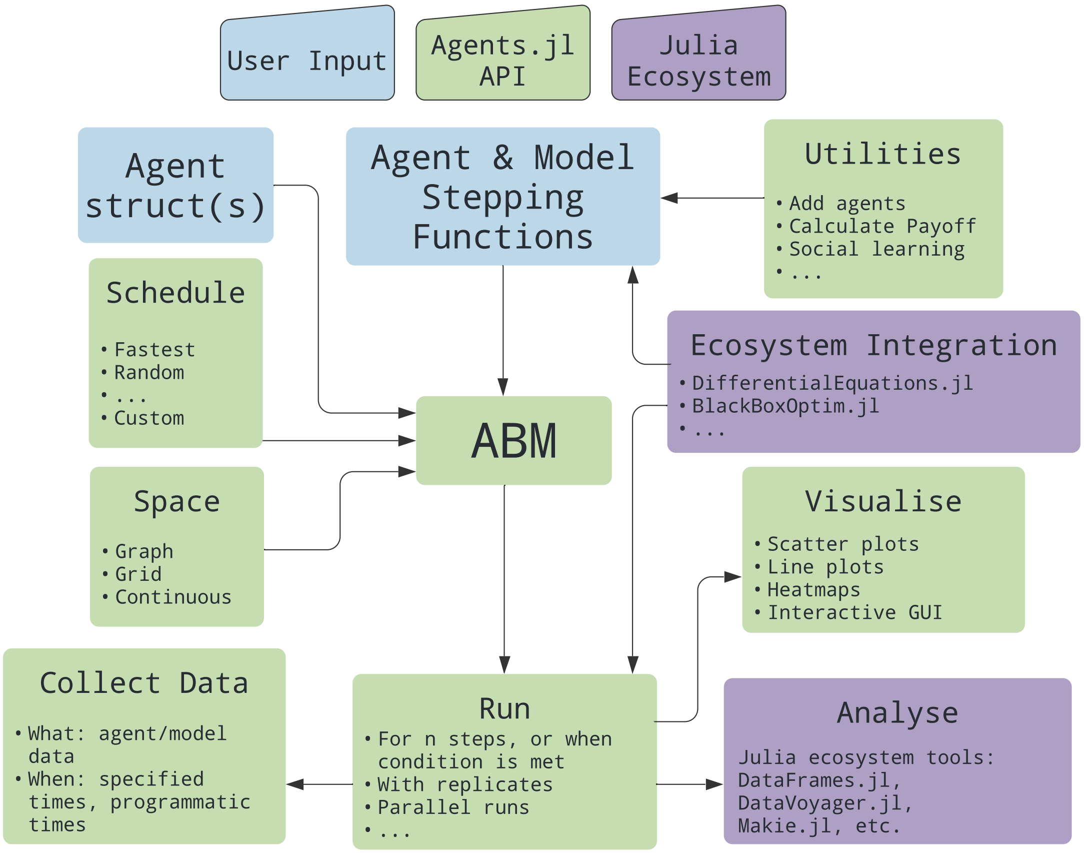

Fig. 2.3 **Flow chart representation of the Agents.jl framework.**

The design of ABM frameworks separates any simulation into simple components, minimising usage complexity. Typically, each of these components integrates with each other through the help of an API. In Figure 2.3, we illustrate in a flow chart one such implementation, in this case for the simple to understand Agents.jl framework. Following the breadcrumbs laid in Section 2.2, we can easily link EGT concepts to agent-based modelling. Agents are players, therefore they will minimally have some parameters describing their fitness and the strategy they follow. The properties of the game being played can be easily represented using payoff matrices, and functions to extract the outcomes of an interaction. The space corresponds to the network of interaction, which governs the connections an individual has, for the purposes of strategic interactions and imitation (more on that shortly). One step is equivalent to one generation in evolutionary time, and the specifics can be adjusted using appropriate stepping functions. One example of data collection is the mean of the frequencies for each strategy being employed in the population, something which can be visualised during runtime, or analysed at a later time, usually by taking into account all the replicates of an experiment.

Let us consider, then, a population of agents, which we have established are equivalent to players in the game theoretic sense. For simplicity, we will assume a simple Prisoner's Dilemma, which has two strategies, (always) cooperate and (always) defect (for a reminder, see Table 2.1). Initially, each agent in the population is assigned one of the two possible strategies, with equal probability. As a baseline scenario, we will assume a complete graph, in which every agent can interact with any other agent in the population. Thus, at each time step or generation, each individual plays the PD with everyone else. The score (fitness) for each agent is the sum of the payoffs in these encounters. We have now reached a decisive point in the modelling phase.

Previously, we had hinted that the replicator equation 2.3 is an useful tool to model social learning and the evolution of certain behaviours. How can we model this iteratively, then? A simple solution would be to consider a *deterministic* update rule. We select a random pairing of agents from the population – for instance a cooperator C and a defector D, with payoffs π_C and π_D, respectively. Under a deterministic update rule (imitation dynamics), the fitter replaces the less fit with probability $p = 1$. Similarly, selection within replicator dynamics always increases the fraction of the fitter strategy. On the other hand, this view is narrow, evolutionary changes are seldom so black and white, and in some cases the less fit individual may replace the more successful one. Occasional errors might be negligible in very large populations, yet may have pivotal effects in finite populations [Nowak et al., 2004]. We can beautifully model these effects by replacing the above probability p with a function derived from statistical physics, where stochastic effects are often described in terms of an effective temperature. Thus, we arrive at the Fermi function [Traulsen et al., 2006]:

$$p = \frac{1}{1 + e^{-\beta(\pi_C - \pi_D)}}, \tag{2.10}$$

which gives the probability that C replaces D, where the payoffs are extracted from a payoff matrix Π, such as one of the many exemplified in the section above. This process is often referred to as a *pairwise comparison* rule. The inverse temperature $\beta \geq 0$ controls the intensity of selection. For $\beta \to \infty$, we recover the same *deterministic* update as mentioned above, the fitter shall always replace the less fit. As β approaches zero, we increase stochastic effects, reducing the impact of fitness and increasing errors in judgement, creating conditions similar to those observed in lab experiments [Grujić and Lenaerts, 2020; Rand et al., 2013; Zisis et al., 2015]. For $\beta = 0$, selection is equivalent to neutral, random drift. Usually, it is assumed that the intensity of selection is an external factor (in relation to players' decision making). Indeed, there is an argument to be made that strategies could adopt different social learning mechanisms, e.g. adopting a different selection intensity based on individual factors, such as relative payoff differences. This could serve as potential future work, although it is beyond the scope of this thesis.

In addition to this, we consider mutation. With a given probability μ, this imitation process is replaced with a mutation, instead. Mutation (in genetics) is equivalent to behavioural exploration, an individual making a stochastic decision to switch to one of the available strategies (i.e. C or D in this example). This evolutionary process is simulated until a stationary state or a cyclic pattern of frequencies is reached.

The iterative model we have described above follows an *asymmetric* update rule. In other words, a single pair of agents is selected for imitation (or mutation) at every time step. We can also consider a *symmetric* (parallel update) rule instead. Instead of randomly allocating only one pair, each agent in the population will randomly select another to observe. Then, social learning will be performed as above, using either a deterministic or pairwise comparison. At the end of each generation, every agent will update their strategy at once.

2.4 Networks and Centrality Measures

Complex networks and nontrivial interaction structures lie at the heart of this thesis. Their significance was observed in the very early years of

EGT [Föllmer, 1974], but the systematic investigation of heterogeneity and structural issues still lies at the cutting edge of research [Szabó and Fáth, 2007]. By interaction structure, we imply a social network of contacts (a graph) [Santos et al., 2006a, 2008; Szabó and Fáth, 2007], which governs not physical (i.e. geographical) space, but social space. If two agents are connected, then they will engage in the strategic game being played, but also in social learning, thus neighbors can imitate one another. We note that scenarios exist in which these two networks (network of interaction and network of social learning) differ [Ohtsuki et al., 2007].

Changing the underlying structure of the interactions is important for several reasons. Firstly, homogeneous (fully connected) networks are idealised for the purposes of mathematical simplicity. In reality, there are few application domains in which individuals interact with everyone else in the population. Introducing spatiality while maintaining the homogeneity in the number of interactions (square lattice populations) is the first and most obvious way of introducing complexity to this previously idealised space. On the opposite side of the spectrum lie scale-free networks, which are extremely heterogeneous in terms of the inequality between nodes. Some individuals are supremely influential, while most are followers or followers of their followers, lowly connected and likewise not very influential. In this sense, scale-free networks allow us to study social diversity, but it is important to note that depending on the application domain, any one of the networks we study could be better suited to model the exact context of the interactions. While scale-free networks are undoubtedly realistic and well-mixed populations are idealised to a fault, it would be incorrect to assume that all real-life interactions are heterogeneous, or that none of them are homogeneous. In insular societies, for instance, it is likely we would observe something akin to homogeneous day-to-day interactions.

Let us first formalise the simplest scenario which we have been employing thus far.

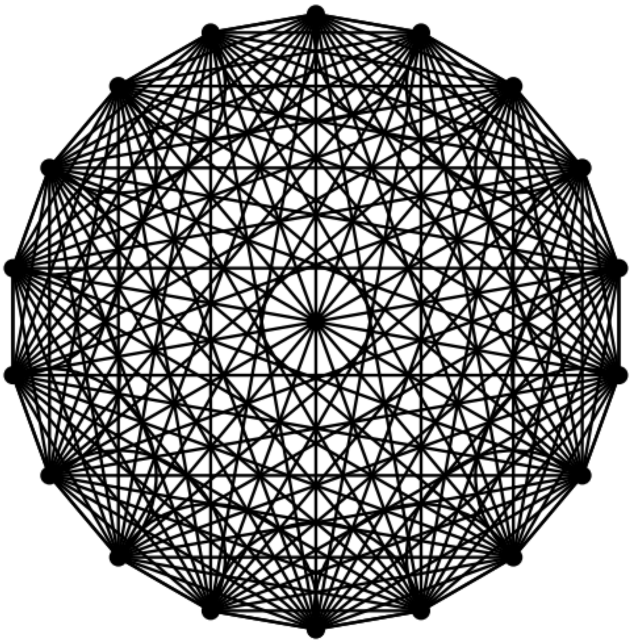

Fig. 2.4 **Complete graph.** Graphical representation of a well-mixed (fully-connected) network. Drawings of this graph date back to the 13th century, and it is sometimes referred to as a *mystic rose.*

2.4.1 Well-Mixed Networks

Infinitely large, well-mixed populations are an idealised version of reality, but nonetheless a useful simplification in the world of differential equations, which rest at the centre of many EGT models [Axelrod, 1984; Maynard Smith, 1982]. Dating back to the origins of graph theory and beyond, complete (*fully-connected*, well-mixed) graphs are the simplest form of network. All nodes (agents) are interconnected, thus interactions and learning occur homogeneously. All complete graphs are evidently their own maximal cliques, so any local observations are by definition global observations. The number of connections in a well-mixed network grows quadratically with the number of agents,

$$c = \frac{n(n-1)}{2}, \tag{2.11}$$

which means the average degree distribution is maximal, in contrast with observations on real networks [Barabasi, 2014].

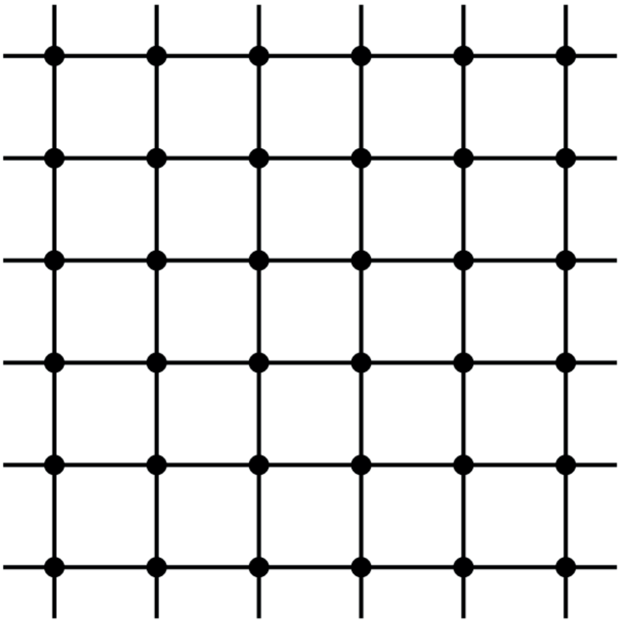

Fig. 2.5 **Structured populations,** represented using a square lattice graph, a regular spatial structure where each agent has four neighbors, with periodic boundaries.

2.4.2 Lattice Networks

If we consider spatiality, the simplest and most popular structure is the square lattice [Szabó and Fáth, 2007]. We can also refer to populations placed in lattice graphs as *structured* populations, although that term can be broader in scope than this narrow example. Commonly, we consider the von Neumann neighbourhood (i.e. four-neighbour lattice with average connectivity $z = 4$), but Moore neighbourhoods are also possible (connections extend to the next-nearest neighbours, $z = 8$). Periodic boundary conditions (i.e. the edges of the graph wrap around to the other sides, to maintain $z = 4$), mean that these systems become translation invariant, which lends itself to mathematical analysis using methods developed in solid state theory and non-equilibrium statistical physics [Szabó and Fáth, 2007]. While still relatively homogeneous, structured populations are not only more realistic settings than well-mixed worlds, but also allow us to consider local measures, as every node is only minimally connected to the others in the network.

Despite their apparent simplicity, square lattices enjoy remarkable popularity and representation in game theoretical models [Perc et al., 2013;

Szabó and Fáth, 2007]. For instance, they were employed to demonstrate that local interactions can maintain cooperation indefinitely in the Prisoner's Dilemma with symmetric updates [Nowak and Sigmund, 1992; Nowak et al., 1994]. These results have raised a lot of additional questions, and inspired a wide breadth of models aiming to understand the properties of spatial games. Later, the coexistence of cooperators and defectors in spatial populations was studied systematically for general, 2×2 payoff games (such as the Prisoner's Dilemma and other games) [Hauert, 2001; Lindgren and Nordahl, 1994]. Ever since, there has been a continued interest in exploring the simple spatial structures provided by a square lattice, using a wide variety of settings and mechanisms for cooperation [Han et al., 2018; Hauert et al., 2002; Lerat et al., 2013; Perc et al., 2013; Pinheiro et al., 2012b].

2.4.3 Scale-Free Networks

Real-world networks are dynamic and inherently heterogeneous [Barabási and Albert, 1999; Dorogovtsev, 2010; Newman, 2003]. Networks evolve with new nodes entering and creating connections to already existing nodes [Dall'Asta et al., 2006]. Several works have unveiled how network structural *heterogeneity* plays a key role in both the evolution of cooperation [Dercole et al., 2019; Poncela et al., 2007; Santos et al., 2006a, 2008] or the emergence of fairness [Sinatra et al., 2009]. Indeed, in the case of cooperation, it may enhance the emergence and resilience of cooperation, inducing cooperative agents (or nodes) to create assortative clusters, where they reciprocate cooperation [Di Stefano et al., 2020; Santos et al., 2006a]. Similarly, in the UG, the presence of highly connected nodes (or hubs) changes the distribution of strategies, due to their ability to get a large reward with a broad range of strategies and thus to rule the final behaviour of the entire population [Sinatra et al., 2009]. One of the main targets of this work is therefore to measure the impact of network properties and structural heterogeneity in the evolutionary dynamics of interference behaviours. To measure these effects, we will make use of two types of scale-free networks [Barabási and Albert, 1999; Dorogovtsev, 2010; Newman, 2003], generated through two growing network models: the *Barabási and Albert* (BA) model [Barabási and Albert,

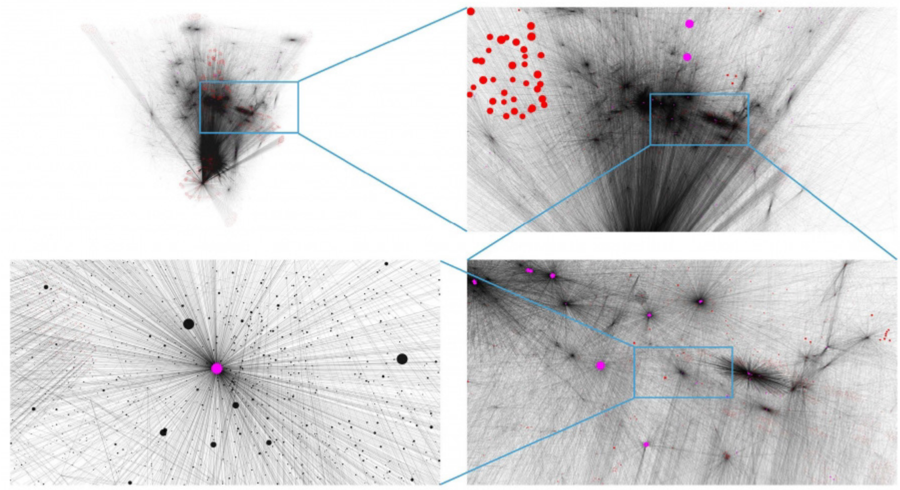

Fig. 2.6 **The Topology of the World Wide Web.** Snapshots of the World Wide Web sample mapped out by Albert et al. [1999]. Each image in the sequence is increasingly magnified. The first panel shows the whole network, offering a global view of the 325729 nodes. Nodes with more than 50 links are coloured in red, and hubs (more than 500 links) in purple. The close-ups highlight the sparsity of the hubs. [Barabási, 2016]

1999], and the *Dorogovtsev-Mendes-Samukhin* (DMS) model [Dorogovtsev et al., 2000].

The Barabási and Albert (BA) model [Barabási and Albert, 1999] is one of the most famous models used in the study of highly heterogeneous, complex networks. The main features of the BA model are that it follows a *preferential attachment* rule, has a small clustering coefficient, and a typical *power-law degree distribution*. In order to explain preferential attachment, let us describe the construction of a BA network. Starting from a small set of m_0 interconnected nodes, each new node selects and creates a link with m older nodes according to a probability proportional to their degree. The procedure stops when the required network size of N is reached. This will produce a network characterised by a power-law distribution,

$$p_k \sim k^{-\gamma}, \tag{2.12}$$

where the exponent γ is its *degree exponent* [Barabási, 2016]. There is a high degree correlation between nodes, and the degree distribution is typically skewed with a long tail. There are few hubs in the network that attract an increasing number of new nodes which attach as the network grows (in a typical *"rich-get-richer"* scenario). The power-law distribution exhibited by BA networks resembles the heterogeneity present in many real-world networks (see Figure 2.6). However, they are also defined by low clustering coefficients, which means they cannot always be used to approximate realistic settings [Su et al., 2016].

To build heterogeneous networks with a large clustering coefficient, Dorogovtsev et al. [2000] have proposed the eponymous DMS model. This model follows a similar method of construction as the BA model, and is also exemplary of the preferential attachment rules and follows a power-law degree distribution. Crucially, each new node connects with the two extremities of m ($m \geq 2$) randomly chosen edges, instead, therefore forming characteristic triangular motifs whenever a new node is added to the network. Since the number of edges arriving to any node reflects its degree, the probability of attaching the new node to an old node is proportional to its degree and preferential attachment is recovered. The degree distribution is therefore the same as the one of a BA model, and the degree-degree correlations are also equal [Dall'Asta et al., 2006]. However, the clustering coefficient is large, and more accurately mimics many realistic social networks [Barrat and Pastor-Satorras, 2005; Su et al., 2016]. The average connectivity for both types of scale-free networks is $z = 2m$.

There exists significant disagreement within the field of network science as to how frequently scale-free networks can be encountered in the real world. Some proponents claim they are ubiquitous [Barabási, 2009], while others argue that they are rare [Broido and Clauset, 2019]. Regardless of their prevalence, it is clear that there exists a great deal of structural diversity between real-world networks, and it is prudent to extend our investigation beyond simple complete graphs. Moreover, in the context of EGT, scale-free networks imply more than the underlying interaction structure. Heterogeneous graphs can portray social diversity [Santos et al., 2008], and the inherent inequality that exists between agents. In that sense, we can think of well-mixed settings as agents being completely equal, and scale-free dynamics as the discrepancy which exists between individuals in the real world.

2.4.4 Centrality Measures

Given the deliberate inequality between nodes, we need a measurement to distinguish important/influential nodes from less important ones. The degree centrality is the oldest measure of importance or influence ever used in network science [Boldi and Vigna, 2014]. It denotes the number of neighbours of the node i, namely it measures the number of edges of node i. By definition, degree centrality is normalised using the total number of nodes, or the maximal possible degree, $n-1$, to obtain a number between 0 and 1. We can therefore define the degree centrality:

Definition 2.4.1 (Degree centrality)
Degree centrality, denoted by NI-deg or x_i^{deg}, is defined as follows:

$$x_i^{deg} = deg_i = \frac{k_i}{n-1}, \tag{2.13}$$

where k_i is the degree of the node i and $n-1$ is the total number of nodes. The degree k_i of a node i is given by: $k_i = \sum_{j=1}^{n} A_{ij}$, where A is the adjacency matrix of a finite graph, populated with pairs of vertices which are adjacent (i.e. connected).

Despite its simple definition, degree centrality is often a highly effective measure of the influence or importance of a node, since people with more connections tend to be more influential in a social network [Bloch et al., 2019; Newman, 2008]. The reason why we define degree centrality by using the previous normalised definition, and not simply the degree, is that it allows us to analyse it in a better way and compare two nodes that belong to two different networks regardless of network size [Saxena and Iyengar, 2020].

Although degree centrality constitutes a simple but effective centrality measure, giving some insight into the connectivity or popularity of node i, it lacks potentially important aspects of the architecture of the network and a node's position in the network. Indeed, it represents a *local* centrality measure, including the local information of the node, but not considering the global connectivity, and only the quantity and not the quality of connections [Di Stefano et al., 2015, 2020; Scatà et al., 2016].

In our work, we consider another measure of centrality, the eigenvector centrality [Bonacich, 2007]. Differently from the degree centrality, the eigenvector centrality is a spectral measure since its definition is based on the spectral properties of the adjacency matrix. Eigenvector centrality represents a related measure of prestige, since the importance of a node i depends on the prestige of its neighbours [Bloch et al., 2019; Perra and Fortunato, 2008]. In other words, this centrality measure acknowledges that not all connections are equal, but connections to nodes who are themselves influential will make a node more influential [Newman, 2008; Perra and Fortunato, 2008]. Eigenvector centrality is computed by assuming that the centrality of node i is proportional to the sum of centrality of node i's neighbours. Central nodes are the most influential nodes which can influence the behaviours of their neighbouring nodes. We can therefore define the eigenvector centrality:

Definition 2.4.2 (Eigenvector centrality)
Eigenvector centrality (also called *eigencentrality* or *prestige score*), denoted by NI-eig or x_i^{eig}, is defined as follows:

$$x_i^{eig} = eig_i = \frac{1}{\lambda} \sum_{j=1}^{n} A_{ij} x_j, \tag{2.14}$$

where λ is a positive constant or proportionality factor.

Defining the vector of centralities $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, we can rewrite the previous definition in matrix form as $\lambda \cdot \mathbf{x} = \mathbf{A} \cdot \mathbf{x}$, hence we see that $\mathbf{x}$ is an eigenvector of the adjacency matrix with eigenvalue λ. Since the centralities must be non-negative, it can be shown (using the Perron–Frobenius theorem) that λ must be the largest eigenvalue of the adjacency matrix and $\mathbf{x}$ the corresponding eigenvector [Newman, 2008]. Thus, from its definition it is clear how the eigenvector centrality depends not only the number or quantity of links or edges of each node, but also the quality of such connections [Perra and Fortunato, 2008; Scatà et al., 2016]. Indeed, even if a large number of connections increases the centrality measure, a node with a smaller number of high-quality nodes may still outrank one with a larger number of low-influential nodes. Eigenvector centrality and its variants have been used in different contexts, e.g. the well-known PageRank centrality [Bianchini et al., 2005; Page et al., 1999; Perra and Fortunato, 2008] used by the Web search engine Google to rank Web pages.

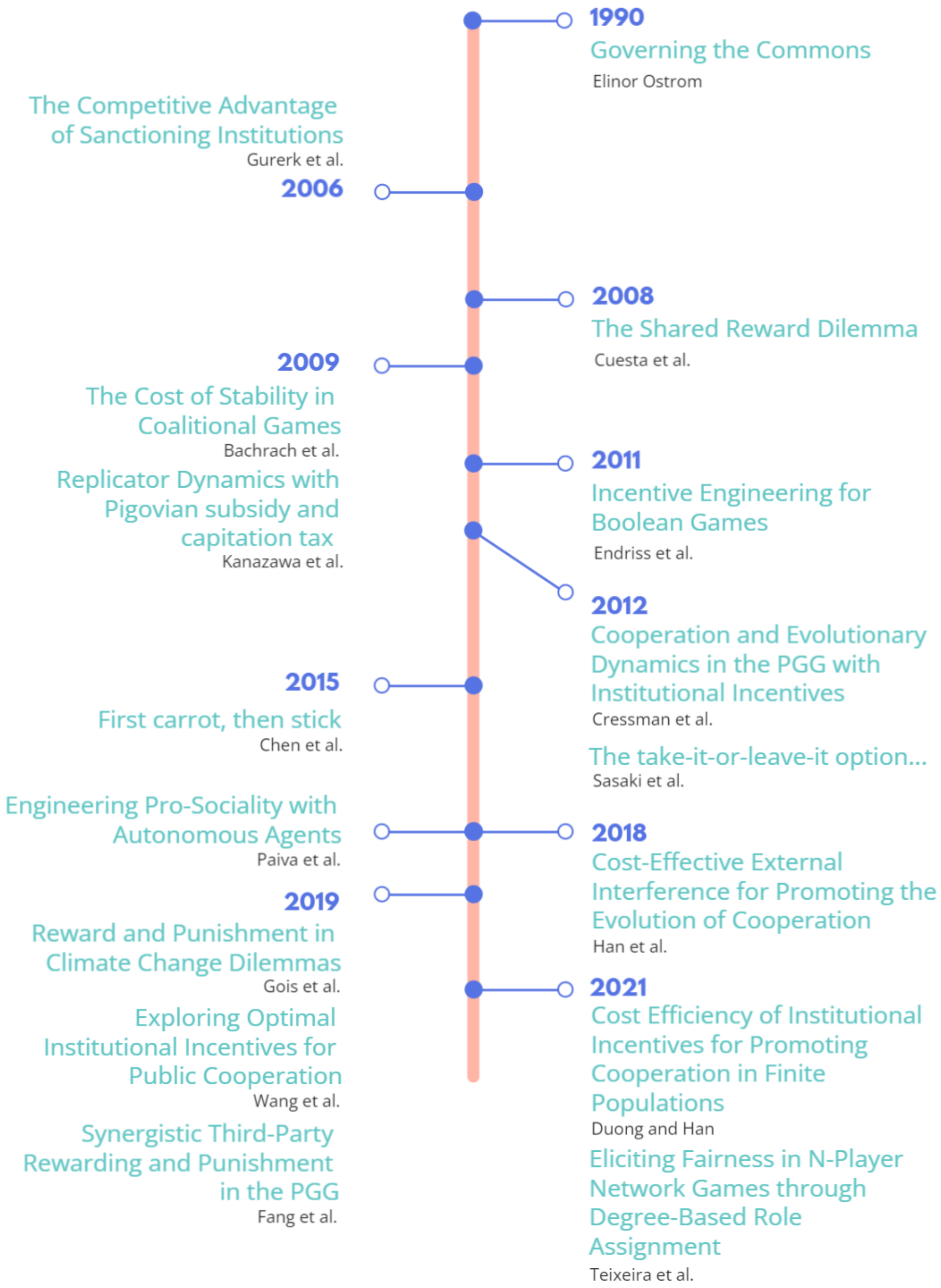
1990
Governing the Commons
Elinor Ostrom
The Competitive Advantage of Sanctioning Institutions
Gurerk et al.
2006
2008
The Shared Reward Dilemma
Cuesta et al.
2009
The Cost of Stability in Coalitional Games
Bachrach et al.
Replicator Dynamics with Pigovian subsidy and capitation tax
Kanazawa et al.
2011
Incentive Engineering for Boolean Games
Endriss et al.
2012
Cooperation and Evolutionary Dynamics in the PGG with Institutional Incentives
Cressman et al.
The take-it-or-leave-it option...
Sasaki et al.
2015
First carrot, then stick
Chen et al.
Engineering Pro-Sociality with Autonomous Agents
Paiva et al.
2018
Cost-Effective External Interference for Promoting the Evolution of Cooperation
Han et al.
2019
Reward and Punishment in Climate Change Dilemmas
Gois et al.
Exploring Optimal Institutional Incentives for Public Cooperation
Wang et al.
Synergistic Third-Party Rewarding and Punishment in the PGG
Fang et al.
2021
Cost Efficiency of Institutional Incentives for Promoting Cooperation in Finite Populations
Duong and Han
Eliciting Fairness in N-Player Network Games through Degree-Based Role Assignment
Teixeira et al.

Fig. 2.7 **A brief chronology of institutions and interference.**

2.5 Overview of Related Work

Institutions, as mechanisms for the evolution of cooperation, have enjoyed much attention in the game theoretic and evolutionary game theoretic literature. Elinor Ostrom [1990] approached this topic using a series of empirical studies aiming to solve the so-called *Tragedy of the Commons* [Hardin, 1968]. In this seminal work, Ostrom laid the foundation for what would become a prosperous endeavour of explaining the emergence of institutions [Aoki, 2001], and the characteristics of successful institutions. Since then, several research agendas have continued to show that institutions can emerge from the collective action of the individuals faced with collective risk, such as ever-depleting common pool resources. Sigmund et al. [2010] have shown that social learning leads to collaborative contributions towards pool punishment, which later led to a concerted effort to study institutions for governing climate change, often envisioned as a common pool resource [Ostrom, 2010; Pacheco et al., 2014; Vasconcelos et al., 2013]. This line of inquiry is still being explored to this day [Garcia and Traulsen, 2019; Sasaki and Uchida, 2014], with continuous efforts being made to explain why institutions emerge, especially as emergent properties of the actions of individuals facing common pool resource problems. While these works explain how institutions come into being, they do not model institutions as exogenous to the system. We have seen that individuals can organise themselves for a collective purpose, but comparatively few studies explore what happens afterwards.

Motivated by findings on the detrimental effects of punishment [Fehr and Fischbacher, 2003], Gürerk et al. [2006] were among the first to consider institutions as exogenous entities, as opposed to coalitions between participants in the interactions being played. While still enforcing the prevalent idea of peer incentives [Fehr and Gachter, 2002; Sigmund et al., 2001], they asked whether sanctioning institutions are vulnerable to the same pitfalls that were suggested to exist in the case of peer punishers in the absence of institutions. Comparing sanction-free and sanctioning institutions experimentally, they found that participants faced with a public-goods dilemma would overwhelmingly choose the latter. In a *voting-with-one's-feet* competition, they showed that sanctioning institutions were the undisputed winner, providing empirical evidence for the existence of strong reciprocators and conformist behaviour. Despite severe individual losses, the choice of a sanctioning institution led to a strong desire for participants to cooperate and punish free-riders. This work served as a fledgling medium between the prominent literature on the emergence of institutions and peer incentives.

External interference in the true sense was not considered until later, when Cuesta et al. [2008] introduced the Shared Reward Dilemma. Starting from a PD, they studied a reward mechanism by which a fixed sum was distributed among all cooperators. This is a particularly interesting setting, as an increase in the number of cooperators appropriately scales down their endowment, producing complex dynamics which cannot easily be studied using a classical (static) analysis. Through evolutionary dynamics (using the replicator equation), they found that cooperation can be promoted only with very high initial rewards, and maintained indefinitely with lower amounts. They conclude that promoting cooperation through rewards is not trivial, which prompted several more works to explore this dilemma in structured populations and random networks [Jiménez et al., 2008, 2009]. Importantly, these works on lattices and random networks show that this reward mechanism is never detrimental to cooperation in these settings, and that it can lead to the emergence of cooperation which is resilient against invasion even after the endowments have ceased [Jiménez et al., 2008, 2009].

Among the first to consider the costs of promoting specific behaviours were Bachrach et al. [2009], who investigated the stability of coalitional games using external payments. In their work, an external party offers a supplemental payment to a coalition, which is then divided among the players who have not deviated from this agreement. Using a classical approach, they provided general bounds on the cost of stabilising several classes of games, focusing especially on weighted voting games, which are used to model political decision-making and cooperation in multi-agent settings. Similarly to the works by Cuesta et al. [2008], the incentive mechanism seeks to divide a fixed sum among all cooperative agents.

Continuing the trend of studying completely rational interactions with a classical game theoretic approach, several authors studied taxation schemes on Boolean games [Endriss et al., 2011; Levit et al., 2013; Wooldridge, 2012]. In a population of self-interested agents with clear goals, they levied taxes in an effort to steer economic equilibria towards a desirable goal. This was mostly resolved using high taxation schemes aimed at undesirable behaviour, but the authors point to several issues that arise with such an incentive scheme, such as an unreasonably high tax burden on society or implied inequity in the scheme. Kanazawa et al. [2009] have studied rate and capitation taxes in selfish routing games, in which either part of the players' payoffs or fixed amounts are removed from some of the players and later redistributed as subsidies. Modelled using replicator dynamics, they also provide several solutions that characterise the equilibria of taxation schemes, among which is a capitation tax which can make a desirable system state asymptotically stable.

Combining an analytical study with experiments on human participants, Cressman et al. [2012, 2013] devised an external incentive program that aimed to increase individual contributions in the Public Goods Game. Deviating significantly from previous works, they put forward a lottery-based system for the eligibility of incentives. Using three types of institutions: reward, punishment and mixed incentives (both reward and punishment), they posited that an individual would be eligible for the incentive as a probability based on their previous action. Contributing high would increase their probability of being selected for a reward, for example, while free-riding would increase their probability of being punished. Their analytical results were partly confirmed experimentally, finding that a combined incentive scheme leads to the highest individual contributions among participants.

Continuing the studies done on Public Goods Games, Sasaki et al. studied both reward and punishment with individual endowments, but with the addition of optional participation [Sasaki and Uchida, 2014; Sasaki et al., 2012]. In contrast to previous studies done on pool punishment [Gürerk et al., 2006], the punishing institution was considered a *Leviathan*-like administration, such as in Hobbes' eponymous book. In other words, a true authority, external to the interactions, but now given the role of enforcing contracts. By paying in the participation fee, players would subject themselves to either reward or punishment, as the institution saw fit, or not play at all. While optional participation only minimally affected the outcomes of rewarding cooperation, it allowed the punishment of free-riders at a greatly reduced cost. This *take-it-or-leave-it* option led to punishing institutions reducing the cost of promoting cooperation in comparison to institutions with positive incentives, making this the first work to contemplate the costs associated with incentives. This argument was limited somewhat, as they only measured per-capita investment amount, which is not always a good indicator of total accumulated costs.

Chen et al. [2015] provided an important contribution to the previously described model [Sasaki and Uchida, 2014; Sasaki et al., 2012], by introducing adaptive hybrid incentives. Maintaining the setting with the exception of elective participation, the institution could now arbitrarily switch from reward to punishment when cooperation rose above a certain threshold. Using the adage *first carrot, then stick*, they showed the effectiveness of reward in establishing cooperation, and of punishing in eliminating defection. This mechanism proved surprisingly effective and widely-applicable in the context of well-mixed and structured populations. Moreover, these findings were later confirmed experimentally using an almost identical setting with human participants [Hou et al., 2019].

Winning the Blue Sky ideas paper award at AAAI'18, Paiva et al. [2018] envisioned a future in which autonomous agents are used to foster pro-social behaviour in a hybrid society of humans and machines. This paper laid the bases for pro-social computing, calling to action a new line of research that aims to understand how pro-sociality can be engineered in these hybrid societies. This marked a period of renewed interest in the topic of external interference, and reinforced the significance and timeliness of this agenda. Shortly thereafter, Han and Tran-Thanh [2018]; Han et al. [2018] bridged the gap between external interference and cost-effectiveness, highlighting the complex relationship between minimising cost and bettering outcomes. The initial analytical study on well-mixed populations showed that cost-efficiency is highly sensitive to changes in the intensity of selection, and proposed a simple, but effective class of interference based on the composition of the population. Later, several complex classes of interference proved to be even more effective at reducing cost in spatial populations, where local neighbourhood information could be used to discern which cooperators would be the best targets for investment [Han et al., 2018]. Recently, the study on well-mixed populations has been expanded to encompass punishment and finite populations, providing a selection-dependant calculation for the optimal cost of both positive and negative incentives [Duong and Han, 2021b].

The problem of how an external decision-maker can steer populations towards a desirable state has not been solved rigorously thus far, and several research programs are keen on answering the many gaps left open in the existing literature. Góis et al. [2019] have successfully applied the *first carrot, then stick* approach to climate change agreements. Fang et al. [2019] have shown the synergistic effects of reward and punishment being applied simultaneously in the spatial PGG, while Wang et al. [2019] approached the well-mixed setting of the PGG using optimal control theory. Franks et al. [2013, 2014] first quantified the role of influencers in networks, though not through external incentives, and Teixeira et al. [2021] provided a similar approach to show that assigning specific roles to lowly-connected nodes can lead to the emergence of fairness. Each of these contributions is bringing us one step closer to understanding which are the optimal approaches available to an institution wishing to promote cooperation, fairness and other pro-social behaviours, but the answers are still out of reach.

2.6 Open Questions

Following the exposition of the existing work on the subject of external interference, we identified several gaps in the literature. In the following chapters, we address each of these broad open questions:

- Do previous findings (from well-mixed and lattice settings) apply in the presence of social diversity? (Chapter 3)
- How does network heterogeneity affect the efficiency of local and global interference schemes? (Chapters 3 and 4)
- Can centrality measures be exploited to reduce interference costs and/or improve pro-social behaviour outcomes? (Chapters 3 and 4)
- How does players' role asymmetry in an interaction impact the choice of interference schemes available to the institution? (Chapter 4)

- What is the effect of non-negligible behavioural exploration or mutation rates on the cost and intensity of interference required to maintain fairness? (Chapter 4)
- Can the threat of punishment serve as a deterrent to defection, thereby reducing costs and improving social welfare? (Chapter 5)

3 Promoting Cooperation in Scale-Free Networks

Be warned that if you wish, as I do, to build a society in which individuals cooperate generously and unselfishly towards a common good, you can expect little help from biological nature. Let us try to teach generosity and altruism, because we are born selfish.

—Richard Dawkins, *The Selfish Gene*

In which we study the effects of social diversity on the design of interference mechanisms aiming to promote cooperation. Here[1], we show that interference on scale-free networks is not trivial. In particular, we show that the inconsiderate distribution of incentives can lead to the exploitation of cooperators. We present which mechanisms are more efficient at fostering cooperation, arguing that social diversity and the network's clustering coefficient both play a key role in the choice of interference mechanisms available to institutions wishing to promote cooperation.

3.1 Introduction

The design of mechanisms that encourage pro-social behaviours in populations of self-regarding agents is recognised as a major theoretical challenge within several areas of social, life and engineering sciences. It is ubiquitous in real-world situations, not least ecosystems, human organisations, technological innovations and social networks [Han et al., 2019; Raghunandan and Subramanian, 2012; Santos et al., 2006a; Sigmund et al., 2001]. In this context, cooperation is typically assumed to emerge from the combined actions of individuals within the system. However, in many scenarios, such behaviours are advocated and promoted by an external party, which is not part of the system, calling for a new set of heuristics capable of *engineering* a desired collective behaviour in a self-organised complex system [Penn et al., 2010]. Among these heuristics, several have been identified as capable of promoting desired behaviours at a minimal cost [Chen et al., 2015; Han and Tran-Thanh, 2018; Han et al., 2018]. However, these studies neglect the diversified nature of contexts and social structures which define real-world populations. Here, we analyse the impact of diversity by means of scale-free interaction networks with dissimilar levels of clustering, and test various interference mechanisms using simulations of agents facing a cooperative dilemma.

For instance, if one considers a near future, where hybrid societies comprising humans and machines shall prevail, it is important to identify the

[1] The model and part of the results presented in this chapter were reported in [Cimpeanu et al., 2019].

most effective incentives to be included to leveraging cooperation in such hybrid collectives [Paiva et al., 2018]. In a different context, let us consider a wildlife management organisation (e.g., the WWF) that aims to maintain a desired level of biodiversity in a particular region. In order to do that, the organisation, not being part of the region's ecosystem, has to decide whether to modify the current population of some species, and if so, then when, and in what degree to *interfere* in the ecosystem (i.e., to modify the composition of the population) [Levin, 2000]. Since a more impactful intervention typically implies larger costs in terms of human resources and equipment, the organisation has to achieve a balance between cogent wildlife management and a low total investment cost. Moreover, due to the evolutionary dynamics of the eco-system (e.g., frequency and structure dependence) [Hofbauer and Sigmund, 1998; Maynard Smith, 1982; Santos et al., 2006a], undesired behaviours can reoccur over time, for example when the interference was not sufficiently strong in the past. Given this, the decision-maker also has to take into account the fact that it will have to repeatedly interfere in the eco-system in order to sustain levels of biodiversity over time. That is, they must find an efficient interference mechanism that leads to their desired goals, while also keeping in mind potential budget concerns.

Specifically, we consider populations of individuals distributed in a scale-free network, who interact with their neighbours via the one-shot Prisoner's Dilemma (PD), where uncooperative behaviour is preferred over cooperation [Santos et al., 2006a; Sigmund et al., 2001]. As an outside decision maker, we aim to promote cooperation by interfering in the system, rewarding particular agents in the population at specific moments. The research question here is to identify when and how much to invest (in individuals distributed in a network) at each time step, in order to achieve cooperation within the system such that the total cost of interference is minimised, taking into account the fact that individuals might have different levels of social connectivity. For instance, we might wonder whether it is sufficient to focus the investment only on highly connected cooperators, as they are more influential. Would targeting influencers reduce overall costs? Do we need to take into account a neighbourhood's cooperativeness level, which was shown to play an important role in square lattice networks [Han et al., 2018]? Also, when local information is not available and only global statistics can be used in the decision making process, how different are the results in heterogeneous networks, in comparison to regular graphs (i.e. homogeneous networks)?

To answer these questions, this chapter will systematically investigate different general classes or approaches of interference mechanisms, which are based on *i)* the global population statistics such as its current composition, *ii)* a node's degree centrality in the network (see Section 2.4.4 in Chapter 2) and *iii)* the neighbourhood properties, such as local cooperativeness level.

3.2 Model and Methods

3.2.1 Prisoner's Dilemma on Scale Free Networks

We consider a population of agents on scale-free networks of contacts (SF NoCs) — a widely adopted heterogeneous population structure in population dynamics and evolutionary games (for a detailed description, see Section 2.4.3). We focus our analysis on the efficiency of various interference mechanisms in spatial settings, adopting an agent-based model directly comparable with the setup of recent lab experiments on cooperation [Rand et al., 2014]. Moreover, we select an initial number of nodes $m_0 = 2$, with two additional edges being created at every time step of network generation (i.e. $m = 2$, for a detailed description please see Section 2.4.3). This produces networks of average connectivity $z = 4$, serving as a direct comparison between this work and other studies performed on structured populations [Han et al., 2018].

Initially each agent in a population of size N is designated either as a cooperator (C) or defector (D) with equal probability. Agents' interaction is modelled using the one-shot Prisoner's Dilemma game (for a discussion, see Section 1.2), where mutual cooperation (mutual defection) yields the reward R (penalty P) and unilateral cooperation gives the cooperator the sucker's payoff S and the defector the temptation T. As a popular interaction model of structured populations [Szabó and Fáth, 2007], we adopt the following scaled payoff matrix of the PD (for row player):

$$\begin{array}{cc} & \begin{array}{cc} C & D \end{array} \\ \begin{array}{c} C \\ D \end{array} & \begin{pmatrix} 1 & 0 \\ b & 0 \end{pmatrix}, \end{array}$$

with b ($1 < b \leq 2$) representing the temptation to defect . We adopt this weak version of the Prisoner's Dilemma in spite of cooperation prevalence shown in previous works on scale-free networks [Santos et al., 2008], so as to have a direct comparison with studies on the effects of rewarding mechanisms in different types of networks [Han et al., 2018].

At each time step or generation, each agent plays the PD with its immediate neighbours. The score for each agent is the sum of the payoffs in these encounters. Before the start of the next generation, the conditions of interference are checked for each agent and, if they qualify, the external decision maker increases their payoff. Multiple mechanisms (i.e. multiple conditions) can be active at once, but the individual investment cannot be applied more than once; the schemes determine the eligibility for investment.

At the start of the next generation, each agent's strategy is updated using one of two social learning paradigms – a *deterministic*, or a *stochastic* rule (for a thorough explanation see Section 2.2 in Chapter 2). Using a deterministic update rule, each agent will choose to imitate the strategy of its highest scored neighbour [Nowak and May, 1992; Szabó and Fáth, 2007]. In the stochastic case, instead of copying the highest scored neighbour, at the end of each generation an agent A with score f_A chooses to copy the strategy of a randomly selected neighbour agent B with score f_B with a probability given by the Fermi rule [Traulsen et al., 2006]: $(1 + e^{(f_A - f_B)/K})^{-1}$, where K denotes the amplitude of noise in the imitation process [Szabó and Fáth, 2007]. In line with previous works and lab experiments [Rand et al., 2013; Szabó and Fáth, 2007], we set $K = 0.1$ in our simulations. Our analysis will be based on this standard evolutionary process in order to focus on understanding the cost-efficiency of different interference mechanisms.

We simulate this evolutionary process until a stationary state or a cyclic pattern is reached. The simulations converge quickly in the case of deterministic update, with the exception of some cyclic patterns which never reach a stationary state. Because this work studies cost effective intervention, these rarely-occurring patterns which inherently invite very large total costs are escaped early by running simulations for 75 generations (deterministic update) and 500 generations (stochastic update), at which point the accumulated costs are excessive enough for this mechanism to not be of interest.

The difference in the final number of generations accounts for the slower convergence time associated with stochastic dynamics. Moreover, the results are averaged for the last 25 generations of the simulations for a clear and fair comparison (e.g. due to cyclic patterns). In order to improve accuracy related to the randomness of network topology in scale-free networks, each set of parameter values is ran on 10 different pre-seeded graphs for both types of SF NOCs. Furthermore, the results for each combination of network and parameter values are obtained from averaging 30 independent realisations. It is important to note that the distribution of cooperators and defectors on the network is different for every realisation.

Note that we do not consider mutations or random explorations when employing a determinstic update rule. Thus, whenever the population reaches a homogeneous state (i.e. when the population consists of 100% of agents adopting the same strategy), it will remain in that state regardless of interference. Hence, whenever detecting such a state, no further interference will be made. Errors can sometimes occur under the presence of stochastic imitation, thus we never preemptively pause these simulations. Given the difference in convergence time, network size and stopping conditions, we do not directly compare the total costs between these two paradigms.

3.2.2 Cost-Efficient Interference in Networks

We aim to study how one can efficiently interfere in spatially heterogeneous populations to achieve high levels of cooperation while minimising the cost of interference. An investment decision consists of a cost $\theta > 0$ to the external decision-making agent/investor, and this value θ is added as surplus to the payoff of each suitable candidate. In order to determine cost-efficiency, we vary θ for each proposed interference strategy, measuring the total accumulated costs to the investor. Thus, the most efficient interference schemes will be the ones with the lowest relative total cost.

Moreover, in line with previous works on network interference [Chen et al., 2015; Han and Tran-Thanh, 2018; Han et al., 2018], we compare global interference strategies where investments are triggered based on network-wide information, local neighbourhood information, and, lastly, node centrality information.

In the *population-based* (**POP**) approach, a decision to invest in desirable behaviours is based on the current composition of the population. We denote x_c the fraction of individuals in the population adopting cooperative behaviour. Namely, an investment is made if x_c is at most equal to a threshold p_c (i.e. when $x_c \leq p_c$), for $0 \leq p_c \leq 1$. They do not invest otherwise (i.e. $x_c > p_c$). The value p_c describes how rare the desirable behaviours should be to trigger external support.

In the *neighbourhood-based* (**NEB**) approach, committing an abuse of notation, a decision to invest is based on the fraction x_c of neighbours of a focal individual with the desirable behaviours, calculated at the local level. Investment happens if x_c is at most equal to a threshold n_c (i.e. when $x_c \leq n_c$), for $0 \leq n_c \leq 1$; otherwise, no investment is made.

As the presence of structural heterogeneity in scale-free networks introduces a level of inequality between nodes in terms of influence, we also examine a *node-influence-based* (**NI**) approach. To achieve this, we make use of degree centrality (for a definition see Section 2.4.4). We denote by x_i the node's centrality measure. The decision-maker invests in a cooperator node C when the value of its degree centrality is above a threshold of influence c_I, for $0 \leq c_I \leq 1$. Otherwise, i.e. when $0 \leq x_i < c_I$, no investment is made. The value c_I describes how influential a cooperator node should be to trigger an investment into its survival.

For the POP and NEB schemes, the threshold signifies an increase in the number of nodes that satisfy the requirements for investment. In other words, a threshold of 1 means always investing in all nodes which follow the desired strategy. Conversely, a lower threshold implies a more careful approach to investment, whereby the exogenous agent is stricter in their selection of suitable candidates. The opposite is true for NI, as a value of 1 implies only the most connected individual(s) is eligible for investment; whereas a value of 0 means investing in every cooperative agent.

Interestingly, we posit that these mechanisms require different levels of information, which may or may not be readily available in the given network. In some cases, such as social networks, the connectivity (i.e. the number of friends) of a node is virtually free information which requires no effort on the part of the external decision maker to discern. On the other hand, neighbour-hood based approaches inherently require more information about the population and the level of cooperativeness in different parts of the network. Thus, POP is a broad mechanism which only requires knowledge about overall cooperativeness, but NEB invites complex information gathering, in order to determine the cooperativeness in each neighbourhood. Combining NI with NEB does not require any additional observation than NEB by itself. Our study of neighbourhood based interference does not directly take into account the cost of gathering information, it is a comparison between perceived gains in cooperation and the associated per-individual cost of interference set out in the interference mechanisms. Our discussion will naturally present these subtle differences in the hierarchy of information gathering, as they signal hidden costs for some application domains.

3.3 Results

In contrast to the study on square lattice networks [Han et al., 2018], we found that performing cost-effective interventions on SF NOCs exhibits complex patterns and presents multiple concerns. In structured populations, more detailed observations resulted in effective interventions with improved outcomes. On the other hand, more knowledge about the population in SF NOCs simply reduces the risk of interfering to the detriment of cooperators. In other words, interfering in SF NOCs without adequate knowledge should be approached cautiously or it could act to the benefit of defectors. This issue is prevalent in SF networks with low clustering (BA model), but also sees some representation in highly clustered (DMS) networks if stochastic dynamics are taken into account.

Successfully investing in BA populations broadly requires heavy-handed investment and large individual endowments (often orders of magnitude higher than similar mechanisms performed on square lattice populations) or a blanketing mechanism that targets all or almost all cooperators, even those which are not necessarily in danger of converting to defection. Converging to 100% C is very difficult unless both of these conditions are met and this introduces multiple concerns in the role of an exogenous interfering party. We avoid focusing on solutions where the per-generation cost is excessive, as it is unlikely for any institution to be able to produce unrealistically high endowments, as required by these heterogeneous networks. Instead we focus on effective intervention with manageable amounts of per-generation cost. In the following subsections, we structure our results based upon the most important findings, and provide relevant references to each studied investment scheme where appropriate. Initially, we will present the results for the deterministic update scheme, then the stochastic update, pointing out any difference between the two. All the main findings are robust irrespective of the social learning paradigm employed.

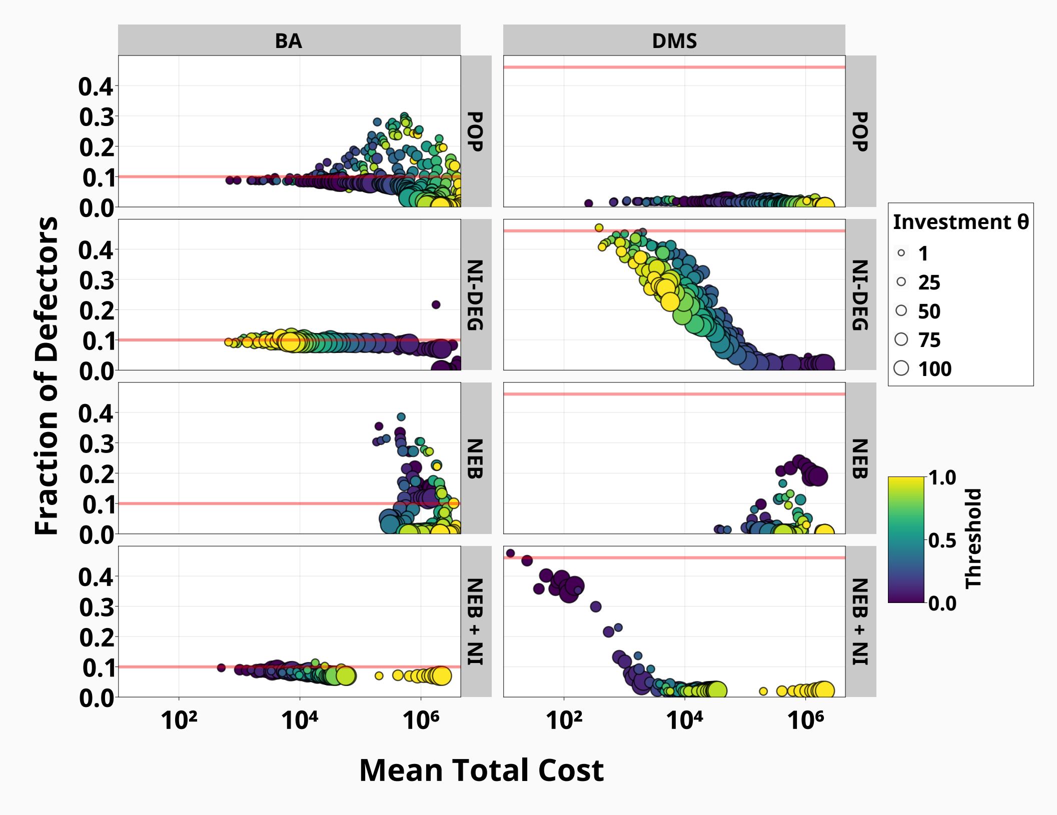

Fig. 3.1 **Fraction of defectors as a function of the mean total cost for each scheme (deterministic update)**. The markers' size is determined by the individual investment θ (grouped to the nearest value), whereas the colour indicates the threshold. Points near the origin indicate the optimal solutions. The horizontal red lines indicate the baseline level of defection in the absence of rewards for either network type (i.e. BA or DMS). Parameters: $b = 1.8$; $N = 5000$.

Careless rewarding leads to the exploitation of cooperators

In direct contrast with previous findings for positive incentives [Chen et al., 2015; Han and Tran-Thanh, 2018; Han et al., 2018], an external decision maker should only interfere in scale-free networks with great care, as investing indiscriminately can lead to the detriment of cooperation (see Figure 3.1). We observe that *inclusive* approaches to interference negatively impact the mean frequency of cooperation if the individual endowments are not sufficient to turn defectors away from the temptation of defecting. By inclusive approaches, we imply high values for the threshold that determines the eligibility of investment (for POP and NEB schemes). If an external investor

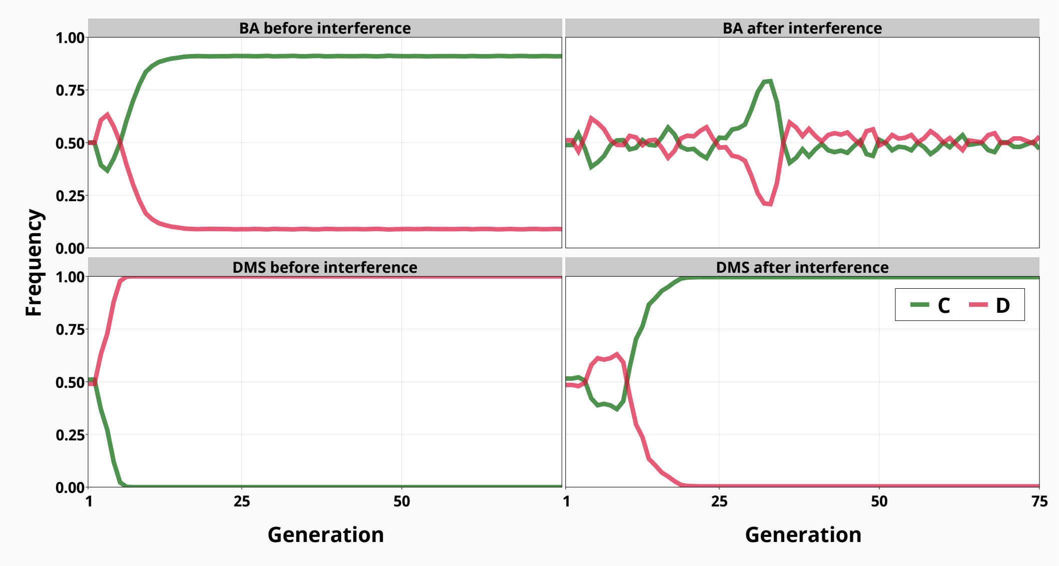

Fig. 3.2 **Typical time-evolution of cooperation,** for $\theta = 5, p_C = 0.8$ (deterministic update). The left column shows the network without interference, while the right one shows the same network after population-based (POP) interference. Some configurations for BA resolve to full C, here we show the scenario in which they do not. Other parameters: $b = 1.8$; $N = 5000$.

hedges their bets, targeting a wide spread of nodes (high threshold) with reduced individual endowments, they risk dooming cooperators. In such a scenario, we see the formation of cyclic patterns, ultimately allowing D players to exploit cooperators (see Figure 3.2). In this way, an investor would be artificially allowing the survival of cooperators in clusters dominated by defectors, abetting the possibility of these sparsely connected clusters to take over larger formations which cannot easily be maintained by defectors. We note that some of these cyclic patterns eventually converge to a stable state, but the accumulated costs of interference at the end of these long-lasting patterns is prohibitively large.

In the presence of deterministic selection, this finding is mostly restricted to classical scale-free networks with low clustering (generated using the BA model), but relaxing the intensity of selection produces similar results even with more realistic levels of clustering (see Figure 3.7). Social diversity changes the inherent nature of the problem of rewarding cooperators effectively. Previous results show the emergence of cooperation in heterogeneous networks [Santos et al., 2006a, 2008] (shown also in horizontal red lines in

Figure 3.1). Compared to homogeneous (well-mixed) and structured populations, there is little improvement to be made in these settings. As the room for improvement narrows, the risk of acting to the detriment of cooperators increases. Individual benefactors prosper temporarily, but the recipients of their naivety are none other than the defectors who exploit them.

Clustering reduces the burden of investment

Real-world networks have been observed to have higher levels of clustering than what normally occurs in typical scale-free networks [Barrat and Pastor-Satorras, 2005; Su et al., 2016]. Nevertheless, several domains, such as the topology of the WWW remain, in which the nodes are sparsely clustered [Albert et al., 1999; Barabasi, 2014; Barabási, 2016]. Thus, it is important to design interference schemes which can target either type of scale-free networks, especially so if there exists a degree of uncertainty about the presence of clusters, or if measuring this factor is unfeasible. We have already mentioned the risks associated with inadequate reward mechanisms, but now we can turn to unveiling the benefits associated with social diversity and clustering in the quest towards engineering pro-social behaviour.

Highly clustered networks often have the most room to improve by receiving endowments (See Figure 3.1). The initial distribution of players in the hubs of the network often determines whether the direction towards which the population will converge. Often, a small nudge can steer the population towards a desirable outcome (see Figure 3.2). Moreover, this can easily be accomplished through a variety of disparate investment paradigms. For instance, metrics on the overall population (POP) can be used to guarantee maximal cooperation regardless of how the endowments are distributed (See Figure 3.3). With the reduction in the complexity of designing an effective scheme, we look towards cost and ways to reduce overspending. Overeager endowments can lead to total costs several orders of magnitude larger than those applied as a last resort. Indeed, even very small endowments applied to few surviving cooperators can *jumpstart* the formation of clusters resilient to invasion. Increasing the threshold for investment guarantees that more cooperators will be eligible for the endowments, thus exacerbating spending. Lowering this threshold guarantees that interference will only be triggered

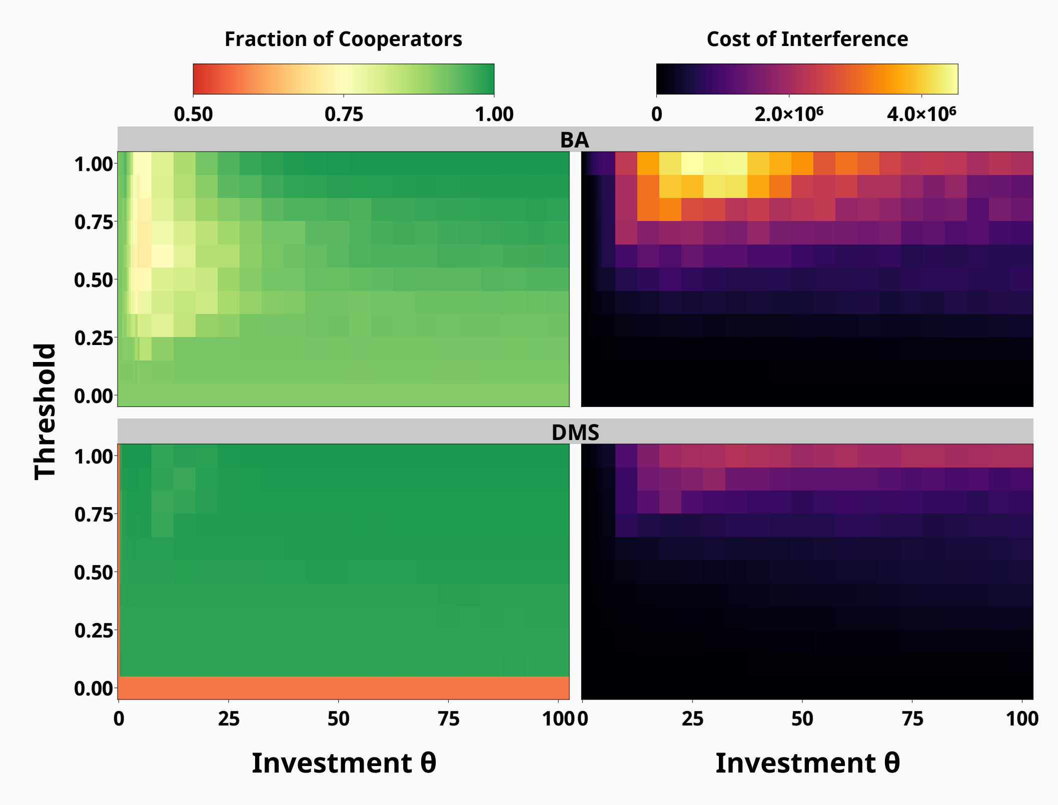

Fig. 3.3 **Fraction of cooperation and total cost for population-based (POP) interference,** using deterministic update. Parameters: $b = 1.8$; $N = 5000$.

if desperately required. Investing in every cooperator as a last resort ensures pro-sociality.

Moreover, local observations can be used to ensure positive outcomes following a variety of pathways (see Figure 3.4). In this case, an external decision maker must target a range of intermediate values for the threshold. Previous results on structured populations showed that investing in cooperator neighbourhoods with exactly one defector was the optimal way of fostering cooperation [Han et al., 2018]. In contrast, our findings suggest that the opposite is true for heterogeneous settings. Indeed, the least expensive routes towards cooperation are those with low or intermediate thresholds, suggesting that investors should focus their attention on ensuring only the survival of cooperators who are in danger of turning. For highly clustered networks, little investment is needed, and provided the threshold is not exceedingly low, maximal cooperation can be reached in any configuration,

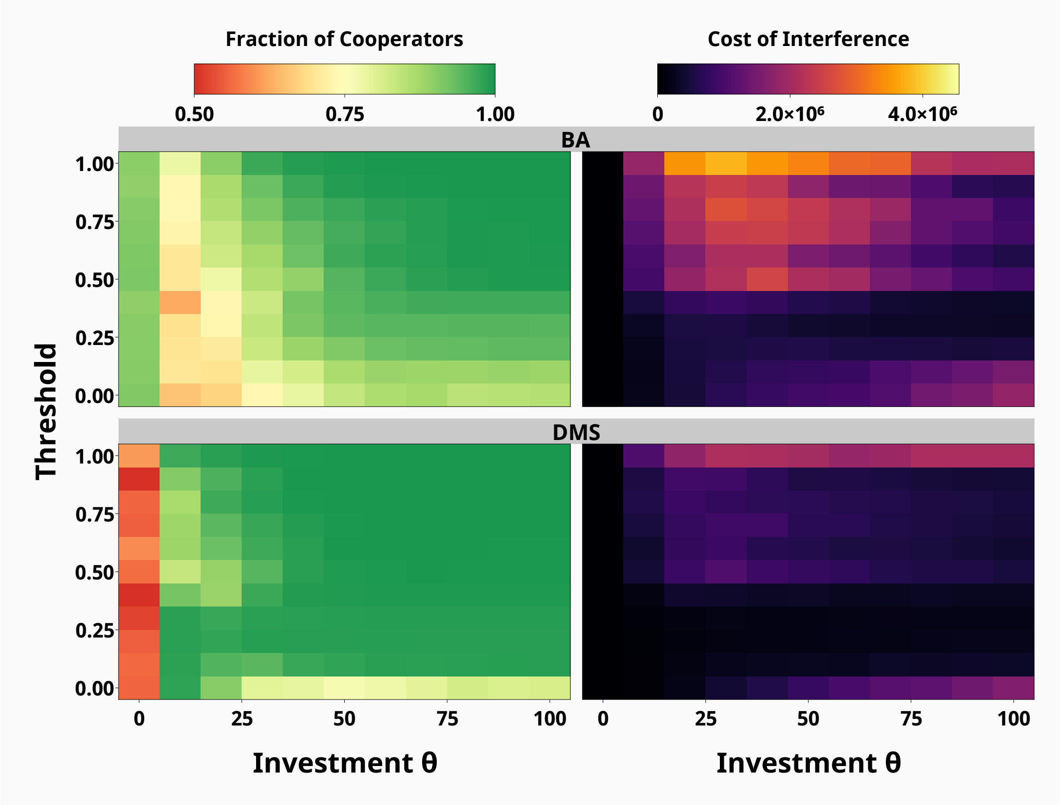

Fig. 3.4 **Fraction of cooperation and total cost for local neighbourhood information (NEB) interference,** using deterministic update. Parameters: $b = 1.8$; $N = 5000$.

without unnecessary expenditure. Lowly clustered networks, on the other hand, require much more deliberate endowments to benefit from investment, with the added risk of causing cooperators to fall victim to exploitation as discussed previously in Figure 3.4.

Heterogeneity and network characteristics play a key role in the design of effective investment mechanisms

Assuming that information about a node's influence can be easily gleamed by an external decision maker, this can provide a partial solution to reducing the risk of deleterious interference. Although comparatively costly, this mechanism has the benefit of never succumbing to the exploitation of cooperators (see Figure 3.5). Notwithstanding, the very nature of influential nodes in scale-free networks (i.e. power-law degree distribution; see Section 2.4.3) implies only exceedingly large endowments are sufficient to sway

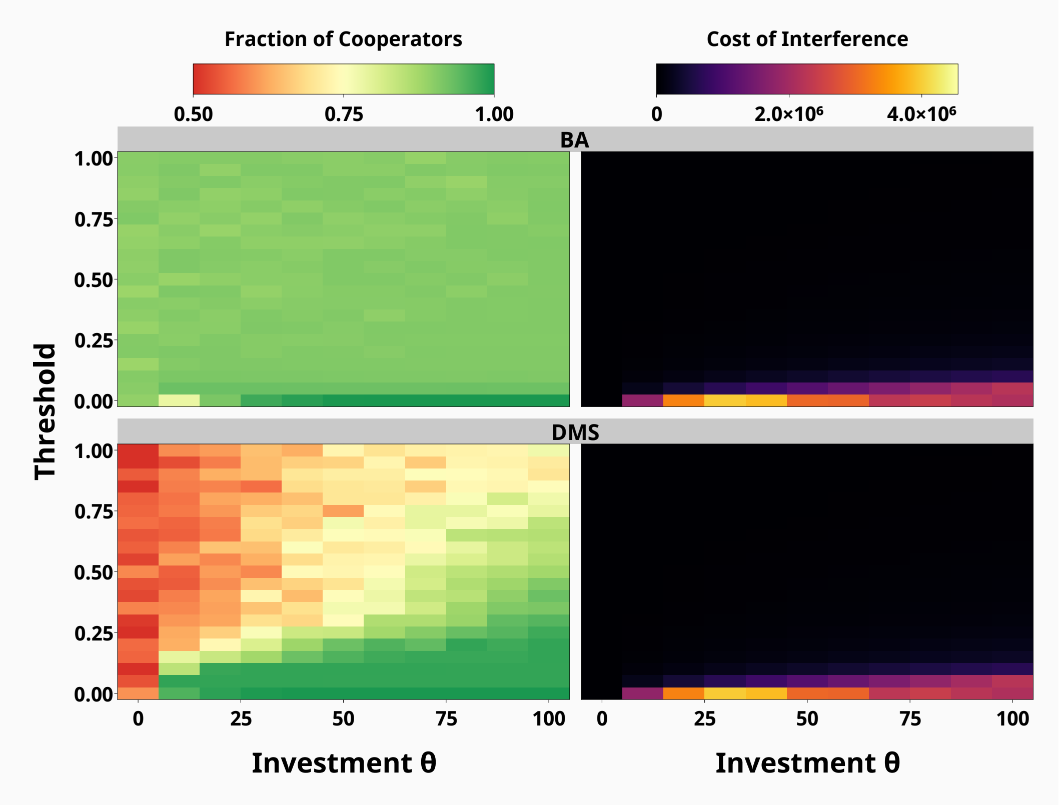

Fig. 3.5 **Fraction of cooperation and total cost for node influence-based (NI) interference,** using deterministic update. Parameters: $b = 1.8$; $N = 5000$.

them. However, the number of cooperators who are eligible for investment is also small; on account of this, overall spending does not scale predictably with the endowment amount. We have previously mentioned that there exist some costs associated with information gathering, which we do not model or measure here. Hence, the assumption that information about influence is readily available suggests this method could prevail in respect to real-world budgeting.

We propose that combining several interference mechanisms can be an effective way of reducing spending while avoiding the pitfalls of pernicious investment. For instance, we might consider taking into account an agent's influence as well as local observations. In Figure 3.6, we explore this possibility, avoiding the least connected nodes (i.e. not investing in the bottom 5% of nodes in respect to degree centrality), and show that this reduces spending compared to either of the two interference schemes taken individually. These

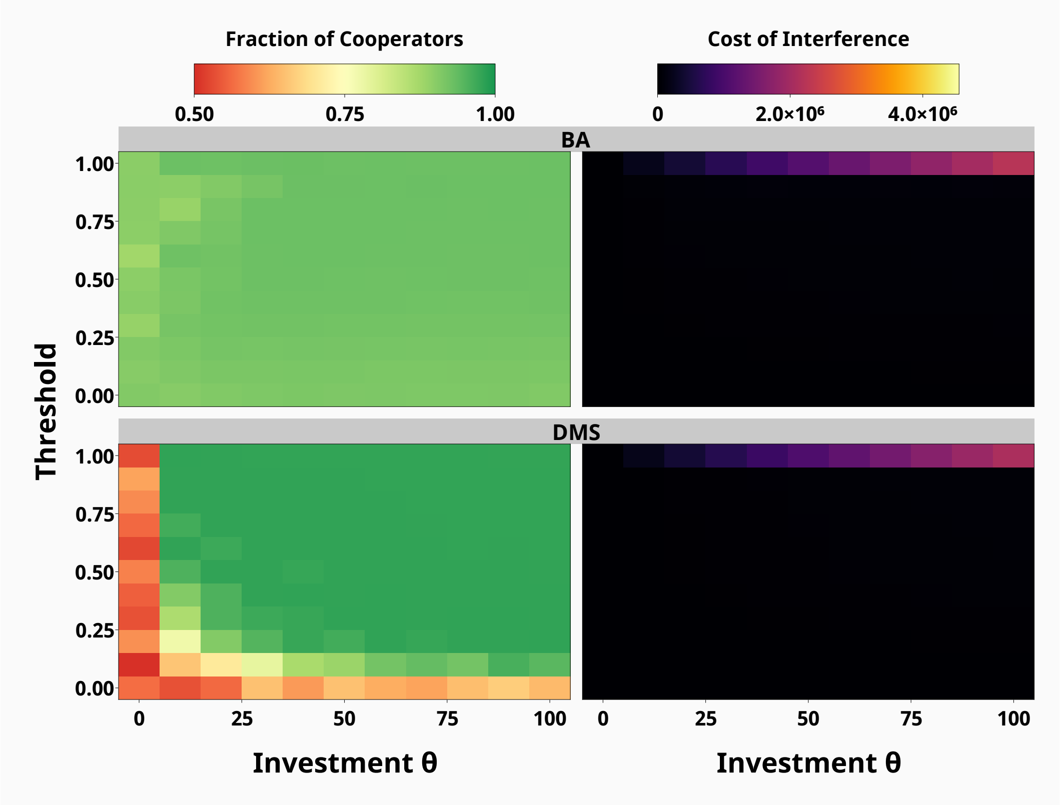

Fig. 3.6 **Fraction of cooperation and total cost for a mixed interference scheme (NEB and NI),** using deterministic update. We fix $c_I = 0.05$, avoiding investing into the least connected nodes (bottom 5%). Parameters: $b = 1.8$; $N = 5000$.

results suggest that hubs play an important role in the emergence of cooperation in highly clustered networks, but that they cannot be effectively used to improve outcomes in their lowly clustered counterparts. Nevertheless, this integrated approach to interference eliminates the possibility of investment being detrimental to cooperation.

We note this conundrum between the two types of heterogeneous networks. Lowly clustered networks have little to benefit from investment, and much to lose if the external investor is negligent in their distribution of endowments. On the other hand, highly clustered networks have much to gain and little to lose, readily responding positively to any tactic, overspending being the only matter of discontent. As investment in the greater context of heterogeneous interactions is not trivial, it would therefore be prudent to first collect as much data on the nature of the network before deciding to dis-

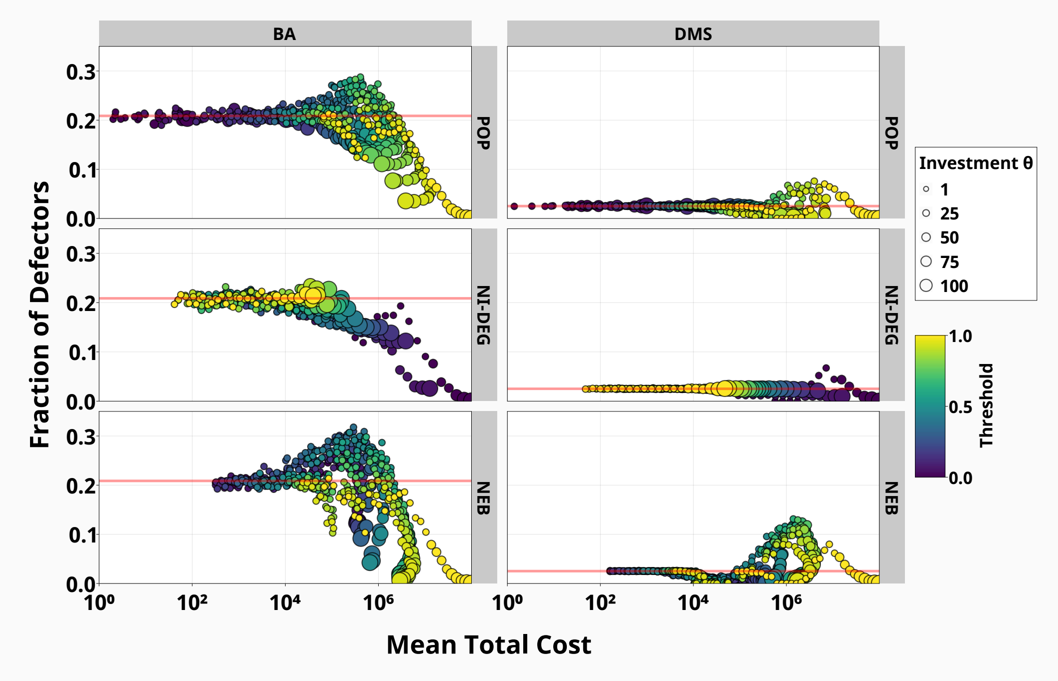

Fig. 3.7 **Proportion of defectors as a function of the mean total cost for each scheme (stochastic update).** The markers' size is determined by the individual investment θ (grouped to the nearest value), whereas the colour indicates the threshold. Points near the origin indicate the optimal solutions. The horizontal red lines indicate the baseline level of defection in the absence of rewards for either network type. Parameters: $b = 1.8$; $N = 2000$; $k = 0.1$.

tribute endowments. Uncertainty about social diversity or clustering carries the additional risk of selecting an improper policy of designing incentive schemes.

Stochastic imitation increases the risk of exploitation

Previously, we had shown that careless rewards might lead to an increase in defectors when interfering in BA networks under a deterministic update paradigm (see Subsection 3.3). Following a transition towards a more realistic, stochastic update rule [Traulsen et al., 2006], we observe a very similar phenomenon and moreover, find that it is no longer limited to lowly clustered scale-free networks (see Figure 3.7). Indeed, investing in DMS networks should be approached with the same due diligence as BA networks, and insufficient endowments often lead to the exploitation of cooperators.

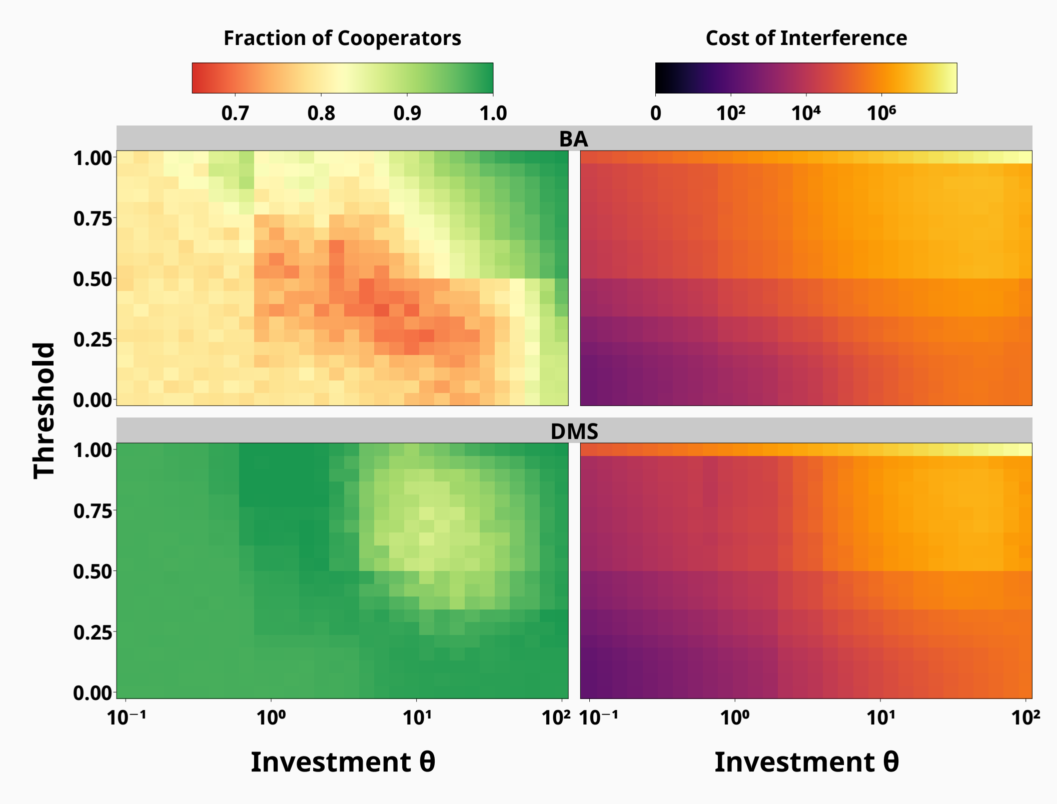

Fig. 3.8 **Fraction of cooperation and total cost for local neighbourhood information (NEB) interference,** using stochastic update. Parameters: $b = 1.8$; $N = 2000$; $k = 0.1$.

Relying solely on local information is most prone to damaging cooperation, in spite of the level of complexity associated with this scheme and the amount of information required to enforce it (See Figure 3.8).

Interestingly, stochastic imitation leads to a significant increase in baseline cooperation prior to interference in highly clustered networks, and conversely a decrease in cooperation in classical scale-free networks (see horizontal lines in Figure 3.7). Nevertheless, the findings discussed in the sections above remain robust. Although the potential gains to be had shift, causing BA networks to benefit from investment more than their highly clustered counterparts, the previous findings still apply in this setting. For instance, we find that DMS networks readily respond to investment, and are not as prone to the pitfalls which befall BA networks. In Figure 3.9, we show that the most efficient interference schemes are consistently one (or

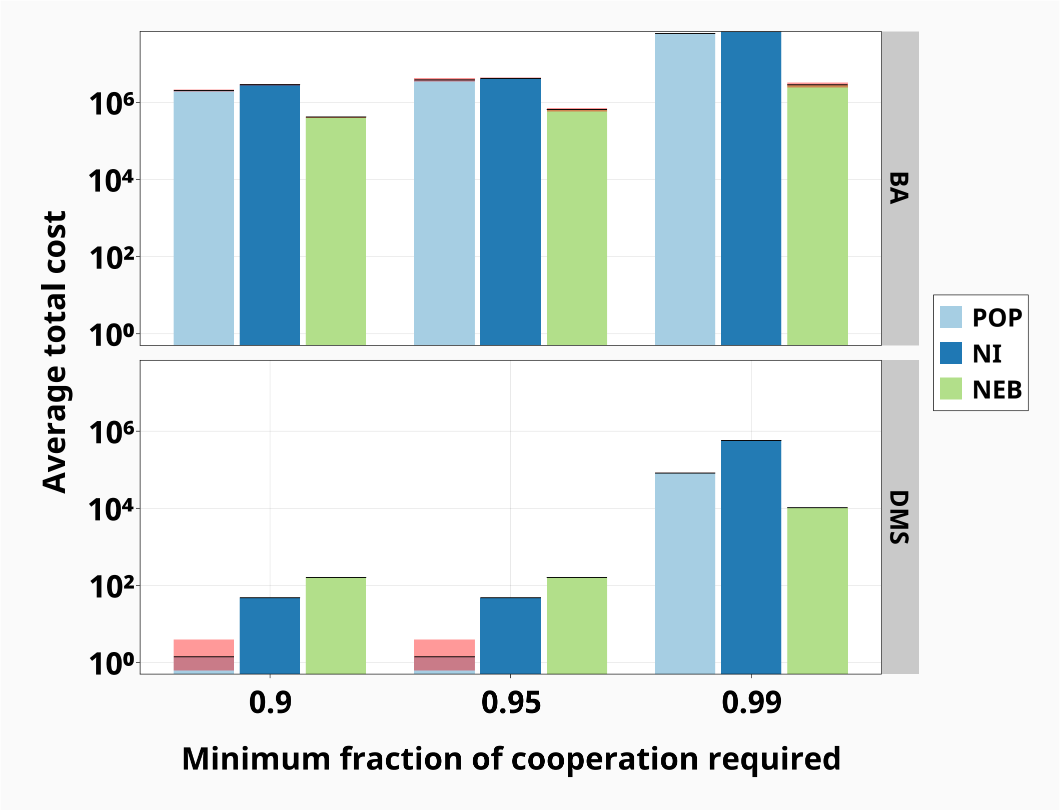

Fig. 3.9 **Mean total costs for the most efficient combinations of threshold and investment amount** θ, using stochastic update. We intentionally avoid configurations in which no endowments are distributed, and select the configurations with the least possible cost for each minimal fraction of cooperation required. Error bars in light red show the standard deviation across all replicates of a configuration.

more) order(s) of magnitude less costly at promoting cooperation in highly clustered networks, regardless of what potential gains the external decision makers is aiming for.

Maximal cooperation gains require significant endowments. Cost-efficiency is a double-edged sword.

Unlike homogeneous populations, heterogeneous interaction structures inherently provide a benefit to cooperators, something usually referred to as network reciprocity [Santos et al., 2006a, 2008]. In practice, this means investment does not lead to outcomes which differ significantly from the baseline. Furthermore, successful attempts at reaching a maximal level of

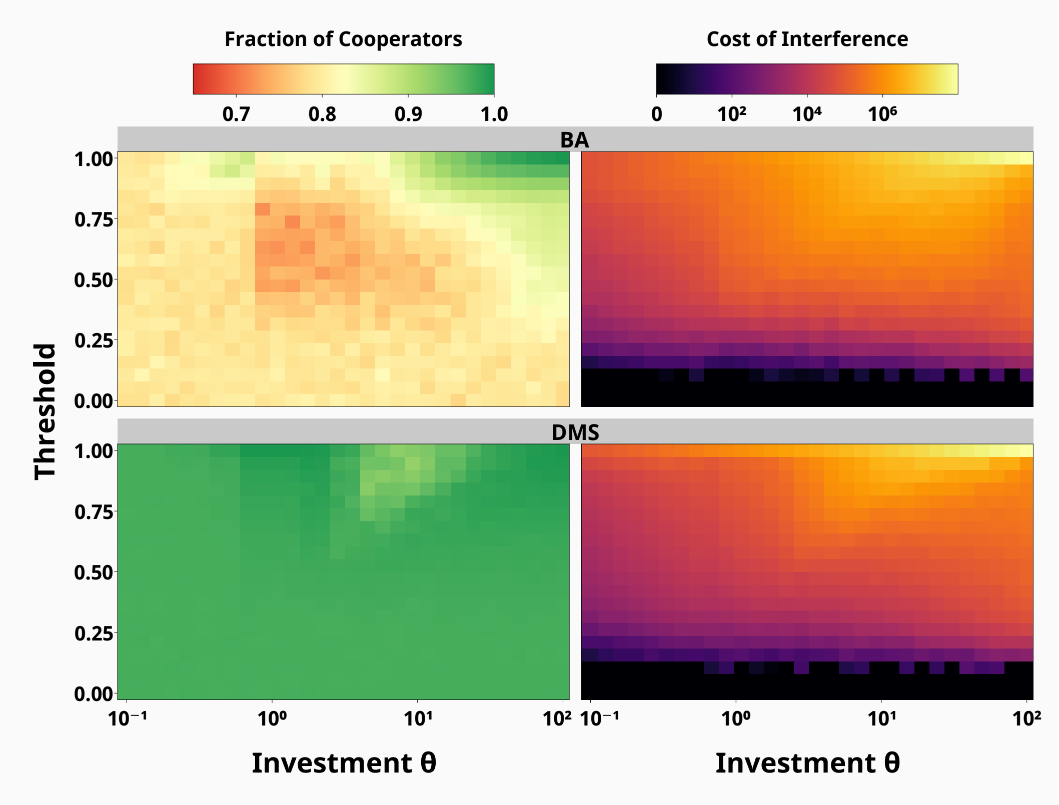

Fig. 3.10 **Fraction of cooperation and total cost for population-based (POP) interference,** using stochastic update. Parameters: $b = 1.8$; $N = 2000$; $k = 0.1$.

cooperation (i.e. little to no defection) require a combination of large endowments and an investment scheme which can target individuals at all levels of the network (see Figure 3.7). Using population level metrics generally fails to improve outcomes unless virtually every cooperator is targeted (see Figure 3.10). Equivalently, relying on degree centrality (i.e. how influential a node is) necessitates an egalitarian distribution of endowments, which naturally increases costs (See Figure 3.11).

Local information (NEB), while risky, also has the potential to best improve outcomes while reducing costs, and this remains true for both BA and DMS networks, if maximal cooperation gains are required (See Figure 3.9). Once again, we intuit the importance of acquiring detailed observations of information about the agents. This approach is a double-edged sword; it is simultaneously the optimal solution, as well as the most prone to errors in decision-making, leading the population to either the most perceived gains

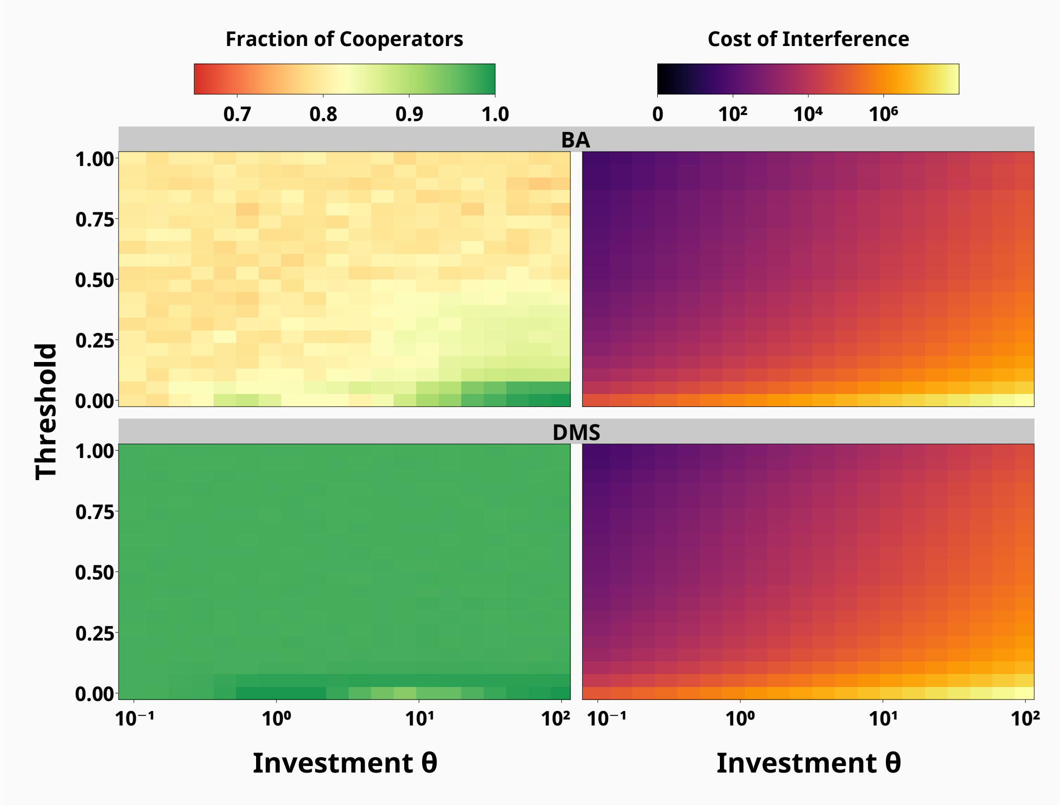

Fig. 3.11 **Fraction of cooperation and total cost for node influence-based (NI) interference,** using stochastic update. Parameters: $b = 1.8$; $N = 2000$; $k = 0.1$.

or the least (See Figure 3.8). This seems to be another dilemma. Investing is risky, and it is likely for endowments to be ineffective or even produce negative results, but only significant sums of capital are likely to lead to desirable outcomes. Social diversity complicates this further, as there exists a great degree of inequality between individuals, and potential errors in decision-making make investment precarious.

3.4 Discussion

In summary, this chapter aims to determine how best an external decision maker could incentivise a population of autonomous agents facing a cooperative dilemma to fulfil a coveted collective state. We build on a previous account which identified the most effective mechanisms to foster cooperative scenarios in spatially distributed systems in regular graph structured populations of agents, but instead we consider two popular models of scale-free networks of contacts. In particular, we try to understand if the insights set out in the context of regular graphs remain applicable to heterogeneous models, as well as explore an additional avenue of interference enabled by the variance in node connectivity. To address these issues, we combine an evolutionary game theoretic model with several incentive mechanisms in two types of pre-generated networks characterised by preferential attachment, with different clustering coefficients. We argue that this problem cannot be solved trivially and we show that transitivity (i.e. the global clustering coefficient) should be the driving force behind the choice of an interference mechanism in promoting cooperation in heterogeneous network structures, as well as its application.

In this chapter, we introduce several incentive mechanisms which are defined formally and mathematically. We note that they do not have to be defined as such, and in fact have many real-life counterparts which are often employed by institutions and investors. For instance, POP-based metrics describe cooperation observed at a global scale. If we consider the Great Recession, or the recent COVID-19 pandemic, an institution might only need to look at the overall state of the economy, or the spread of an infection, before deciding that action is required. Neighbourhood-based metrics represent local schemes, which are almost ubiquitous when considering social inequality. Whether it is housing schemes, incentives to stop smoking, homelessness, education, etc., local governments often decide to invest based on the level of economic and social deprivation in a specific area, and that is precisely what we have tried to capture with NEB-based schemes. Finally, we have looked at centrality (influence) metrics. If we consider social media, a company might wish to use influencers to market its products, or an institution might decide to specifically target someone in the public eye in order to increase the visibility of its incentives, whether they were positive or negative. The mechanisms we have chosen are by no means exhaustive choices, but they serve as a fundamental starting point to our discussion, and they are arguably the most common and most easily implementable mechanisms that we observe in the real world.

We find that impetuously rewarding cooperators can lead to cyclic patterns which damage cooperation in the long run, enabling the exploitation of cooperators to the benefit of defectors. We argue that detailed information gathering about the networks and agents prior to the distribution of endowments can prevent these mistakes. Using two social learning paradigms, we show the robustness of these findings and observe that clustering lowers the risk of deleterious investment, easing the strictness of distributing incentives. Moreover, we show that ignoring lowly connected individuals leads to unprofitable and even futile intervention irrespective of network transitivity.

Our comparison between the two types of scale-free networks provides valuable insights regarding the importance of clustering in the outcome of cooperation. We find that a large clustering coefficient allows for successful, cost-effective interference, indeed even when partly disregarding a full comprehension of the population and its tendencies. Furthermore, transitivity lessens the burden on external investors, lowering the total cost required to enforce cooperation. These results are of particular interest, given that most SF networks portray high clustering, such as in the case of social ties where friends are likely to be friends of each other [Newman, 2018]. This scope encompasses heterogeneous scenarios inhibited by spatial constraints (e.g. in highly urbanised areas or even the allotment of rangelands such as pastures), where high clustering is also imposed.

In this work, we do not consider the possibility of detecting the existence of a certain type of interference from an external party. In reality, individuals could be aware of active interference and react by changing their behaviour, either to become suitable candidates for reward or to avoid sanctions. In Chapter 5, we introduce the idea of the threat of punishment, but in future work we could also test whether evolutionary dynamics might lead to the capacity for individuals to detect external interference.

Transitioning towards a more realistic, stochastic imitation rule [Traulsen et al., 2006], we measure a shift between the two network types, whereby lowly clustered networks prescribe a greater need for investment, and vice-versa. Maximal cooperation gains in either paradigm can generally be achieved using large individual endowments. Notwithstanding, highly clustered networks respond more readily to interference, and we provide several insights about ways in which cost could be reduced further.

4 | Promoting Fair Proposers, Responders or Both?

The lowly have small ambitions, and are satisfied with small indulgences. They need not get fair treatment. They need only think that they do...

—Joe Abercrombie, *The Last Argument of Kings*

In which we foster fairness in structured populations. Here[1], we show the importance of strictly targeting individuals who are fair in both their offers and responses. Additionally, we find that these measures can be relaxed through increased information gathering, or in the presence of social diversity. Crucially, role asymmetry and social diversity open up novel mechanisms available to institutions wishing to promote pro-social outcomes.

4.1 Introduction

Fairness has a deep impact on decision-making and individuals often prefer fair outcomes over payoff maximising ones [Nowak et al., 2000; Rand et al., 2013]. For example, in group interactions, fairness concerns emerge when agents must decide upon outcomes possibly favouring different parts unequally [Teixeira et al., 2021]. This is true for many domains, such as automated bargaining [Jennings et al., 2001], conflict resolution [Pritchett and Genton, 2017] or multiplayer resource allocation [Chevaleyre et al., 2006]. Moreover, with the advent of autonomous technology, it is crucial to determine how best to engineer pro-social behaviour in a hybrid society of humans and machines [Paiva et al., 2018]. While several mechanisms have been identified to explain why fairness is widespread in human decision-making, the introduction of machines and artificial agents in society could result in vastly different responses. Engineering fairness in such a context might hinge on exogenous agents or institutions able to engage in the distribution of incentives. In these scenarios, external decision makers need to find a trade-off between the cost of the investment and its effectiveness in ensuring high levels of fair behaviour. In this chapter, we provide novel insights towards robust solutions for the grand challenge of engineering pro-sociality in dynamical multi-agent systems.

The literature on external interference in evolving, dynamical systems (or populations) has so far focused on cooperation dilemmas, namely the Prisoner's Dilemma (PD) [Cimpeanu et al., 2019; Han and Tran-Thanh, 2018; Han et al., 2018] and the Public Goods Game (PGG) [Chen and Perc, 2014;

[1] The model and results presented in this chapter are also reported in Cimpeanu et al. [2021a,b].

Chen et al., 2015; Sasaki et al., 2012; Wang et al., 2019]. In these games, the interactions are symmetric and the players' roles are equivalent. However, many real-world and MAS interactions are asymmetric, where players may have different baseline characteristics and/or play different roles in the interactions [McAvoy and Hauert, 2015; Ogbo et al., 2021; Tuyls et al., 2018]. Examples include conflict resolution [Selten, 1980; Smidt, 2020], technology adoption by firms [Ogbo et al., 2021], and multiparty resource allocations [Chevaleyre et al., 2006; Lerat et al., 2013], where participants might have different roles (e.g. proposers/dictators vs responders) or bargaining power in the decision making process. In this asymmetric setting, the external decision maker might need to take into account the difference among players' underlying characteristics, such as their roles in the interactions, in order to optimise the cost and the level of desired behaviour. In particular, we might ask, is it enough to target a subset of the roles to already achieve a sufficiently good outcome, since collecting information about all the roles might be (very) costly and time consuming?

Optimising fairness becomes especially challenging when analysing dynamical systems that incorporate diverse stochastic effects and uncertainty factors, such as a non-deterministic behaviour update. For evolving dynamical systems such as those in the above-mentioned examples, system dynamics are shaped by various stochastic and random effects, such as those resulting from behavioural updates and mutation (behavioural exploration) [Rand et al., 2013; Traulsen et al., 2009]. With behavioural updates, such as through social learning or reproduction [Nowak, 2006a; Sigmund, 2010], undesired behaviours might resurface over time whenever interference was not sufficiently strong in the past. Through mutation, these behaviours might do so even when they were extinct. Hence, the external decision maker needs to take into consideration that they will have to repeatedly interfere in the system, in order to sustain the desired behaviour over time. Note however that, for simplicity, previous works have either omitted mutation [Han and Tran-Thanh, 2018; Wang et al., 2019], or assumed that it is infinitely small (for analytical treatment) [Duong and Han, 2021b; Han and Tran-Thanh, 2018]. Mutation (behavioural exploration), where agents can freely experiment with new behaviours, is usually non-negligible in real populations and has been shown to play an important role in enabling cooperation in the context of social dilemmas [Antal et al., 2009; Duong and Han, 2019, 2021a; Han et al., 2012; Rand et al., 2013; Traulsen et al., 2009]. Thus, this chapter will also advance the state-of-the-art in this respect, where we will closely examine how different regimes of mutation, or agents' propensity for behavioural exploration, influence the manner in which external interference should be carried out.

The aim of this chapter is to contribute to the timely challenge of pro-social computing [Paiva et al., 2018], providing robust solutions towards optimising the cost of engineering fairness in a real-world multi-agent setting, in the presence of social diversity and incomplete information. We resort to the Ultimatum Game [Nowak et al., 2000] as a suitable mathematical approach to modelling fair decision making. In the Ultimatum Game, one of the players can decide on how to split a sum of money. Thus, in this setting, offers close to an even split are considered fair. In an uneven split, in which the proposer gets to keep most of the money, is considered unfair. Because the proposer has asymmetric power in the interaction, the only way in which they can be "punished" is if the responder declines an unfair offer, thus causing neither individual to receive anything from the original sum. Supremely fair individuals would always propose an even split, and always decline unfair offers to prevent selfish players from receiving part of the endowment. In contrast, very unfair individuals would propose to keep most if not all of the endowment while accepting anything they are given. To avoid the interesting discussion of what is a fair response (to keep anything that is offered or to "punish" the other by declining), we have chosen to measure fairness by the number of fair offers in the population (see e.g. [Nowak et al., 2000; Page et al., 2000; Rand et al., 2013]), disregarding the responses in this metric. We determine how these heterogeneous network characteristics can be exploited to reduce costs while maintaining high standards of fairness, providing insights regarding how the presence of social diversity alters the complexity of engineering fairness and how it can be done efficiently.

4.2 Models and Methods

4.2.1 Ultimatum Game (UG)

Agents' interaction is modelled using the one-shot Ultimatum Game (UG) [Nowak et al., 2000; Page et al., 2000]. In the UG, two players are offered a chance to win a certain sum of money, normalised to 1, which they must divide between each other. One player is elected proposer, and suggests how to split the sum, while the other, the receiver (responder) can accept or reject the deal. If the deal is rejected, neither player receives any part of the initial sum. As in [Nowak et al., 2000; Page et al., 2000], we assume that a player is equally likely to perform in one of the roles (proposer or receiver). A player's strategy is defined by a pair of parameters, p and q. When acting as proposer, the player offers the amount p, whereas in a receiver's role, the player rejects any offer smaller than q.

As we focus in this chapter on the effect of having multiple roles on interference decision making, we consider a baseline UG model where proposers have two possible strategic offers, a low (L, with $p = l$) and a high one (fair) (H, with $p = h$), where $l < h \in [0,1]$. On the other hand, receivers have two options, a low threshold (L, with $q = l$) and a high threshold (H, with $q = h$). Thus, overall, there are four possible strategies HH, HL, LH and LL (i.e. HL would denote proposing high and accepting any offers, etc.). Fairness is measured by calculating what percentage of the population is representative for either the HH or HL strategies (i.e., fair proposers), and this allows us to have a clear comparison with previous works—in terms of the level of population fairness achieved—that have studied the evolution of fairness in the UG, see e.g. [Nowak et al., 2000; Page et al., 2000; Rand et al., 2013]. Unlike our work, they did not study the cost-efficiency of interference strategies for enhancing fairness.

Given evidence from several behavioural experiments [Güth et al., 1982; Rand et al., 2013], in which people (almost) never offered more than half of the sum in UG, we assume $h \leq 0.5$. Particularly, we set $h = 0.5$ and $l = 0.1$ in homogeneous populations, as shown in [Page et al., 2000]. In this scenario, the strategy LL is the most frequent strategy in the population. We also

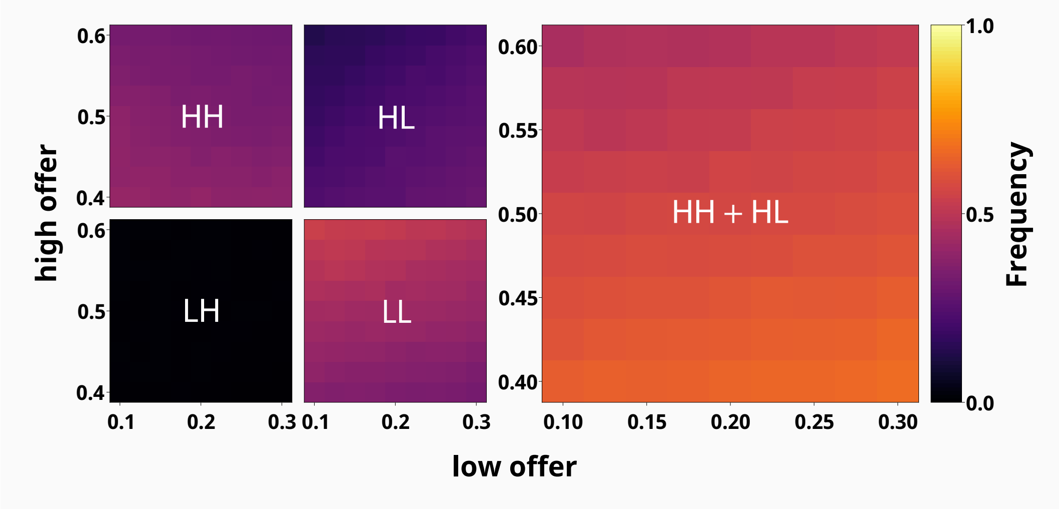

Fig. 4.1 **Baseline frequencies for each strategy in scale-free networks (BA)**, with a separate panel for overall fairness (fair offers). $\mu = 0$.

confirm this result in our simulations, as shown in Figure A.1 in Appendix A and we note that this result is true for several mutation rates. For scale-free, heterogeneous populations, we set $h = 0.6$ and $l = 0.1$, as this represents the environment with (roughly) the lowest frequency of fair proposals (see Figure 4.1). We intentionally choose a more competitive UG environment for heterogeneous populations, given the increase in fairness observed in the absence of investment.

As we focus in this chapter on the effect of having multiple roles on interference decision making, we consider a baseline UG model where proposers have two possible strategic offers, a low (L, with $p = l$) and a high one (fair) (H, with $p = h$), where $l < h$, with $l, h \in [0,1]$. On the other hand, receivers have two options, a low threshold (L, with $q = l$) and a high threshold (H, with $q = h$). Thus, overall, there are four possible strategies HH, HL, LH and LL (e.g., HL denotes a strategy that offers high and accepting any offers). The payoff for the four strategies HH, HL, LH and LL reads (for row player):

For example, an HH player encountering an HL player results in the payoff $\frac{1}{2}$ for either player, as both of them propose and accept a fair split (i.e. one interaction results in the payoff $1 - h$ for the proposer, and h for the receiver, and vice-versa for when the roles are reversed).

	HH	HL	LH	LL
HH	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1-h}{2}$	$\frac{1-h}{2}$
HL	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1-h+l}{2}$	$\frac{1-h+l}{2}$
LH	$\frac{h}{2}$	$\frac{1+h-l}{2}$	0	$\frac{1-l}{2}$
LL	$\frac{h}{2}$	$\frac{1+h-l}{2}$	$\frac{l}{2}$	$\frac{1}{2}$

4.2.2 Population structure and dynamics

We consider a population of agents or individuals distributed in a network of contacts. Among these, we study well-mixed (WM) populations, square lattices (SL), as well as two types of scale-free (SF) networks, the Barabási and Albert (BA) [Barabási and Albert, 1999] and the Dorogovtsev-Mendes-Samukhin (DMS) [Dorogovtsev et al., 2000] models (for a detailed description please see Section 2.4 in Chapter 2). We focus our analysis on the efficiency of various interference strategies in spatial settings, adopting an agent-based model directly comparable with the setup of recent lab experiments on cooperation [Rand et al., 2014].

Initially each agent is designated as one of the four strategies (i.e. HH, HL, LH, HH), with equal probability. At each time step or generation, each agent plays the UG with its immediate neighbours. In the well-mixed baseline, each agent plays the UG with every other agent in the population. The score for each agent is the sum of the payoffs in these encounters. At the end of each generation an agent A with score f_A chooses to copy the strategy of a randomly selected neighbouring agent B with score f_B with a probability given by the Fermi function (i.e. *stochastic update*) [Traulsen et al., 2006]:

$$(1 + e^{(f_A - f_B)/K})^{-1},$$

where K denotes the amplitude of noise in the imitation process [Szabó and Fáth, 2007]. Varying K allows us to capture a wide range of update rules and levels of stochasticity, including those used by humans, as measured in lab experiments [Rand et al., 2013; Zisis et al., 2015]. In line with previous works and lab experiments [Rand et al., 2013; Szabó and Fáth, 2007; Zisis et al., 2015], we set $K = 0.1$ in our simulations. With a given probability μ, this process is replaced instead by a randomly occurring mutation. A mutation is equivalent to behavioural exploration, where the individual makes a stochastic decision to switch to one of the four available strategies.

Although our analysis below will focus on the stochastic update rule (in order to examine how stochasticity affects interference, as discussed above), we will also provide results for a *deterministic update* paradigm to have a clear comparison with previous works (see e.g. [Han et al., 2018]). For the deterministic update, an agent's strategy is always changed to that of its highest scoring neighbour [Nowak and May, 1992; Szabó and Fáth, 2007]. This is a way of approximating the stochastic update rule where the stochastic effect is infinitely small, i.e. $K \to 0$.

We simulate this evolutionary process until a stationary state or a cyclic pattern is reached. Similarly to [Nowak and May, 1992], all the simulations in this work (described in next sections) converge quickly to such a state. For the sake of a clear and fair comparison, all simulations are run for 500 generations. Moreover, for each simulation, the results are averaged over the final 25 generations, in order to account for the fluctuations characteristic of these stable states. When shown in figures, the error bars represent the standard error of the mean between replicates. Below, we outline the chosen parameters for the experiments performed on each network type:

i) Homogeneous populations

For baseline results performed on well-mixed populations (complete graph), we chose a population size $N = 100$. We set $L = 100$ for our experiments on lattices, resulting in a population size $N = 10^4$. Furthermore, to improve accuracy, for each set of parameter values, the final results are obtained from averaging 30 independent realisations for WM and SL. As the baseline experiments on well-mixed networks converge readily, we run the simulations for 100 generations and average these results over the final 10. For each interference strategy, we study four different mutation rates, for $\mu \in \{10^{-4}, 10^{-3}, 10^{-2}, 2 * 10^{-1}\}$, as well as $\mu = 0$ for the deterministic update. We will explicitly state the values of mutation rates in all figures' captions for homogeneous populations.

Note that in the special case of deterministic update (where we also do not consider mutations), simulations can stop early when the proportion of fair proposers reaches 100%. We note that when maximum fairness is not reached, investment can still be ongoing beyond 500 generations and thus, the total cost of interference is dependent on the chosen stopping point. However, our results show that the average investment at the 500 generation mark is never more than 0.2% of the average total investment, for all types of interference. Thus, this arbitrary number has a limited effect and should not affect these results qualitatively.

ii) Heterogeneous populations

For all of our experiments on scale-free networks, we seed 10 different networks (of each type, BA and DMS) of size $N = 2000$ for robustness, with an average connectivity of $z = 4$, to easily compare against SL graphs. Furthermore, to improve accuracy, for each set of parameter values, the final results are obtained from averaging 20 independent realisations for SF networks, accounting for the additional network seeding and intensive computational requirements. For heterogeneous populations, we will only consider a stochastic update rule with mutation-rate $\mu = 0$ as we will have already thoroughly explained the role of varying mutation rate in the homogeneous case.

Given the differing levels of heterogeneity, the discrepancy between the average connectivity z of WM populations and the other network types, and the variance in the maximal number of generations, total costs cannot be compared across all the networks studied. Indeed, we are systematically exploring the cost-efficiency of certain schemes given an existing population and the emergent fairness dynamics, and assume that the underlying network of interactions cannot be influenced by the external investor.

4.2.3 Cost-Efficient Interference in Networks

We aim to study how one can efficiently interfere in well-mixed, structured and spatially heterogeneous populations to achieve high levels of fairness while minimising the cost of interference. As mentioned above, the level of fairness is measured by the fraction of fair offers in the population [Rand et al., 2013]. An investment decision consists of a cost $\theta > 0$ to the external decision-making agent/investor, and this value θ is added as surplus to the payoff of each suitable candidate. In order to determine cost-efficiency, we vary θ for each proposed interference strategy, measuring the total accumulated costs to the investor. Thus, the most efficient interference schemes will be the ones with the lowest relative total cost.

We examine different approaches to interference, where fairness is advocated for either role or both, leading to different desirable behaviours to be targeted:

(i) ensure all proposals are fair, thus investing in HH and HL (**Target: HH, HL**);

(ii) ensure only fair offers are accepted, thus investing in HH and LH (**Target: HH, LH**);

(iii) ensure both (i) and (ii), i.e. investing in HH only (**Target: HH**).

Moreover, in line with previous works on network interference [Chen et al., 2015; Cimpeanu et al., 2019; Han and Tran-Thanh, 2018; Han et al., 2018], we compare global interference strategies where investments are triggered based on network-wide information, local neighbourhood information, and, lastly, node centrality information.

In the *population-based* (**POP**) approach, a decision to invest in desirable behaviours is based on the current composition of the population. We denote x_f the fraction of individuals in the population with a desirable behaviour, given a targeting approach, i.e. (i), (ii) or (iii) as defined above. Namely, an investment is made if x_f is at most equal to a threshold p_f (i.e. $x_f \leq p_f$), for $0 \leq p_f \leq 1$. They do not invest otherwise (i.e. $x_f > p_f$). The value p_f describes how rare the desirable behaviours should be to trigger external support.

In the *neighbourhood-based* (**NEB**) approach, a decision to invest is based on the fraction x_f of neighbours of a focal individual with the desirable behaviours, calculated at the local level. Investment happens if x_f is at most equal to a threshold n_f (i.e. $x_f \leq n_f$), for $0 \leq n_f \leq 1$; otherwise, no investment is made.

As the presence of structural heterogeneity in scale-free networks introduces a level of inequality between nodes in terms of influence, we also examine a *node-influence-based* (NI) approach. Here, we build upon the existing literature by resorting to two measures for defining a node's influence, *degree centrality* (**NI-DEG**) and *eigenvector centrality* (**NI-EIG**). For a comprehensive description of these two measures of centrality, see Section 2.4.4 in Chapter 2.

We denote by x_i the node's centrality measure (e.g. for NI-DEG: $x_i = deg_i = x_i^{deg}$). The nodes are sorted in ascending order based on their influence x_i, and the threshold i_f denotes the fraction of nodes that will be selected for interference, if their behaviour satisfies the given targeting approach. For instance, given a network of size 1000 and a threshold $i_f = 0.001$, this would mean selecting only the most influential node in the network for investment.

Irrespective of the interference scheme or the targeted behaviour, the threshold signifies an increase in the number of nodes that satisfy the requirements for investment. In other words, a threshold of 1 means investing in all nodes which follow the desired strategy. Conversely, a lower threshold implies a more careful approach to investment, whereby the exogenous agent is stricter in their selection of suitable candidates. Moreover, the target selection also affects the number of candidate nodes. Stricter schemes, such as targeting individuals who are fair when proposing, and also when responding (HH), narrow the search for nodes which satisfy the requirements even further.

4.3 Results

When choosing to invest in a population of individuals in an effort to ensure some form of desirable outcome, an external decision maker must first consider several factors before any decision is made. Among these, we consider and aim to resolve the questions regarding what sort of behaviour they should invest in, how large the individual endowment must be, but also what an investor can do when information about the population or the environment is incomplete, or even unknown. As such, we consider that the simplest form of information gathering evaluates the overall population (in the form of some metrics measuring fairness on average), as opposed to fine-grained observations on individual neighbourhoods. Likewise, we consider that ensuring all proposals are fair (i.e. investing in HH or HL) is less demanding on an external decision-maker than ensuring that only fair offers are accepted (i.e. investing in HH and LH), which is, in turn, a simpler endeavour than for both the former and latter to be strictly enforced (choosing to invest in HH only). In this way, we can conceptualise a hierarchy of investment strategies, in terms of complexity, some of which may simply be impossible for an investor to follow, merely due to lack of information, funding, or a combination of the two.

We consider that there exists a *minimal level of fairness* which the external decision maker is aiming to enforce in regards to the population's behaviour [Han and Tran-Thanh, 2018], and we study the least expensive investment strategies for differing preferences of such an acceptable fairness. We will first systematically present our results for homogeneous populations, namely well-mixed and square-lattice graphs, and then we will present our main findings from more realistic, heterogeneous graphs.

4.3.1 Homogeneous Populations

Firstly, we explore the simplest class of investment strategies, using a macroscopic metric of the population, measuring average fairness in the whole system (population).

Well-mixed populations (baseline)

As the foundation of this analysis, we first introduce a baseline analysis of this interference on well-mixed (complete graph) populations, in Figure 4.2. We notice an increase in fairness for all three different targets, if the threshold for investment is sufficiently high, but there are marked differences in the cost of interference. Specifically, targeting both fair responses and proposals (HH), as well as only fair responses (HH LH), reduce the accumulated costs of interference for the external investor for a broader range of parameters than targeting only fair proposals (HH HL). Furthermore, the threshold for investment is the deciding factor for ensuring high levels of fairness for all

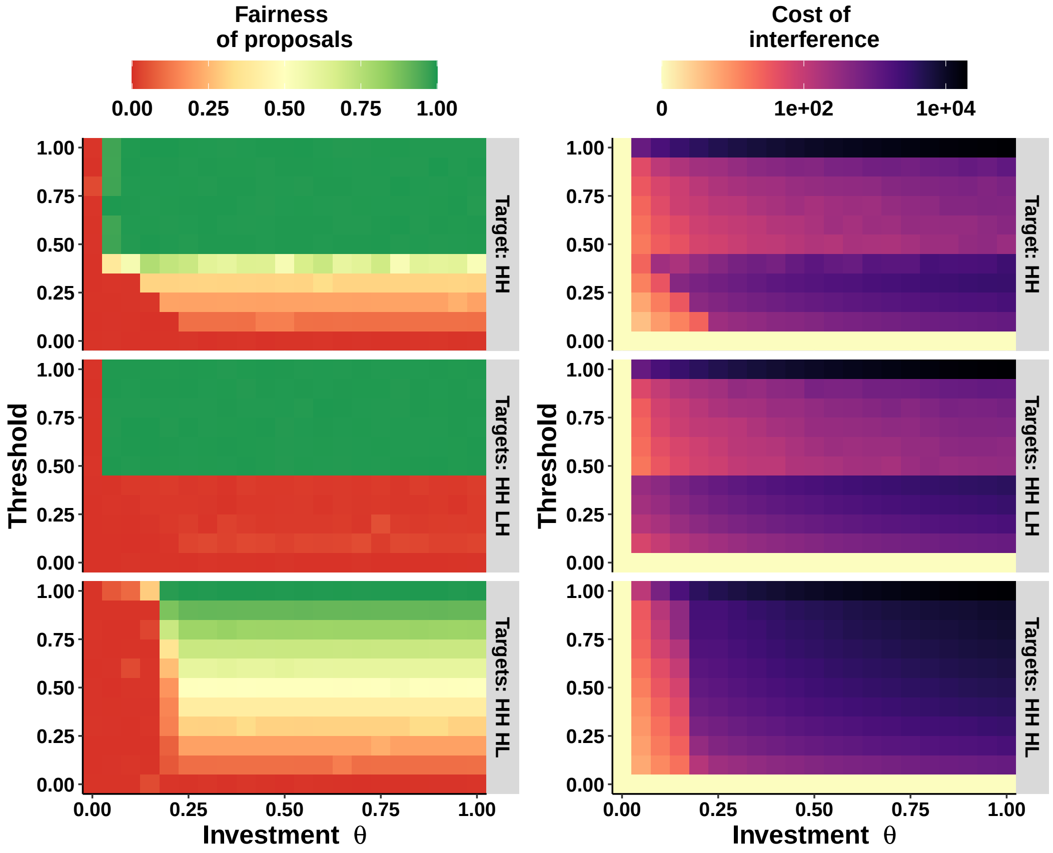

Fig. 4.2 **Baseline results after interference in a well-mixed scenario**. Average fairness (left) and average cost of interference (right) as a function of the individual endowment θ and the threshold p_f (population-based, well-mixed network, $\mu = 0.01$, stochastic update). Each row represents a different targeting scheme. The cost of interference is shown on a logarithmic scale.

cases. This suggests that if certain levels of fair behaviour are maintained, then the population will converge to fairness without requiring further investment.

Based on the amount of information available to the external decision maker, we confirm that more information gathering leads to a more flexible investment approach. Respectively, the strictest approach (targeting HH only) leads to the highest levels of fairness with lowest accumulated costs, followed by ensuring fair responses, and, lastly, promoting fair proposals. Targeting both roles or only fair responses produce almost indistinguishable results if the chosen threshold is sufficiently high ($p_f \gtrapprox 40\%$), whereas only targeting fair proposers is very costly regardless of minimal fairness requirements. These results show that fair responders drive the dynamics of the system in the well-mixed scenario, and they should be targeted correspondingly by an external decision maker.

Population-based interference

We now consider that the population is structured and that individuals interact only with their neighbours. Figure 4.3 shows the results for different population-based interference scheme and clearly demonstrates the difference between the three targets for investment. We would like to point out the higher levels of fairness obtained using the HH targeting scheme, especially for a lower threshold p_f. We also notice an increase in the threshold for investment p_f in order to achieve similar levels of fairness. When it comes to the accumulated cost of interference, we see that HH is the most cost-effective solution, due to the previously perceived lower threshold required to maintain fairness.

Figure 4.4 further exemplifies the finding that targeting HH is the optimal scheme for population-based interference. Each row (portraying the different targeting schemes), drifts further away from the cost-optimal bottom left. As the threshold increases, so does the total cost, so the regions of high fairness for a lower threshold observed in Figure 4.3 coincide with the maximal savings (while still achieving desired levels of fairness).

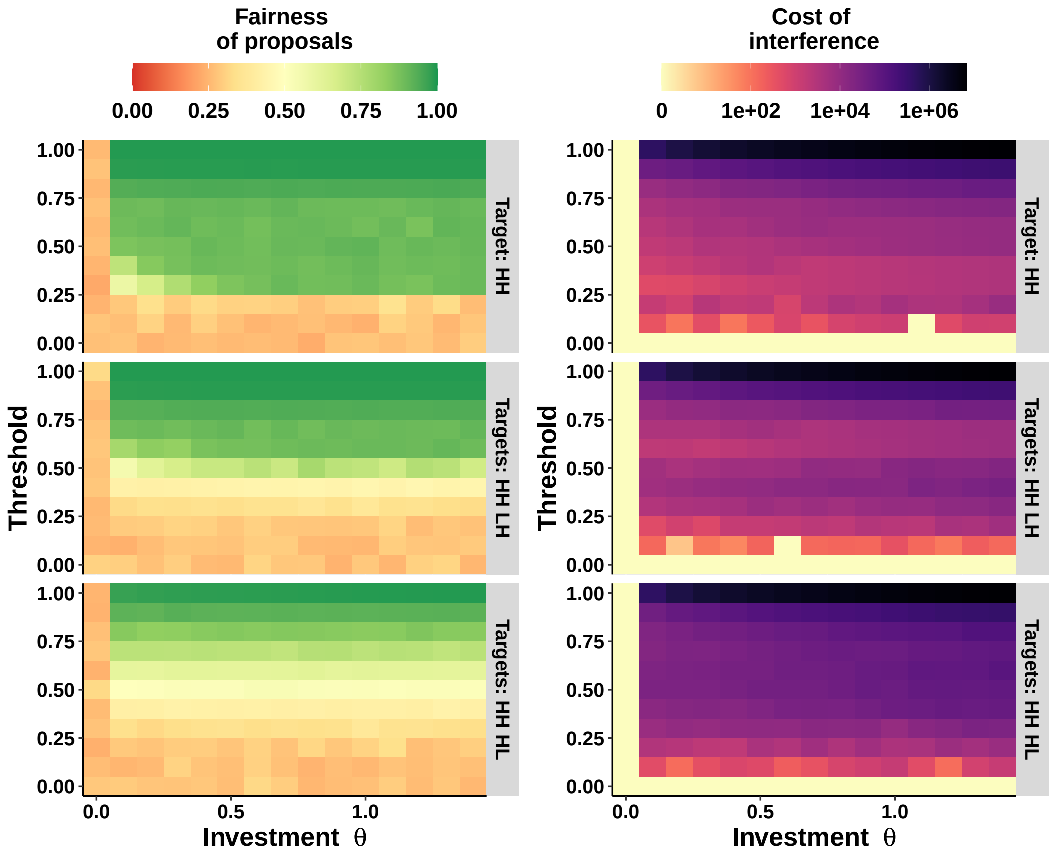

Fig. 4.3 **Population-based interference in square lattice populations**. Average fairness (left) and average cost of interference (right) as a function of the individual endowment θ and the threshold p_f ($\mu = 0.01$, stochastic update). Each row represents a different targeting scheme. The cost of interference is shown on a logarithmic scale.

Table 4.1 **Most cost-efficient POP schemes to reach a minimum fairness of proposals for different mutation rates in SL populations (stochastic update)**. There exist no schemes which satisfy the higher minimum fairness requirements in the case of very high mutation rate, written as '–' in the table.

Mutation rate	Minimum fairness	Target	Threshold	θ	Cost (mean ± 1.96 se)
10^{-4}	75%	HH	0.3	0.1	530 ± 5
10^{-4}	90%	HH	0.3	0.1	530 ± 5
10^{-4}	99%	HH	0.3	0.4	999 ± 7.6
10^{-2}	75%	HH	0.3	0.3	750 ± 5.4
10^{-2}	90%	HH	0.3	0.7	1747 ± 11.2
10^{-2}	99%	HH	1	0.1	487514 ± 93.6
0.2	75%	HH	0.6	0.2	358089 ± 650
0.2	90%	–	–	–	–
0.2	99%	–	–	–	–

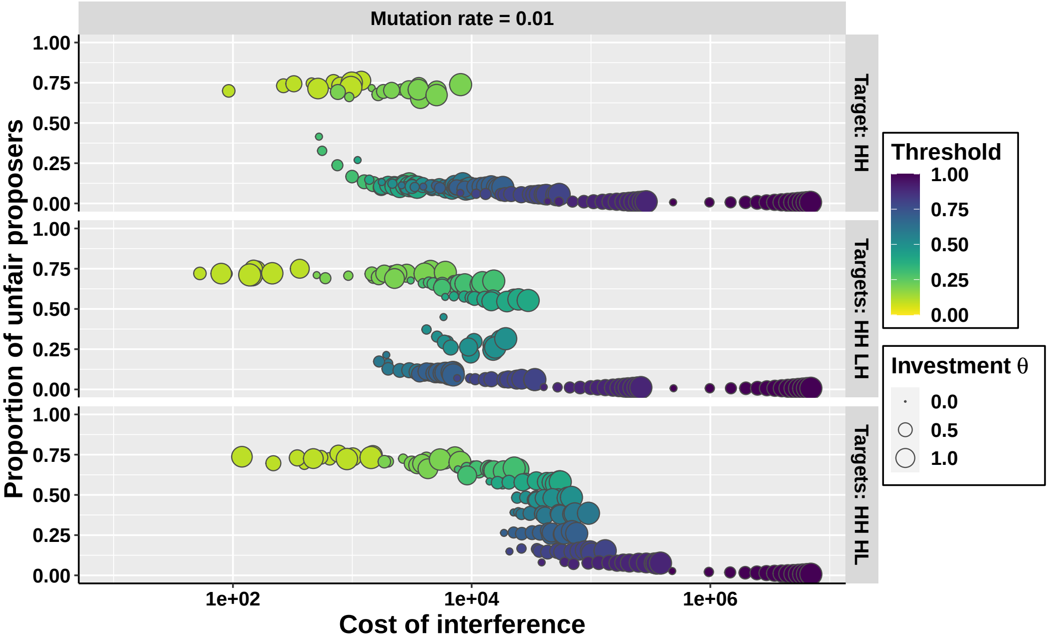

Fig. 4.4 **Pareto fronts for population-based interference in square lattice populations**. Proportion of unfair proposers as a function of average cost of interference for different targeting schemes ($\mu = 0.01$, stochastic update). The size and colour of the circles correspond to investment amount and threshold of investment, respectively. We note that the most desirable outcomes are closest to the origin.

Table 4.1 shows the most cost-efficient schemes for ensuring specific standards of fairness when only a population-based approach is possible, under differing rates of mutation (μ). We observe a definitive bias towards the most complex investment scheme (i.e. targeting HH players), which reiterates our previous observation. We note that, in order to maintain a desired level of fairness, an external decision maker must increase the threshold at which they resume their investment, but also the individual endowment (θ). It becomes increasingly difficult to maintain standards of fairness when the population is exposed to high degrees of behavioural exploration and this naturally attracts an increase in the total cost for the investor. We report similar figures for other values of μ in Figures A.2, A.3, A.4, in Appendix A.

Moreover, we observe an increase in fairness for all schemes of interference, across most values of individual endowment θ, which bodes well when the external decision maker possesses limited knowledge. If reducing cost is not the main objective, fairness can be maintained using any targeting scheme (i.e. any relevant observations made about the population), by increasing the minimum threshold p_f.

When the external decision maker is limited to the macroscopic metrics associated with population-based interference, interference is characterised by its strictness. To elaborate, information gathering should be the main goal for the investor, as ensuring that proposals and responses are simultaneously fair (i.e. targeting HH) is the optimal outcome. In this way, the minimum threshold can be kept low, reducing the accumulated cost. These findings are robust when compared to well-mixed populations, although it is easier for an investor to maintain fairness in the case of structured populations, when targeting fair proposers is the only option for investment.

Neighbourhood-based interference

Previous works on the PD have shown that the greatest gains in cooperation (while maintaining a minimal investment cost) require very detailed observations of individual neighbourhoods, coupled with overly strict investment schemes [Han and Tran-Thanh, 2018; Han et al., 2018]. In order to decipher

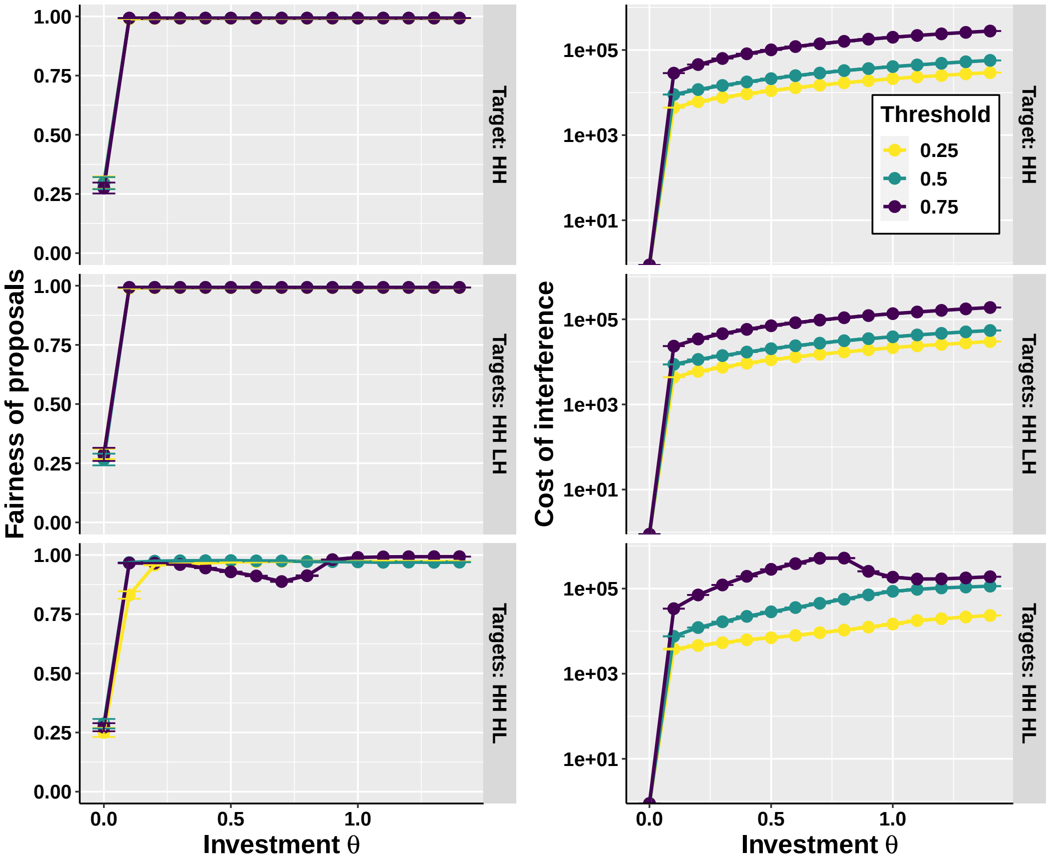

Fig. 4.5 **Neighbourhood-based interference in square lattice populations**. Average fairness (left) and average cost of interference (right) as a function of θ and threshold n_f ($\mu = 0.01$, stochastic update). Each row represents a different targeting scheme. The cost of interference is on a logarithmic scale.

whether or not these findings hold for the spatial Ultimatum Game, we study the outcome when an investor can perceive fairness at the local level.

Figure 4.5 reports the relationship between gains in fairness and increases in cost for an external investor, with diverse targets for receiving investments. We observe that fairness is more easily achieved than in population based interference, with only a very low investment required to sustain a majority of fair proposals. Further investment increases the cost of interference, but only slightly. If different thresholds result in fairness, Figure 4.5 shows that a threshold of 25% is the most cost-efficient. Similarly to population-based interference, the external decision maker should invest only when a large proportion of unfair individuals are present to limit the cost of investment. Finally, there are no significant differences between targeting schemes.

Similarly to our findings using a population-based approach, we observe that the more prohibitive option, HH, is also the most cost-effective. On the other hand, high fairness can be achieved in all three cases for the same values of endowment. Ensuring that all proposals are fair (thus investing in HH and HL), can lead to an increase in cost of interference, and a decrease in fairness gains (relative to the other two interference strategies). While all investment schemes evidently succeed in promoting the evolution of fairness, only ensuring the equitable proposals is not as reliable as encouraging discerning responses to offers or both. We note that this effect can only be seen when the threshold for investment is very high (i.e. an investor only invests in neighbourhoods with three or more fair proposers). As discussed earlier, investing in neighbourhoods with at most one fair agent and not investing otherwise, solves this dilemma.

Markedly, it is not effective to invest in neighbourhoods with a high percentage of fair proposals. These results point to a key observation, that it is more important to invest in fair proposers when there are few of them in a specific neighbourhood. In this sense, the lonely fair individuals require aid in otherwise competitive, unjust entourages. This result can further be seen in Figure 4.5. By being very selective with which neighbourhoods the external investor chooses to invest in (i.e. only choosing very fair neighbourhoods), they inadvertently produce a much higher final cost to their own selves. An external decision-maker would then unwittingly keep investing

Table 4.2 **Most cost-efficient NEB schemes to reach a minimum fairness of proposals for different mutation rates in SL populations** (stochastic update). There exist no schemes which satisfy the higher minimum fairness requirements in the case of very high mutation rates, written as '–' in the table.

Mutation rate	Minimum fairness	Target	Threshold	θ	Cost (mean ± 1.96 se)
10^{-4}	75%	HH	0.25	0.1	1395 ± 36.9
10^{-4}	90%	HH	0.25	0.1	1395 ± 36.9
10^{-4}	99%	HH	0.25	0.1	1395 ± 36.9
10^{-2}	75%	HH HL	0.25	0.1	3794 ± 200.1
10^{-2}	90%	HH LH	0.25	0.1	4352 ± 56.2
10^{-2}	99%	HH LH	0.25	0.2	5957 ± 60.7
0.2	75%	HH	0.25	0.4	150777 ± 121.5
0.2	90%	–	–	–	–
0.2	99%	–	–	–	–

in fair proposals ad infinitum because fairness is eventually reached in the Ultimatum Game, even when individual endowment is relatively low. It is clear, therefore, that to reduce potential costs, only players in unfair groups should be eligible for investment. Therefore, the defining characteristic of neighbourhood-based interference is the low threshold for investment (25%).

By varying minimal fairness requirements and rates of mutation, we can gain further insight into which investment strategies are the most robust and cost-effective. Table 4.2 highlights some surprising findings. We see that neighbourhood based interference can result in a higher total cost than the optimal population-based interference schemes (see Table 4.1). Previous works have shown that more specific and restrictive intervention schemes are more effective in the PD [Cimpeanu et al., 2019; Han et al., 2018], but by being able to target different roles in the Ultimatum Game, these differences can be mitigated. Furthermore, mutation rate serves as an equaliser between the investment targets, and we observe that less specific schemes (HH & HL and HH & LH) are the most cost-efficient options. We note that the differences between results are small enough that different runs could yield any outcome in the case of high or intermediate mutation rates. The lack of significant variability among the distinct targeting schemes contrasts strongly with the findings on the PD [Cimpeanu et al., 2019; Han and Tran-

Thanh, 2018]. We report similar figures for other values of μ in Figures A.6, A.7, A.8, in Appendix A.

Evolution of strategies over time

We make use of the optimal parameter values identified in Tables 4.1 and 4.2 to explore the evolution of fairness over time for all the strategies in the population, as well as any associated accumulated costs. Through this analysis, we clarify some of the dynamics differentiating the different decisions for investment, as well as the effects of varying mutation rates upon the outcomes and the options available to investors.

The effects of mutation on the optimality of different interference schemes can be seen in Figure 4.6. As the mutation rate (μ) increases, the capacity of maintaining a threshold of fairness decreases (as also seen in Table 4.2). An external investor must increase their individual investment amount in order to meet these new demands set out by the increased mutation rates, and by doing so they can maintain fairness levels to a respectable standard.

To better highlight the sharp increases in the cost associated with the non-optimal threshold (i.e. when it is greater than 25%) for neighbourhood-based interference, we show such typical runs for varying mutation rates for the 50% threshold in Figure 4.7. When comparing Figures 4.6 and 4.7, we note the relative differences in total accumulated costs attributed to the choice of the threshold for investment n_f. We also note that increasing rates of behavioural exploration (mutation) amplifies this discrepancy.

We show how less specific interference strategies, which require less information gathering, can be effective in facilitating the evolution of fairness, when local monitoring is possible (Figure 4.9). Promoting fair proposals may often not be sufficient for low individual investment budgets (which are also the optimal solution) — in such cases fairness does not evolve. This occurs due to the inability of indiscriminate fair proposers to protect themselves against unfair proposers. Investing in fair proposers, in this case, artificially protects them against very competitive selection pressures.

Figure 4.8 showcases how different mutation rates call for different approaches to interference. As shown previously, optimal interference strate-

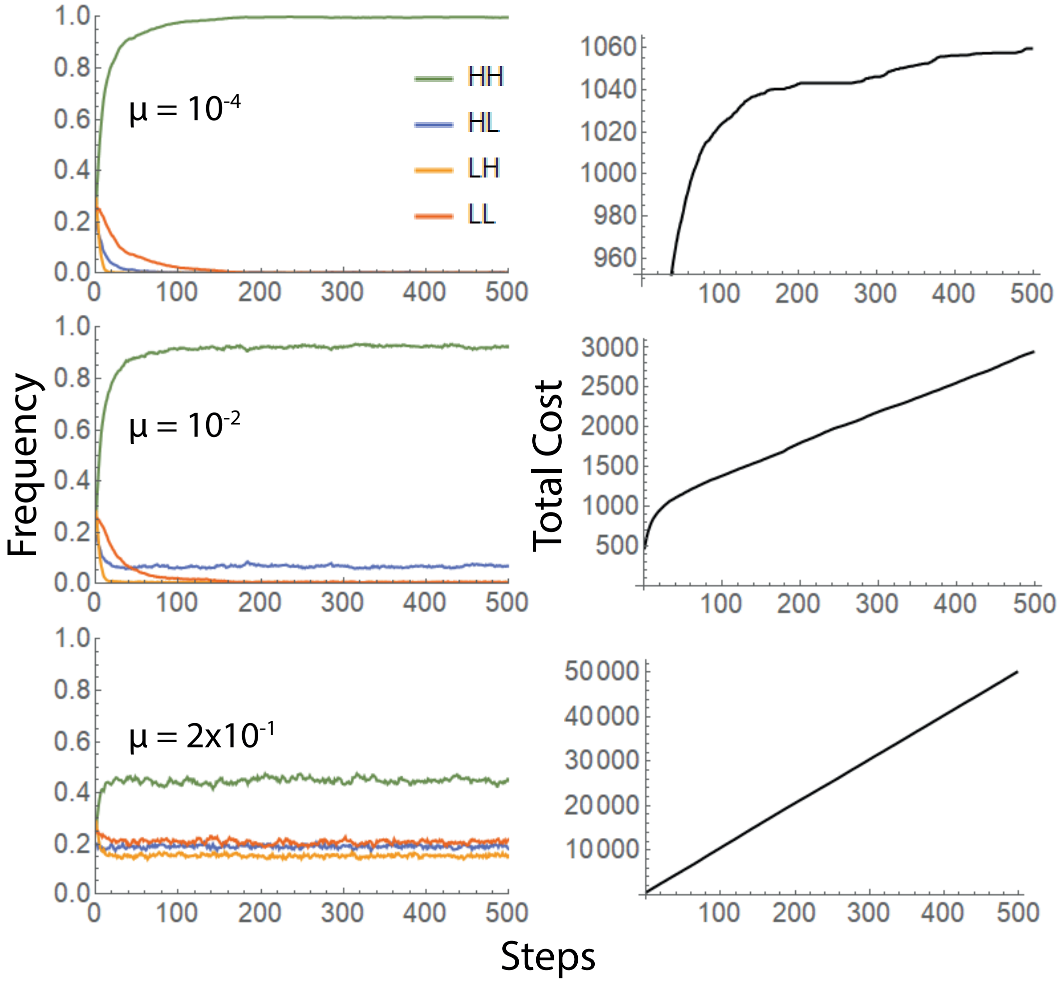

Fig. 4.6 **The effect of mutations in the evolution of fairness (low threshold).** Typical runs showing the evolution of fairness and the associated total cost of interference for various mutation rates in SL populations (top row $\mu = 10^{-4}$, middle row $\mu = 10^{-2}$, bottom row $\mu = 2 * 10^{-1}$; neighbourhood-based, stochastic update). Parameters: $n_f = 0.25$, $\theta = 0.1$, Target = HH. The choice of parameter values was motivated by selecting the optimal solutions in Table 4.2.

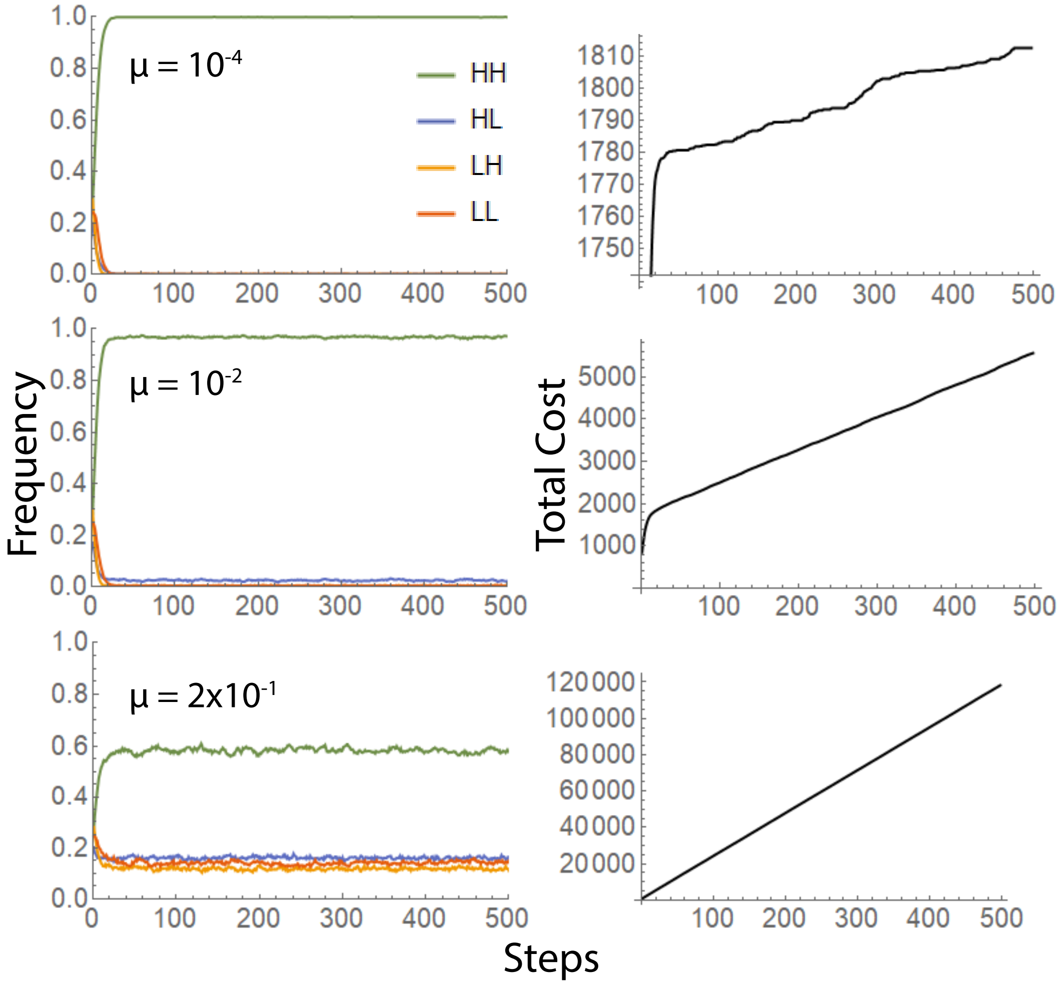

Fig. 4.7 **The effect of mutations in the evolution of fairness (high threshold)**. Typical runs showing the evolution of fairness and the associated total cost of interference for various mutation rates in SL populations (top row $\mu = 10^{-4}$, middle row $\mu = 10^{-2}$, bottom row $\mu = 2 * 10^{-1}$; neighbourhood-based, stochastic update). Parameters: $n_f = 0.5$, $\theta = 0.1$, Target = HH. The choice of parameter values was motivated by selecting the optimal solutions in Table 4.2.

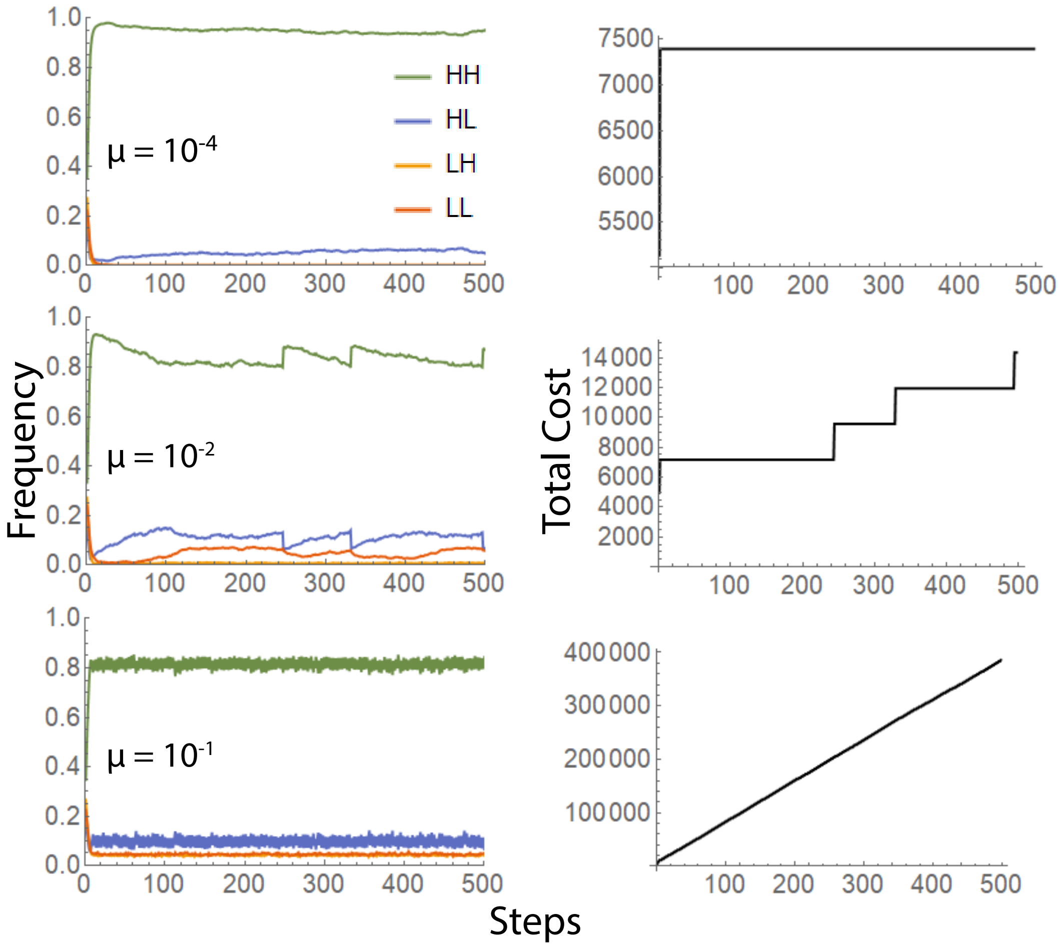

Fig. 4.8 **Higher mutation rates leads to an increasing need for interference over time**. Typical runs showing the evolution of fairness and the associated total cost of interference for various mutation rates in SL populations (top row $\mu = 10^{-4}$, middle row $\mu = 10^{-2}$, bottom row $\mu = 2 * 10^{-1}$; population-based, stochastic update). Parameters: $p_f = 0.8$, $\theta = 0.3$, Target = HH. The choice of parameter values was motivated by selecting the optimal solutions in Table 4.1.

Table 4.3 **Most cost-efficient population-based schemes (deterministic update) to reach a minimum fairness of proposals in SL populations.**

Minimum fairness	Target	Threshold	θ	Cost (mean ± 1.96 se)
75%	HH	0.5	0.5	1251 ± 10.8
90%	HH	0.6	0.9	2228 ± 22.6
99%	HH	0.9	1.1	5488 ± 22.9

Table 4.4 **Most cost-efficient neighbourhood-based schemes (deterministic update) to reach a minimum fairness of proposals in SL populations.**

Minimum fairness	Target	Threshold	θ	Cost (mean ± 1.96 se)
75%	HH	0.25	0.8	2146 ± 56.3
90%	HH	0.25	0.8	2146 ± 56.3
99%	HH	0.25	1	2513 ± 16.4

gies vary according to the mutation rate. We point out the three different cases in which an investor might find themselves in. First, when few initial rounds of investment are enough for the system to converge and stabilise to a desired state. Second, an investor might be required to reinvest when the population tends to revert back to its initial condition. Lastly, constant investment is required to maintain a desired level of fairness, with the total cost skyrocketing accordingly. To some extent, a fair population can better deal with unfair invaders and this explains the need for a sufficiently high initial investment when mutation rates increase.

Finally, behavioural exploration motivates the manner or strength (in terms of individual endowment) of any initial efforts to moderate unfair behaviour. Figures 4.6, 4.7 and 4.8 show that the increase in cost is linear and ever-growing for high mutation-rates and gradually sharper at the beginning for lower mutation, eventually plateauing when the population is exposed to little or no behavioural exploration.

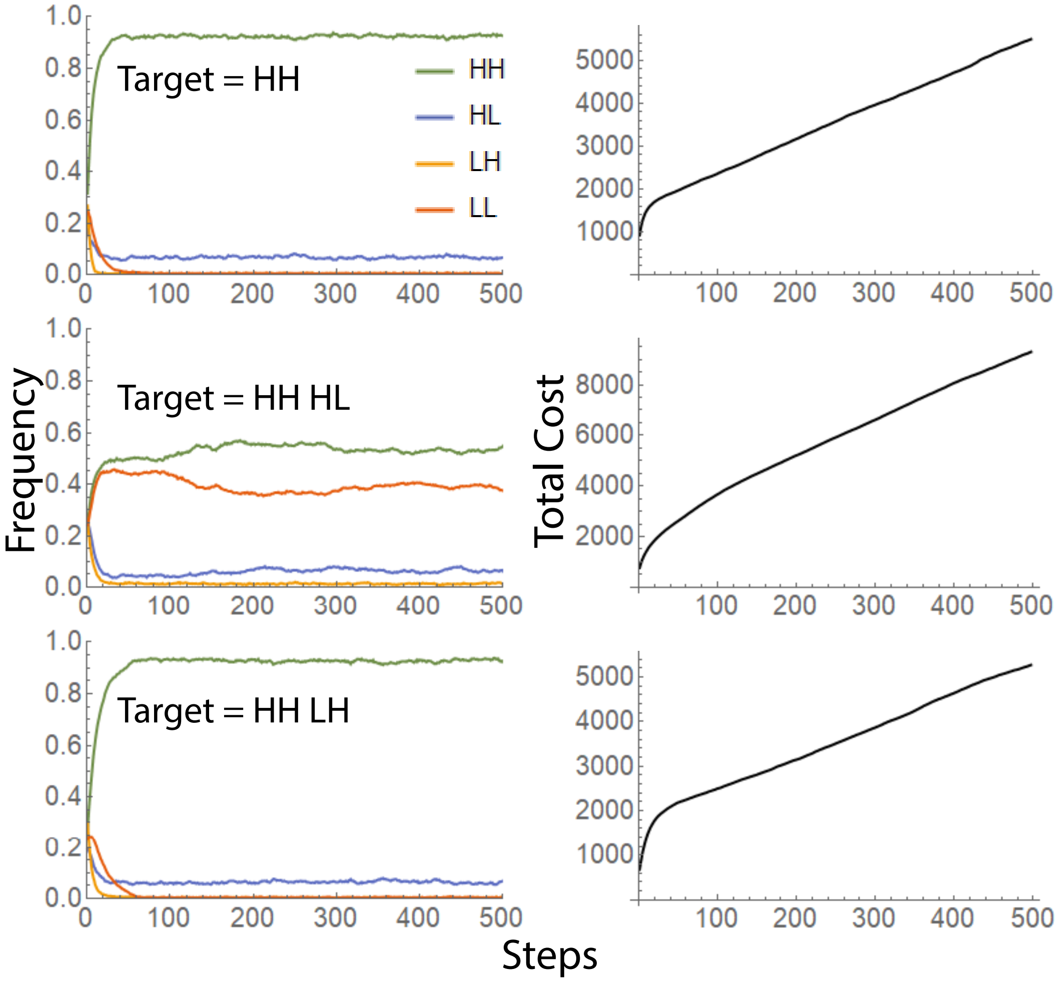

Fig. 4.9 **The choice of target influences the number of eligible candidates and the total costs of investment.** Typical runs showing the evolution of fairness and the associated total cost of interference for various targeting schemes in SL populations (neighbourhood-based, stochastic update). Parameters: $n_f = 0.25$, $\theta = 0.2$, $\mu = 10^{-2}$. The choice of parameter values was motivated by selecting the optimal solutions in Table 4.2.

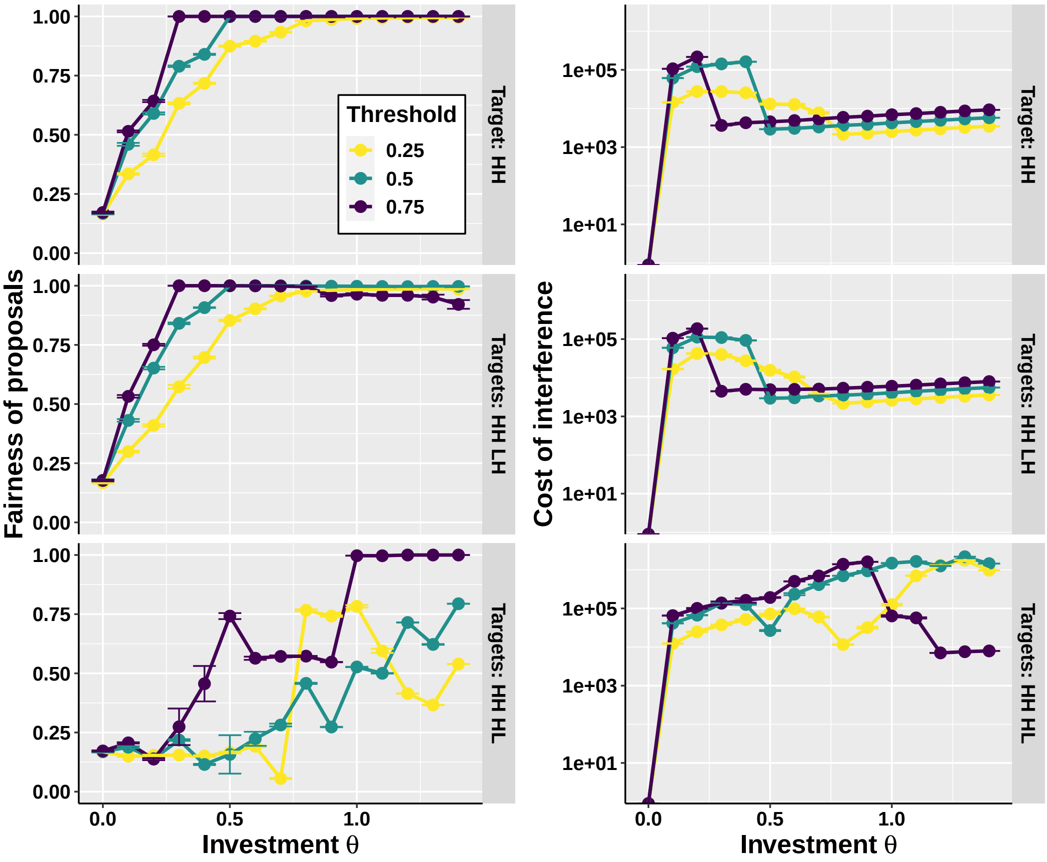

Fig. 4.10 **Neighbourhood-based interference in square lattice populations (deterministic update)**. Average fairness (left) and average cost of interference (right) as a function of θ and threshold n_f. Each row represents a different targeting scheme. The cost of interference is on a logarithmic scale for clarity.

Table 4.5 **Most cost-efficient interference schemes to reach a minimum fairness of proposals in BA networks**. For each minimal standard of fairness, we highlight (in bold) the least costly options across schemes.

Scheme	Minimum fairness	Target	Threshold	θ	Cost (mean ± se)
POP	75%	HH	0.2	56.23	168655 ± 14592
POP	90%	HH	0.4	74.98	176377 ± 14389
POP	99%	HH LH	0.8	56.23	293956 ± 20785
NEB	75%	HH LH	0.7	56.23	112870 ± 944
NEB	90%	HH LH	0.7	56.23	**112870 ± 944**
NEB	99%	HH LH	0.7	56.23	**112870 ± 944**
NI-DEG	75%	HH	0.005	17.78	**66891 ± 2185**
NI-DEG	90%	HH LH	0.007	23.71	143260 ± 2000
NI-DEG	99%	HH	0.017	31.62	512727 ± 1885
NI-EIG	75%	HH	0.003	31.62	**66252 ± 2623**
NI-EIG	90%	HH	0.003	74.98	190862 ± 3974
NI-EIG	99%	HH	0.017	42.16	705906 ± 2352

Deterministic update

The results and findings reported so far were based on the stochastic update rule. We now take a step back and consider whether our findings would still hold true for the deterministic rule (see again Section 4.2.2). It is not only for the sake of a direct comparison with a previous analysis reported in [Han et al., 2018], where cost-efficient interference was studied for the spatial PD in a deterministic setting (with no mutation). It would also allow us to examine if the findings above would remain robust for the deterministic update, a popular approximation for rare stochastic effect (or infinite intensity of selection) that is regularly used in the literature [Szabó and Fáth, 2007]. Tables 4.3 and 4.4 report results for the optimal interference strategies, in population-based (for a full report, see Figure A.9, in Appendix A) and neighbourhood-based schemes, respectively. We observe that for both schemes, targeting HH is always the best option. This is the same as the stochastic approach for population-based schemes but different from the neighbourhood-based ones. However, for the latter ones, the optimal threshold $n_f = 0.25$ remains the same as in the case of stochastic update, see also Figure 4.10. This is in stark contrast with the PD results where $n_f = 0.75$ was always the optimal choice.

4.3.2 Heterogeneous Populations

We structure the subsections that follow according to key insights derived from the results, and refer to previous results on homogeneous populations. We will highlight the key differences that arise in the presence of diversity (in the form of spatial heterogeneity), and mention similarities where appropriate.

Diversity reduces interference complexity

Diversity introduces several challenges which must be overcome by an external decision maker, but these bring with them opportunities to exploit the inherent mix of strategies that can be successful, according to different initial network conditions. If we consider a hierarchy of complexity based on the inherent costs of gathering information as explained above, we show that two targets have the potential to be optimal in a wide range of schemes and fairness requirements (see Table 4.5). Ensuring both offers and responses are fair is the strict, but also intuitive approach to investment. Nevertheless, several configurations in which rewarding fair responders (i.e. HH LH) succeed as the most cost-effective avenues towards fairness.

In the presence of diversity, we see that strictness, while generally effective, is not necessarily optimal. Heterogeneity allows for the coexistence of several strategies in a cluster, and relaxing the eligibility conditions for investment can allow an external decision-maker to reinforce positive behaviour, ultimately producing the optimal outcomes shown in Table 4.5. These results contrast starkly with previous observations in structured populations, for which stringent information gathering (i.e. targeting HH) leads to the lowest total costs and the highest levels of fairness, in almost all cases. Previously, for structured populations, relaxing targeting conditions was only desirable in the presence of high mutation-rates. Diversity acts in a similar fashion to the noise associated with high mutation-rates, which also allows the coexistence of several strategies.

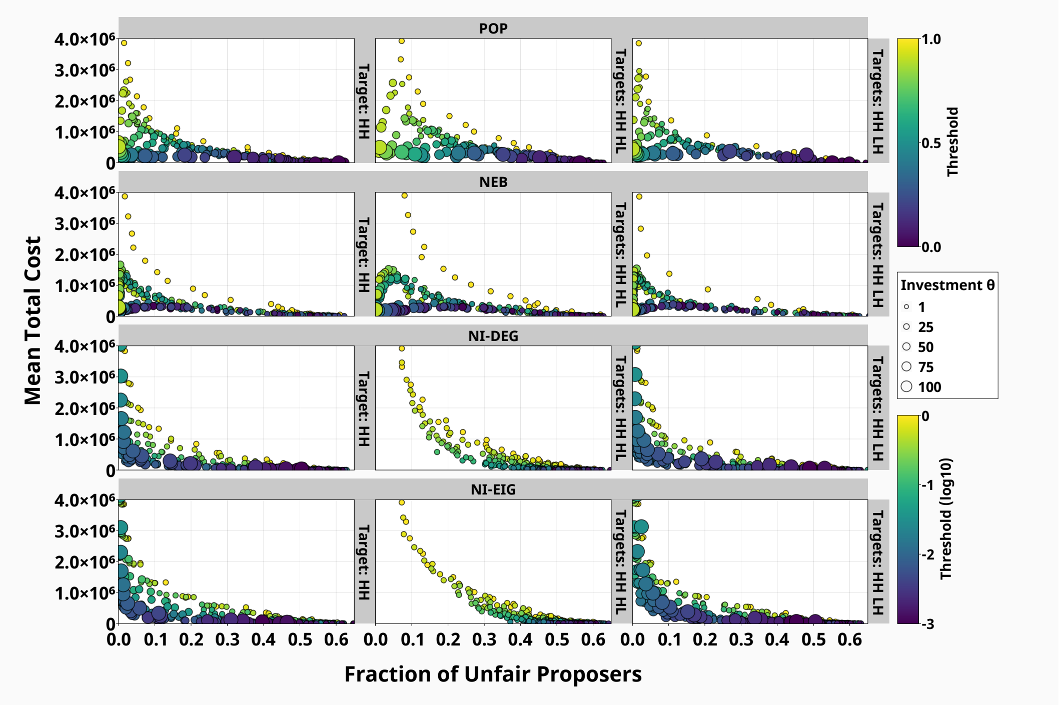

Fig. 4.11 **Pareto fronts for each scheme in BA networks**. Proportion of unfair proposers as a function of the average interference cost for each scheme and target combination. The markers' size is determined by the individual investment θ (grouped to the nearest value), whereas the colour is determined by the threshold. Markers near the origin indicate the optimal solutions. Note that we only show the most cost-effective solutions, by limiting the maximum total cost.

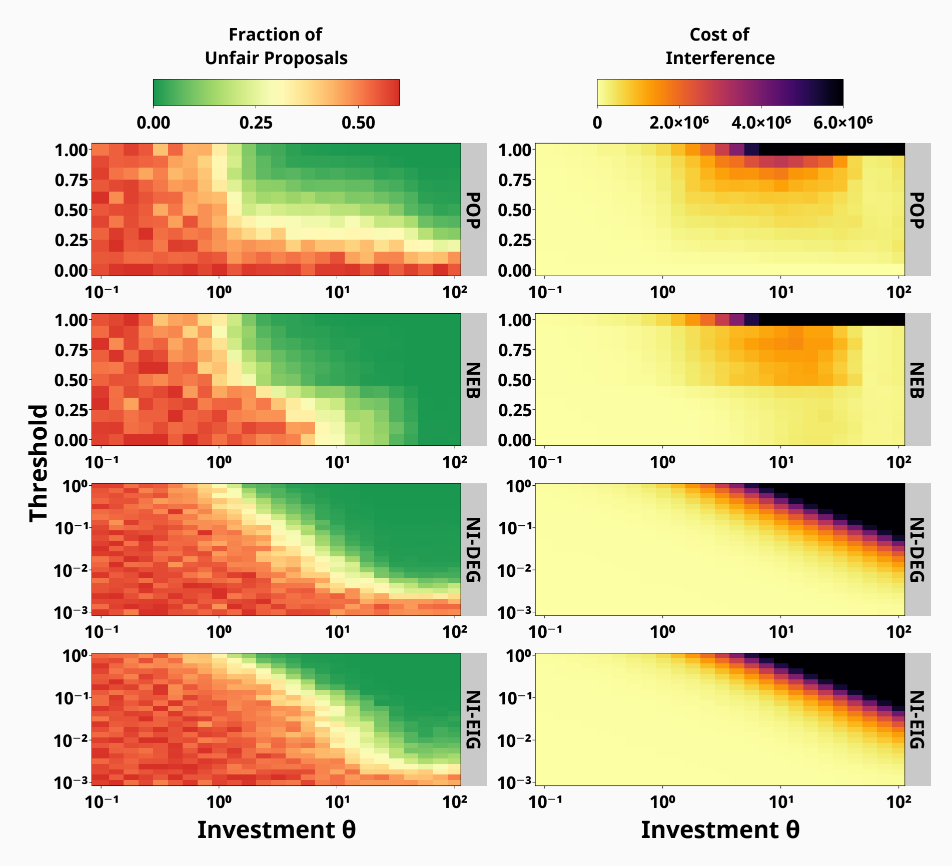

Fig. 4.12 **Targeting fair responders in BA networks**. Proportion of unfair proposers and total costs of investment for each scheme, while targeting fair responses (HH and LH).

Polarisation towards fairness

Departing significantly from previously discussed results on homogeneous populations, we observe a tendency towards polarisation (see Figures 4.11 and 4.12). Across a wide spectrum of both interference schemes and various targets, we show that a very large θ can propel the system towards fairness, while also minimising cost. After having reached close to 100% fairness, it is difficult for unfair strategists to invade the network, due to the heterogeneous dynamics at play, so no further investment is required.

By varying the threshold for investment, the amount of funding that is required for the system to shift towards fairness also changes. In a phenomenon akin to energy landscapes encountered in physical systems, the "energy" required to push fairness across the local maxima increases the further away it gets from the global minimum. Lowering the threshold can be beneficial, as overspending is avoided. On the other hand, the obstacle that arises is two-fold. As the goal is ultimately to reach a high level of fairness, a low threshold allows unfair individuals to thrive before fair individuals are eligible for investment. In other words, the system dips further away from the global minimum, and the investment amount required increases appropriately.

Theoretically, this would allow for a number of viable investment approaches in such a system, but there are also several practical concerns of note. Institutions wishing to employ the practice of increasing their endowment amounts would be expected to have access to a considerable amount of initial funding, as opposed to spreading out the costs over multiple investment rounds. Moreover, there exists an intermediary region where overspending becomes problematic (see Figure 4.12). This issue occurs when the investment amount is large enough to be effective at inducing the change towards fairness, but not at the level where it does so rapidly. Given a relatively high threshold, many candidates become suitable for assistance, thus leading to excessive funds being deployed in locations of the network which are unimpactful.

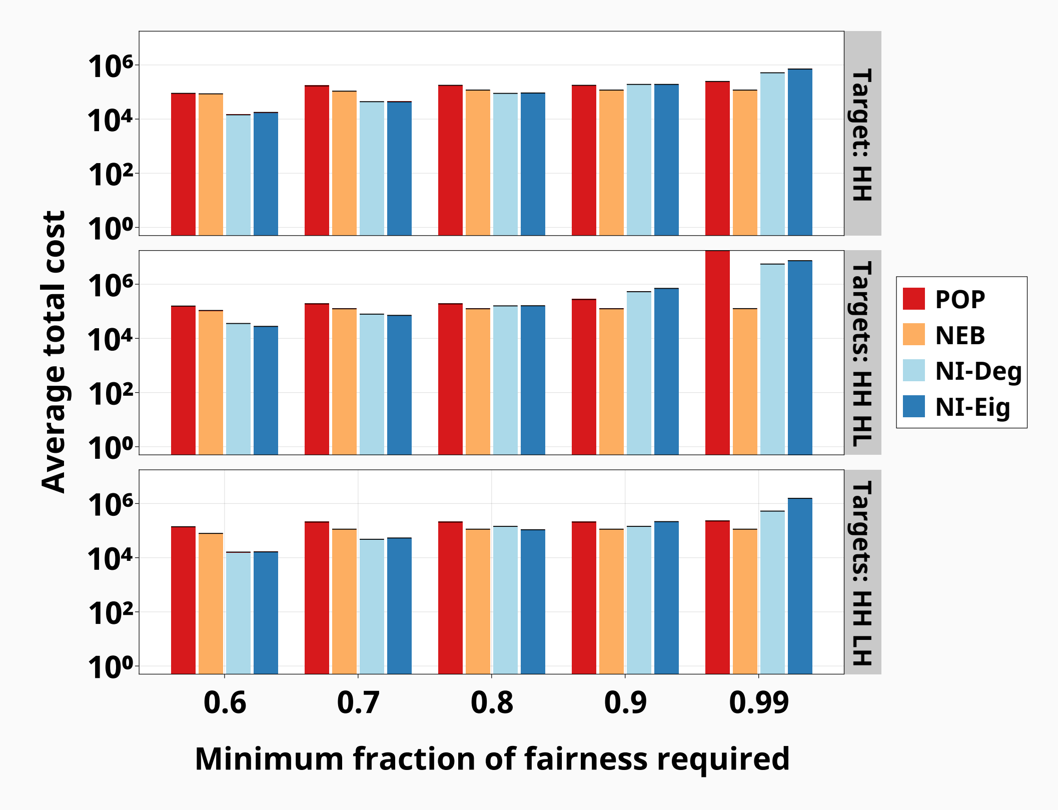

Fig. 4.13 **Most efficient schemes in BA networks**. Mean total costs (scaled by log10) for the most efficient combinations of threshold and investment amount θ for each possible target and scheme. Error bars are shown in light red.

Standards of fairness stipulate divergent approaches

Fundamentally, institutional incentive schemes, in the context of asymmetric interactions and the Ultimatum Game, are diametrically different to cooperative dilemmas where heterogeneity naturally promotes pro-sociality [Cimpeanu et al., 2019; Santos and Pacheco, 2005; Santos et al., 2008]. As the UG allows for a larger room for improvement, due to relatively low baseline fairness levels (see again Section 4.2.1 and Figure 4.1), an investor can modify their goals and opt for modest improvements. They could, for instance, decide to invest only in large hubs, and ignore any potential outliers. Whether due to budget constraints, lack of information, or even uncertainty of network characteristics, they could adjust their margin for improvement and prevent unnecessary spending. The solutions presented in Figure 4.11 are by definition equally optimal, always implying a trade-off between reducing cost or seeing an increase in unfair proposals. Here, we attempt to answer how best to choose a solution that maximises the efficiency of the chosen scheme based on a desired standard of fairness. Figure 4.13 shows the average total cost required for optimal investment schemes across a wide range of goals. We note the differences in the scales of the y-axes, which imply that on average, the strictest target (HH) is also the cheapest, followed by ensuring fair responses (HH LH) and finally only ensuring fair proposals (HH HL), which is significantly more expensive on average. This variance increases substantially with higher requirements for fairness.

Given a low enough minimum fairness requirement, targeting hubs (highly connected nodes) prevents over-spending. As these requirements increase, we observe that the hubs' spheres of influence do not extend far enough towards the leaves of the graph for them to make a marked improvement. Local observations, while comparatively expensive for most minimal fairness requirements, have the benefit of being able to extend their reach to lowly connected nodes.

Moreover, this relationship between the two schemes (NI and NEB) is exacerbated for more demanding targets. Ensuring only fair responses implies the suitability of a swathe of nodes, thus being comparatively more effective at promoting fairness than the stricter target (HH), which restricts node candidacy. An external decision maker should be strict about which hubs to

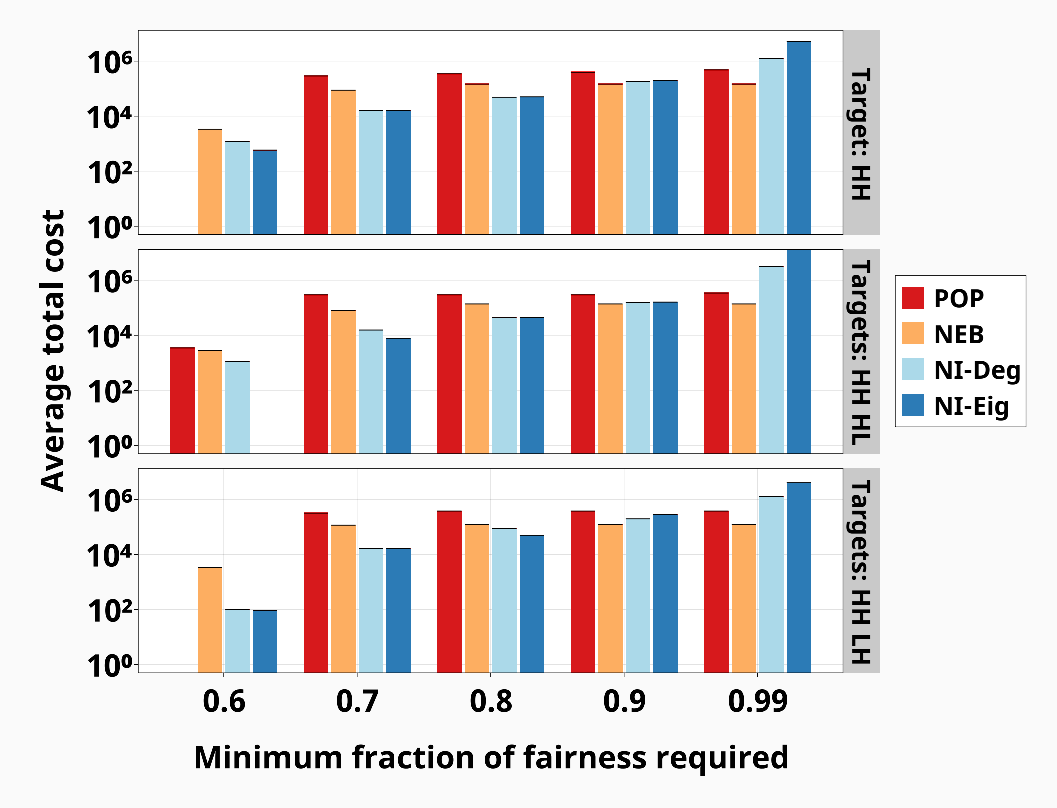

Fig. 4.14 **Most efficient schemes in DMS networks**. Mean total costs (scaled by log10) for the most efficient combinations of threshold and investment amount θ for each possible target and scheme. Error bars are shown in light red. If a certain scheme is missing, no investment was triggered for to each that desired standard of fairness.

invest into, but can be more lenient in the selection of sparsely connected individuals. This is a promising result, as the cost of information gathering is assumed to scale with the number of subjects. Investors could potentially afford to spend the implied additional costs to scrutinise influential individuals, while allowing for a much broader classification if they opted for local observations, instead.

Clustering further reduces the burden on investors

Real-world scale-free networks often have high transitivity (i.e. clustering), a feature missing in BA networks [Barrat and Pastor-Satorras, 2005; Su et al.,

Table 4.6 **Most cost-efficient interference schemes to reach a minimum fairness of proposals in DMS networks.**

Scheme	Minimum fairness	Target	Threshold	θ	Cost (mean ± se)
POP	75%	HH HL	0.8	74.98	294102 ± 34137
POP	90%	HH HL	0.8	74.98	294102 ± 34137
POP	99%	HH HL	0.9	74.98	342927 ± 30915
NEB	75%	HH LH	0.3	74.98	123497 ± 7776
NEB	90%	HH LH	0.3	74.98	123497 ± 7776
NEB	99%	HH LH	0.3	74.98	123497 ± 7776
NI-DEG	75%	HH LH	0.004	10.00	25173 ± 1060
NI-DEG	90%	HH HL	0.004	42.16	158095 ± 1113
NI-DEG	99%	HH	0.031	42.16	1260950 ± 5993
NI-EIG	75%	HH HL	0.001	17.78	14995 ± 376
NI-EIG	90%	HH HL	0.004	42.16	160555 ± 1343
NI-EIG	99%	HH LH	0.177	23.71	4026960 ± 32543

2016], so it is crucial to measure the effects of network transitivity on the choice of investment policies. In the absence of investment, high transitivity positively influences fairness (see Figure A.20 in Appendix A), thereby lowering the total amount of costs required to reach minimal standards of fairness (see Figure 4.14). Moreover, eigenvector centrality (NI-EIG) can be employed to reduce overall spending for moderate fairness requirements, and hubs can be exploited to reduce spending for all but the most strict fairness regimes, unlike in lowly clustered networks. Comparing optimal investment schemes, we show that costs remain similar across all but the minimal desired standards of fairness (see Figure 4.14 and Table 4.6). Strict fairness regimes can be enforced using local information without overspending, similarly to BA networks, yet we also show that population-based metrics are less risky in highly clustered networks. Thus, network transitivity acts as an equaliser between the different schemes and targets.

Previously, we had seen that targeting proposers and responders (HH) and solely proposers (HH and HL) were both sensible approaches towards leveraging fairness. In the presence of clustering, targeting fair responders (HH LH) becomes viable and in some cases, optimal (see Table 4.6). In fact, we see that diversity and transitivity act similar to noise or behavioural exploration in the choice for investment, but on a much broader scale. Both of these factors open up novel mechanisms of engineering fairness, while minimising the risk of choosing inappropriate candidates for endowments. With this reduction in complexity, we also see a very slight increase in the overall costs required to promote fairness. Paradoxically, a higher baseline of fairness also requires more endowments to reach fully positive outcomes. Heterogeneity acts as an equalizer in the truest sense, aiding in the quest towards a fair society, but also fostering a small minority of unfair individuals.

4.4 Discussion

In summary, this chapter addresses an asymmetric interaction setting, in the form of the Ultimatum Game, in which players have different roles within interactions. We have shown that it is crucial to consider the roles' asymmetry to provide cost-efficient investment strategies, an important feature which was not possible to identify in previous works where symmetric games were studied [Chen and Perc, 2014; Chen et al., 2015; Cimpeanu et al., 2019; Duong and Han, 2021b; Han and Tran-Thanh, 2018; Han et al., 2018; Wang et al., 2019]. Moreover, we have incorporated realistic levels of mutation or behavioural exploration in our analysis and have shown that they strongly affect the manner in which interference should be carried out. Previous works have always omitted mutation or assumed that it is infinitely small, thereby being unable to address this important issue for real-world populations and applications. Finally, we have introduced varying levels of social diversity and network complexity and measured the effects of heterogeneity on external interference, taking into account cost optimisation, limited information and standards of fairness.

We have identified several key features that are required to minimise costs while ensuring positive outcomes. Global information is characterised by strictness, and targeting both roles is required to promote fairness effectively, whereas local endowments are characterised by flexibility, and investment is only required as a last resort. We found that diversity reduces the need for complex information gathering, and allows for less strictness in the eligibility criteria for receiving endowments. Exploiting different measures of centrality can be an useful way to reduce spending if standards of fairness are lowered. Beyond the scope of this work, there still exists a large array of potential alternatives for measures of centrality, such as PageRank, betweenness, closeness, Katz, etc. The two metrics discussed in this work, degree and eigenvector centrality, present little qualitative difference between them. Future work could perhaps explore this relationship further, and test whether the other centrality measures lead to similar findings. These results, regardless of the underlying interaction structure, stand out in stark contrast with previous works on cooperation dilemmas, in which interference schemes require an exceptionally strict investment approach.

These findings indicate that the presence of diversity reduces the complexity of engineering fairness in multi-agent systems, shedding new light on the critical, open problem of engineering pro-social behaviours therein [Paiva et al., 2018].

5 | Making an Example: Signalling Threat in the Evolution of Cooperation

There are three things all wise men fear: the sea in storm, a night with no moon, and the anger of a gentle man.

—Patrick Rothfuss, *The Wise Man's Fear*

In which we introduce the threat of punitive acts in social and institutional settings. Here[1], we lay the foundations for a comprehensive and systematic analysis of the effects of signalling the threat of peer and institutional punishment. We show that fearful defectors can emerge through evolutionary dynamics, leading to an increase in social welfare and cooperation, even when punishment would be ineffective otherwise. Moreover, we find that institutions can enforce cooperation by advertising punitive acts, especially when populations are socially diverse, improving pro-sociality while minimising expenditure. Our results provide a compelling argument suggesting that the threat of punishment is an effective deterrent in social and institutional settings.

5.1 Introduction

Punishment has been suggested as one of the most relevant explanations to understanding how selfish individuals self-organise and enforce cooperation or compliance to social norms in various societies [Boyd et al., 2003; Fehr and Gachter, 2002; Herrmann et al., 2008; Powers et al., 2012; Sigmund et al., 2001]. Numerous empirical studies show human proclivity towards punishing unjust behaviour or violations of social norms, often at great cost to their own selves [De Quervain et al., 2004; Fehr and Gachter, 2002; Herrmann et al., 2008]. Although in modern societies sanctioning systems have been widely implemented in the hopes of enforcing laws, many social norms continue to be upheld by the effects of private sanctions [Fehr and Gachter, 2002]. Moreover, third-party punishment has also been implemented in various online systems, such as virtual agent societies [Savarimuthu et al., 2009] and vendor marketplaces [Michalak et al., 2009], as a mechanism of enhancing pro-social behaviour and norms compliance, by both customers and sellers [Michalak et al., 2009].

Cooperation often emerges under the influence of social punishment (i.e. punishing wrongdoers) [Clutton-Brock and Parker, 1995; Fehr and Gachter, 2002; Han, 2016], but this fails to explain how punishment evolves,

[1] The model and results presented in this chapter are also reported in Cimpeanu and Han [2020a,b].

especially if it is highly costly to punish others. Indeed, it has been concluded that punishment is often maladaptive within game theoretic settings, and that punishment can only evolve if it is cost-effective to do so (i.e. the offender suffers much more from retribution compared to the aggrieved party) [Dreber et al., 2008; Ohtsuki et al., 2009; Wu et al., 2009].

Refusing low offers in the ultimatum game (another form of punishment) in the presence of observers, made the wilful (punishers) more likely to receive higher offers in future interactions [Fehr and Fischbacher, 2003]. The fear of having a low offer refused increased the tendency to present higher offers to obstinate individuals and this may help explain which mechanism(s) allow the promotion of punishment when it is costly to do so. In addition, pre-play signalling has been shown to open new avenues for cooperation to emerge, even when such signals are meaningless [Santos et al., 2011]. One area, to our knowledge, which has not yet been explored is using signalling [Huttegger et al., 2014] to explain the emergence of punishment as a viable strategy in evolutionary games, whereby punishers make their retributive deeds well-known in the population, as a deterrent to malefactors. Threat of punishment has also been indicated as one form of making credible commitments [Han et al., 2012; Nesse, 2001], which becomes another reasonable explanation to the dilemma of social punishment.

The effect of threat of punishment by costly signalling may provide key insights into policy making in the context of distributed systems or artificial intelligence. Indeed, it has been concluded that increasing the probability of developing super-intelligent agents is incompatible with using safety methods that incur delays or limit performance [Bostrom, 2017]. What is more, when technological supremacy can be achieved in the short to medium term, the significant advantage gained from underestimating or even ignoring ethical and safety precautions could lead to serious negative consequences [Armstrong et al., 2016; Cave and ÓhÉigeartaigh, 2018]. One proposed solution to mitigating this dangerous behaviour is to look towards intrinsic measures of encouraging AI research communities to want to pursue safe, beneficial design methodology [Baum, 2017]. Our results show that threat signalling may serve as one intrinsic factor to prevent catastrophic consequences in that regard.

In this chapter, we propose and analyse a novel approach towards explaining the evolutionary advantage of punishers in the context of anonymous interactions [Sigmund, 2010] (without relying on reputation). We make use of evolutionary game theoretic models [Hofbauer and Sigmund, 1998; Sigmund, 2010] (see Section 5.3) to show that signalling acts of punishment can promote the emergence of cooperation in the selfish environment of the one-shot Prisoner's Dilemma (PD) [Sigmund, 2010]. We recall that the PD is also the most difficult pairwise social dilemma for cooperation to emerge in [Hofbauer and Sigmund, 1998] (for a detailed discussion please see Section 2.2 in Chapter 2). Threat of punishment can reduce defection from others without having to go through with punitive acts, and we show that social welfare in this regime is much higher than what traditional social punishment models suggest. We provide a comprehensive view of the outcomes of external factors, such as cost of signalling or the effectiveness of punishment, and we show that expensive signalling can still provide meaningful gains to cooperation when punishing others is costly.

Furthermore, we extend the proposed model of social punishment, introducing the signalling of institutional acts of punishment. Thus, whenever an individual is sanctioned by an institution external to the interactions, this act is advertised to their peers, deterring them from anti-social behaviour. We explore fearful defectors with varying levels of sensitivity to this signal, as well as changing the underlying networks of interaction. We find that cooperative outcomes can be increased dramatically when fearful defectors adequately respond to this signal, which also lessens the financial burden placed on sanctioning institutions, but that these findings are restricted to heterogeneous networks of interaction. Overall, our proposed models and analysis set a foundational basis in the study of signalling threat in the evolution of cooperation, showing that the fear of punishment can be a useful tool to improve cooperative outcomes and social welfare, serving as a deterrent to defection and retaliatory acts of punishment, in both peer and institutional settings.

5.2 Related Work

Punishment has been a major explanation for the evolution of cooperation in the context of the one-shot interaction [De Quervain et al., 2004; Fehr and Gachter, 2002; Powers et al., 2012]. A critical condition for cooperation to be sustainable in evolutionary models [Boyd et al., 2003; Hauert et al., 2007; Sigmund et al., 2001], as well as observable in lab experiments, requires the punishment to be cost-efficient, i.e. the effect it has on the wrongdoer should be sufficiently large compared the cost issued towards the punisher.

Signalling within and between organisms has been investigated using game theoretic models in areas of biology, economics and philosophy and it has been suggested that certain qualitative aspects are common to many real-world interactions [Huttegger et al., 2014]. Furthermore, it has been shown that signalling is a robust mechanism for promoting cooperative action in certain collective quorums [Pacheco et al., 2015]. In the presence of meaningless (no pre-defined meaning or behaviour) pre-play signals, cooperation has been shown to emerge as a result of individuals learning to discriminate between different signals and reacting accordingly [Santos et al., 2011]. Pre-play signalling has also been studied in the context of the evolution of honest signalling, showing that honest signalling only emerges when signalling is costly [Catteeuw et al., 2013]. To the best of our knowledge, no work so far has studied how signalling theory could explain the prevalence of social punishment by advertising acts of punishment after the fact.

Reputation has been suggested as an approach towards addressing this puzzle [dos Santos et al., 2011; Raihani and Bshary, 2015], whereas agents' actions consolidate in the eyes of observers some assumption of future behaviour. In this manner, social punishers can benefit indirectly through maintaining a reputation of punishing unjust behaviour [Hilbe and Traulsen, 2012]. However, the assumption that agents' actions are not anonymous proves unrealistic in many social contexts or application domains [Sigmund, 2010], i.e. in very large societies or when observation is difficult.

The simple presence of an audience has been shown experimentally to increase human propensity for moralistic punishment, causing an increase in costly punishment as a response to perceived moral violations [Kurzban et al., 2007]. Participants did not expect to encounter audience members again and the results hold for anonymous interactions or when the only observer was the experimenter. This suggests that there is at least some type of benefit to increasing the observability of one's willingness to punish, beyond reputation. Participants were generally not self-aware of the reasons for which they decided to punish, so in the context of self-organised societies [Boes and Migeon, 2017], this would explain why some individuals act towards the interests of society as a whole, irrespective of their intentions to do so [Stewart, 2017]. To this end, we are further motivated to study inherent normative mechanisms that have developed as a result of indirect evolutionary advantages.

Survey data on contribution norms in homogeneous and heterogeneous groups has demonstrated that uninvolved individuals hold well defined, yet conflicting normative views on equality, equity and efficiency [Reuben and Riedl, 2013]. That being the case, it has also been shown experimentally that punishment can help groups overcome this collective action problem, through the emergence of strong and stable contribution norms [Reuben and Riedl, 2013]. With regard to self-organised systems, punishment may help self-organising agents come to collective agreements on normative standards for efficiency, equity and equality.

Institutional interventions and interference have often been suggested as one effective mechanism to enforce cooperation, either by rewarding pro-sociality or sanctioning anti-social behaviour. Positive and negative institutional incentives have both been discussed in the literature, as well as combinations of the two, showing, for instance, how an adaptive approach to external incentives can be an effective way of promoting cooperation [Chen et al., 2015; Duong and Han, 2021b]. For a detailed reminder of these works, please see Section 2.5 in Chapter 2.

Finally, punishment and sanctioning have been studied extensively in the multi-agent system (MAS) literature, see e.g. [Balke and Villatoro, 2011; Villatoro et al., 2011]. Differently from our work, these studies aim at using the cooperation enforcing power of the mechanism for the purpose of regulating individual and collective behaviour, formalizing different relevant aspects of these mechanisms (such as norms and conventions) in a MAS.

Moreover, to the best of our knowledge, no work exists in the literature that analyses how sending costly threat of punishment can improve cooperation. As we show later, this mechanism can significantly enhance cooperation even when punishment is not highly cost-efficient.

5.3 Models and Methods

We adapt the Prisoner's Dilemma (PD), first by integrating the option of costly punishment as a benchmark, followed by describing the main models, and the different configurations which we explore using replicator dynamics and simulations. By choosing the most competitive social dilemma [Hofbauer and Sigmund, 1998], we explore the toughest environment for the emergence of cooperation, therefore increasing the relevance of any observed effects.

5.3.1 Prisoner's Dilemma (PD)

For ease of comparison, we recall that the one-shot PD is characterised by the following payoff matrix (for row player):

$$\begin{array}{c c} & \begin{array}{cc} \mathsf{C} & \mathsf{D} \end{array} \\ \begin{array}{c} \mathsf{C} \\ \mathsf{D} \end{array} & \begin{pmatrix} R & S \\ T & P \end{pmatrix} \end{array}.$$

Players experience, in pairs, a cooperation dilemma. In an interaction, individuals can decide whether to cooperate (play C) or defect (play D). Mutual cooperation (mutual defection) yields the reward R (penalty P), whereas unilateral defection provides a defector with the temptation T and the cooperator with the sucker's payoff S ($T > R > P > S$) [Sigmund, 2010]. The game is considered one-shot, in other words there is no memory of past actions or prior knowledge about the interactions. We note that an act of cooperation, i.e. *playing* C is different from a player adopting the strategy C. For the latter, a player will always play C and this is likewise true for acts of defection and D players. In all our experiments, we set $T = 2,\ R = 1,\ P = 0,$

and $S = -1$, a standard choice of a hard PD for cooperative behaviour to evolve [Han et al., 2013; Nowak, 2006a].

5.3.2 Social Punishment

We extend the PD by allowing a special type of C player the option of costly punishment, thereby becoming a punisher (P). After the normal interaction has taken place, a P player chooses to punish those opponents who played D during the interaction. A punishment act consists in paying a cost p to make their opponents incur a penalty q. Contrary to previous work that focuses mostly on efficient punishment [Boyd et al., 2003; Rand et al., 2010], we include the case where $p > q$, in order to better understand whether and when highly costly or inefficient punishment can still act as a promoting mechanism of cooperation. The newly defined P strategy always cooperates with C (as well as other P) players and always punishes D players. By including this strategy, we can analyse the evolutionary dynamics of punishment strategies and their viability in the evolution of cooperation. The 3×3 payoff matrix for the strategies P, D and C (for row player), is given by:

$$\begin{array}{c|ccc} & \mathsf{P} & \mathsf{D} & \mathsf{C} \\ \hline \mathsf{P} & R & S-p & R \\ \mathsf{D} & T-q & P & T \\ \mathsf{C} & R & S & R \end{array}$$

Next, we extend standard social punishment by introducing the signalling of an act of punishment and responding to such signals. Firstly, we consider a new type of punisher (denoted by PT) who, upon punishing a defector, can advertise this act by paying a cost θ, thereby alerting future opponents to their willingness to punish (and to the consequences of defecting against them). As such, a new type of defector arises (denoted by DT), who, once receiving the threat of punishment, will react and thus cooperate with the signalling punishers (to avoid punishment). PTs cooperate between each other, whereas DTs defect against each other, in similar fashion to P and D players. For infinite population size we can derive the 4×4 payoff matrix for

PT, D, DT and C (for row player) as follows:

$$\begin{array}{c} \\ \text{PT} \\ \text{D} \\ \text{DT} \\ \text{C} \end{array} \begin{array}{c} \begin{array}{cccc} \text{PT} & \text{D} & \text{DT} & \text{C} \end{array} \\ \begin{pmatrix} R & S-p-\theta & R & R \\ T-q & P & P & T \\ R & P & P & T \\ R & S & S & R \end{pmatrix} \end{array}$$

In order to derive the payoff matrix for infinite populations, notice that we can disregard the initial encounter between a PT player and either type of defector. Given some probability dependent on the composition of the population, the PT player can enact a punishment upon a DT player. We explain this interaction in-depth in Section 5.3.4 and provide average payoffs in the case of finite populations (the above payoff matrix for infinite populations can then be recovered at the limit of increasing the population size to infinity). As this population is infinitely large, the infinitesimally small initial interaction can be safely forgone for the sake of simplification.

5.3.3 Institutional Punishment

Departing from the scenario described above, in which punishment is enacted by individuals participating in the interactions, we also explore the setting of institutional punishment. Here, we consider a sanctioning institution which is exogenous to the interactions. In other words, punishment is delegated to an external agent, which punishes any act of defection. In this case the external institution pays a cost p to decrease the payoffs of defectors by an amount q. The one-shot PD, with institutional punishment, is characterised by the following payoff matrix:

$$\begin{array}{c} \\ \text{C} \\ \text{D} \end{array} \begin{array}{c} \begin{array}{cc} \text{C} & \text{D} \end{array} \\ \begin{pmatrix} R & S \\ T-q & P-q \end{pmatrix}. \end{array}$$

Moreover, we also consider that this institution signals acts of punishment, and DT players who observe this punishment switch to cooperation, as described above. In this setting, we study two types of observations: *direct* (i.e. first-hand) and *indirect* (i.e. second-hand) experience. If we assume first-hand experience of the signal, then only DT strategists directly involved in an interaction will observe the signal. Thus, DT strategists will always experience punishment in their first interaction with another strategist, regardless of their strategy. On the other hand, if DT strategists can observe interactions in their vicinity, then it may be the case that they observe others being punished before having the opportunity to defect. If such a signal has been detected, with some sensitivity ρ (i.e. the fear of being punished), the DT player will instead behave as a cooperator for the remainder of the generation. The sensitivity ρ ($0 \leq \rho \leq 1$) denotes the probability that DT players will switch to cooperation after witnessing an act of punishment in their vicinity. We assume a random order of interactions for each generation, and determine this order prior to calculating payoffs in the interactions. We note that we do not consider the cost of signalling in our analysis, as this does not affect evolutionary dynamics and can be included *post hoc*, but mention this in our findings when appropriate.

5.3.4 Methods

All the analyses and numerical results in this chapter are obtained using evolutionary game theoretic methods, using replicator dynamics for infinite populations [Hofbauer and Sigmund, 1998] (social punishment) and agent-based simulation for finite populations [Nowak et al., 2004; Sigmund, 2010]. For a detailed description, please refer to Section 2.2 in Chapter 2. In this setting, the payoff for each agent represents their fitness or social success. Evolutionary dynamics are then shaped by social learning [Hofbauer and Sigmund, 1998; Sigmund, 2010], whereby the most successful individuals tend to be imitated more often by others.

We consider a population of agents or individuals distributed in a network of contacts. Among these, we study well-mixed (WM) populations, square lattices (SL), as well as two types of scale-free (SF) networks, the Barabási and Albert (BA) [Barabási and Albert, 1999] and the Dorogovtsev-

Mendes-Samukhin (DMS) [Dorogovtsev et al., 2000] models (for a detailed description please see Section 2.4 in Chapter 2). We study social punishment on WM populations, to check whether evolutionary dynamics allow for the existence of fearful defectors. In our analysis of institutional punishment, we include all the network topologies as this setting is more closely related to the aims of this thesis. In the future, we aim to include a detailed analysis of social punishment in structured populations, as well.

Social Punishment

Replicator dynamics are used to study the growth of each fraction (of strategies) in the population, as a function of their frequency and relative fitness, where the fitness in this case corresponds to their payoffs [Hofbauer and Sigmund, 1998; Sigmund, 2010]. Considering a three-strategy game with PT, D and DT, we denote x_{PT}, x_D and x_{DT} the fraction of each strategy, respectively. Therefore, $x_{PT} + x_D + x_{DT} = 1$. The average payoff ($\Pi$) for each strategy reads:

$$\begin{aligned} \Pi_{PT} &= (1 - x_D)R + x_D(S - p - \theta), \\ \Pi_D &= x_{PT}(T - q) + (1 - x_{PT})P, \\ \Pi_{DT} &= x_{PT}R + (1 - x_{PT})P. \end{aligned} \tag{5.1}$$

In order to calculate the relative fitness, we determine the average fitness ($\bar{\Pi}$) in the population:

$$\bar{\Pi} = x_P \Pi_{PT} + x_D \Pi_D + x_{DT} \Pi_{DT}. \tag{5.2}$$

We can then calculate the gradients of selection for each strategy with the replicator equations:

$$\begin{aligned} \dot{x}_{PT} &= x_{PT}(\Pi_{PT} - \bar{\Pi}), \\ \dot{x}_D &= x_D(\Pi_D - \bar{\Pi}), \\ \dot{x}_{DT} &= x_{DT}(\Pi_{DT} - \bar{\Pi}). \end{aligned} \tag{5.3}$$

According to replicator dynamics, whenever a gradient is positive (i.e. $\dot{x} > 0$), the frequency of that particular strategy grows in the population. We can similarly describe the replicator dynamics for the case of three strategies P, D and C and and that of four strategies PT, D, DT and C. We use replicator dynamics to analyse homogeneous populations (i.e. well-mixed networks) with and without social punishment.

Moreover, we also study the case of finitely-sized homogeneous populations (with and without social punishment), where we can simplify the model using a statistical average of the payoffs for each conditional strategy, as opposed to carrying out the random ordering of interactions at the start of each game. The average payoffs for social punishment without threat remain the same as above.

To derive the average payoffs, we consider two distinct sequences of events, for each agent acting out a conditional strategy (PT and DT). Firstly, we consider the case when one PT player encounters a D player at the start of the generation. In this instance, the PT will punish the D player, while also incurring the cost θ for signalling the act of punishment. Each subsequent interaction with DT players will result in a reward for both the players, as DTs will react to the signal and avoid defecting against that PT player. Conversely, we consider how the payoffs change when a PT player encounters a DT player as the first defection against that PT in that generation. As the PTs signal is unbeknownst to the DT, it will defect. In turn, the PT will carry out their act of punishment, causing both players to miss the opportunity of cooperating.

The probability of either sequence happening first is dependent on the composition of the population at each first interaction for PTs. The payoffs for all other strategies remain unaffected. Let n_1, n_2 and n_3 denote the numbers of PT, D and DT players in the population, respectively. We have $n_1 + n_2 + n_3 = N$. We denote $\Pi_{A,B}$ the payoff received by a player following the strategy A when facing players following strategy B (some payoffs are equivalent e.g. $\Pi_{C,C} = \Pi_{PT,C} = \Pi_{C,PT} = \Pi_{PT,PT} = R$). The average payoffs for PT, D and DT read:

$$\begin{aligned}
\Pi_{PT} &= \frac{1}{N-1}\left((n_1+n_3-1)\Pi_{C,C} + n_2\Pi_{PT,D} + \frac{n_3(\Pi_{PT,D}-\Pi_{C,C})}{n_2+n_3}\right),\\
\Pi_{D} &= \frac{1}{N-1}\Big(n_1\Pi_{D,PT} + (n_2+n_3-1)\Pi_{D,D}\Big),\\
\Pi_{DT} &= \frac{1}{N-1}\left(n_1\Pi_{C,C} + \frac{n_3(\Pi_{D,PT}-\Pi_{C,C})}{n_2+n_3} + (n_2+n_3-1)\Pi_{D,D}\right).
\end{aligned}$$

Note that at the limit of infinite population size, $N \to \infty$, $x_P = \frac{n_1}{N}$, $x_D = \frac{n_2}{N}$, $x_{DT} = \frac{n_3}{N}$, we recover the equations (5.1) for infinite population sizes.

For the four-strategy game (PT, D, DT, C), we introduce n_4 as the number of C players. Therefore, we have $n_1 + n_2 + n_3 + n_4 = N$. The payoffs for PT, D, DT and C then become:

$$\begin{aligned}
\Pi_{PT} &= \frac{1}{N-1}\left((n_1 + n_3 + n_4 - 1)\Pi_{C,C} + n_2\Pi_{PT,D} + \frac{n_3(\Pi_{PT,D} - \Pi_{C,C})}{n_2 + n_3}\right),\\
\Pi_D &= \frac{1}{N-1}\left(n_1\Pi_{D,PT} + (n_2 + n_3 - 1)\Pi_{D,D} + n_4\Pi_{D,C}\right),\\
\Pi_{DT} &= \frac{1}{N-1}\left(n_1\Pi_{C,C} + \frac{n_3(\Pi_{D,PT} - \Pi_{C,C})}{n_2 + n_3} + (n_2 + n_3 - 1)\Pi_{D,D} + n_4\Pi_{D,C}\right),\\
\Pi_C &= \frac{1}{N-1}\left((n_1 + n_4 - 1)\Pi_{C,C} + (n_2 + n_3)\Pi_{C,D}\right).
\end{aligned}$$

Note that the payoffs for unconditional strategies are never affected by the ordering of interactions. Analytically, the payoff for punishers who threaten depends on the number of defectors who respond to threatening signals in the population, specifically the ratio between the two different defecting strategies (i.e. $\frac{n_3}{n_2+n_3}$). We contend that this could remain true in practical scenarios. In other words, it is better for signalling punishers when future defectors can discriminate signals precisely, and this indirectly increases the payoff for sensitive defectors, as well. This would suggest that there is a synergistic relationship between signalling punishers and fearful defectors and, to an extent, neither would prevail without the other.

Institutional Punishment

Using agent-based simulations, we study institutional punishment with and without signalling threat using all the network topologies described above. Unlike the previous scenarios, we cannot simplify the average payoffs, as the ordering of the interactions is particularly important in the context of networks of contacts. In Algorithm 1, we describe the process of determining a random order of interactions with direct experience (as introduced in Section 5.3.3).

Algorithm 1 Randomise order of interactions with *direct* observations

```
Require: Initialize a randomly ordered list of all DT agents
for agent in agents do
    $agent.partner \leftarrow 0$
end for
for agent in list do
    if $agent.partner = 0$ then
        $other \leftarrow$ random adjacent agent
        $agent.partner \leftarrow other.id$
        if $other.partner = 0$ then
            $other.partner \leftarrow agent.id$
        end if
    end if
end for
```

After executing Algorithm 1, each DT player will have an assigned partner that determines the first individual they have interacted with. The signal has no effect on C (unconditional cooperator) or D (unconditional defector) strategists. The process of extracting the payoff for each agent (we assume $\rho = 1$ for simplicity of representation) is described in Algorithm 2. Note that we assume DT players will cooperate unless they have not witnessed an act of punishment when calculating payoffs, as they will defect at most once per generation. In other words, DTs are assumed to cooperate with the exception of the first interaction they have participated in.

Algorithm 2 Calculate payoff of one agent with *direct* observations

```
Require: random $agent$ and $list$ of adjacent agents
for $neighbor$ in $list$ do
    $x \leftarrow agent.strategy$
    $y \leftarrow neighbor.strategy$
    if $x = DT$ then
        if $agent.partner = neighbor.id$ then
            $x \leftarrow D$
        end if
    end if
    if $y = DT$ then
        if $neighbor.partner = agent.id$ then
            $y \leftarrow D$
        end if
    end if
    $agent.fitness \leftarrow agent.fitness + \Pi_{x,y}$
end for
```

For robustness, we also study the case (as introduced in Section 5.3.3) where DT strategists can observe the threat of punishment after any interaction in their neighbourhood. In Algorithm 3, we describe the process of determining a random order of interactions with second-hand experience.

Algorithm 3 Randomise order of interactions with *indirect* observations

```
Require: Initialize a randomly ordered $list$ of all D and DT agents
for $agent$ in $agents$ do
    $agent.marker \leftarrow 0$
end for
for $agent$ in $list$ do
    if $agent.marker = 0$ then
        $agent.marker \leftarrow 2$
        $other \leftarrow$ random adjacent agent
        $temp \leftarrow other.marker$
        for $neighbor$ in adjacent agents of $agent$ do
            if $neighbor.strategy \neq C$ then
                $neighbor.marker \leftarrow 1$
            end if
        end for
        if $temp = 0$ then
            $other.marker \leftarrow 2$
        end if
        if $other.strategy = D \vee (temp = 0 \wedge other.strategy = DT)$ then
            for $neighbor$ in adjacent agents of $other$ do
                if $neighbor.strategy \neq C \wedge neighbor.marker \neq 2$ then
                    $neighbor.marker \leftarrow 1$
                end if
            end for
        end if
        if $agent.strategy = DT \wedge other.strategy = DT \wedge temp = 1$ then
            $agent.marker = 3$
            $other.marker = 4$
        end if
    end if
end for
```

After executing Algorithm 3, each DT agent will have had a marker assigned to them. These marker values denote whether a DT player has seen the signal before interacting with another agent, or, if they have been sanctioned, they will have a marker that coincides with the marker assigned to an adjacent agent. If two DT strategists interact, but one of them has already witnessed an interaction in the past, then two special values for the markers account for this exceptional case, to recall which agent acted as the defector. Thus, in Algorithm 4 we describe the process used to calculate the payoff of an agent after determining this sequence of interactions.

Algorithm 4 Calculate payoff of one agent with *indirect* observations

```
Require: random $agent$ and $list$ of adjacent agents
for $neighbor$ in $list$ do
    $x \leftarrow agent.strategy$
    $y \leftarrow neighbor.strategy$
    if $agent.marker = 2 \wedge neighbor.marker = 2$ then
        if $x = DT$ then
            $x \leftarrow D$
        end if
        if $y = DT$ then
            $y \leftarrow D$
        end if
    end if
    if $agent.marker + neighbor.marker = 7$ then
        if $agent.marker = 3$ then
            $x \leftarrow 2$
        else
            $y \leftarrow 2$
        end if
    end if
    $agent.fitness \leftarrow agent.fitness + \Pi_{x,y}$
end for
```

We note that direct observations almost always imply more acts of defection, as each DT agent will have had to be punished before observing the signal. If we consider indirect observation, then DT players have potential chances of observing their neighbours being punished before they adjust their behaviour. In Figure 5.1, we depict typical cases of this ordering, showing that indirect observations can avoid many cases of initial punishment.

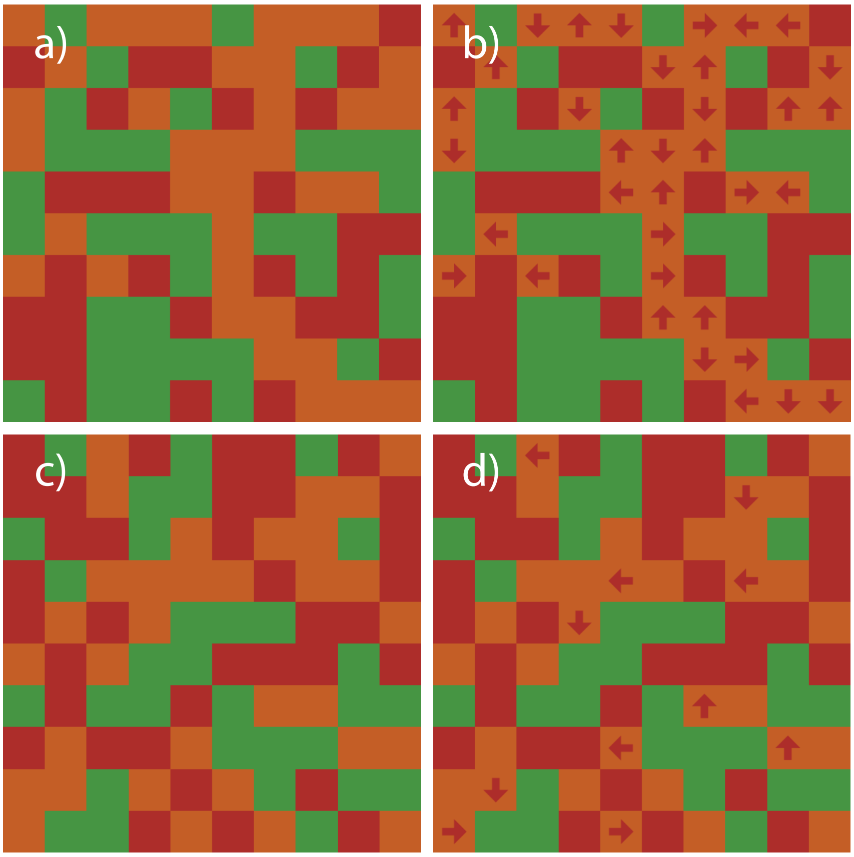

Fig. 5.1 **Indirect observations can enable fearful defectors to avoid punishment.** Panels show populations on 10×10 square lattices before (a) and after (b) direct observations and similarly (c and d) for indirect observations. Patch colours are green for C, red for D and orange for DT. Arrows depict which adjacent agent the DT players would defect against before observing the threat of punishment, as determined by algorithms 1 and 3.

Agent-Based Simulations

For our simulations, we adopt a population size $N = 100$ for WM, $N = 900$ ($L = 30$) for SL, and $N = 1000$ for either type of SF network. At the beginning of the game, each agent is randomly assigned a strategy from all the available strategies for that experiment. At each time step (generation), each agent plays the PD with every other agent in their immediate neighbourhood. The fitness for each agent is the sum of their payoffs from each interaction.

Social learning is modelled using the pairwise comparison rule [Traulsen et al., 2006], a standard approach in studying evolutionary dynamics in evolutionary game theory, which states that a player A with fitness f_A can imitate another player B with fitness f_B with a probability given by the Fermi function, i.e. $P_{A,B} = (1 + e^{-\beta(f_B - f_A)})^{-1}$, where β represents the intensity of selection, i.e. how strongly the agents value the difference in fitness between them and their opponents (for a more detailed discussion, see again Section 2.3 in Chapter 2). For $\beta = 0$, we obtain neutral drift (random decisions), whereas large β values lead to increasingly deterministic imitation. We assume at most one imitation can happen per generation (asymmetric update).

In the absence of exploration or mutations, evolution inevitably leads to monomorphic states. Once such a state has been reached, it cannot be escaped solely through imitation. With a given probability μ, the process of imitation is replaced instead by a randomly occurring mutation. A mutation is equivalent to behavioural exploration, where the individual makes a stochastic decision to switch to one of the other available strategies. In line with previous works and lab experiments [Rand et al., 2013; Szabó and Fáth, 2007; Zisis et al., 2015], we set $\beta = 1$ and $\mu = 0.001$ in our simulations.

For social punishment, we simulate the evolutionary process for 10^4 generations and average our measurements over the final 10^3 steps for a clear and fair comparison (for example due to cyclic patterns). Furthermore, the results for each combination of parameters are obtained from averaging 500 independent realisations, with the exception of typical run patterns.

For institutional punishment experiments, which converge more slowly, we simulate the evolutionary process for 5×10^4 generations for WM, and 2.5×10^5 generations for SL and SF networks, and average our measurements over the final 10% generations for a clear and fair comparison. Moreover, we seed 10 different networks for each type of scale-free networks (BA and DMS), and for each combination of parameters, we obtain results from 20 individual replicates.

5.4 Results

We independently study social and institutional punishments, using the former to show that fearful defectors can emerge in the setting with peer punishment only. Following this, we go beyond peer (i.e. social) punishment, delegating the sanctions to an external institution, instead. The model with institutional punishment is then systematically explored on a variety of network topologies.

5.4.1 Social Punishment

We study the potential of punishment and signalling strategies and their effects on evolutionary dynamics using the three scenarios described in Section 5.3.4: no threat (P, D, C), threat without cooperators (PT, D, DT) and threat with freely available strategies (PT, D, DT, C).

Replicator dynamics for infinite populations

Following our analysis on infinite populations, we find that introducing threat signalling introduces a type of beneficial dynamic which fosters cooperation. The relationship between signalling punishers and defectors remains very similar to the one found in standard social punishment models (See Figure 5.2a and 5.2c). Increasing the efficiency ratio of punishment (i.e. q/p) is the only way in which social punishment can remain relevant and this can often lead to undesirable consequences. On the other hand, the dynamic created between PTs and DTs naturally promotes cooperation. DTs lose out to PTs, as DT players do not cooperate amongst each other, but they do better than their defecting brethren, by reaping the rewards of their reluctance to defect against PTs. Even when the fraction of cooperators

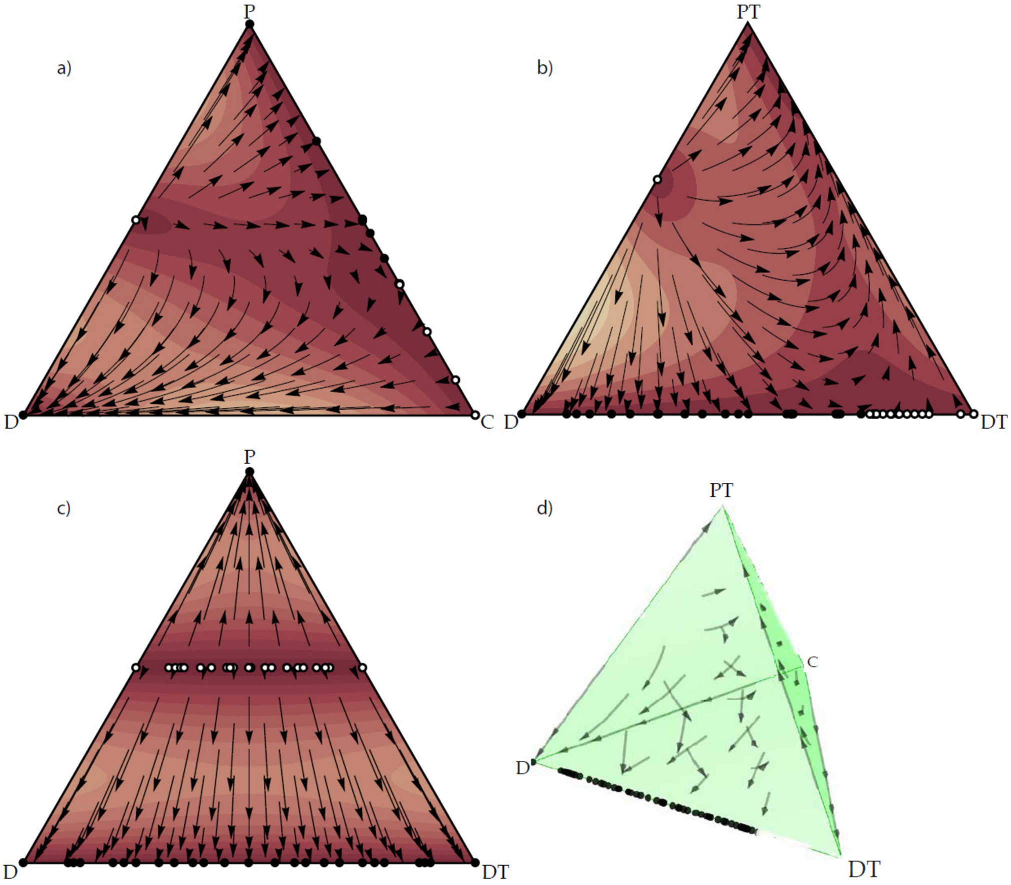

Fig. 5.2 **Replicator dynamics analysis of social punishment with and without threat.** Phase diagram comparison between standard social punishment (left column) and social punishment with threat (right column) using replicator dynamics. Vertices represent specific strategies, whereas solid and empty dots stand for stable and unstable rest points, respectively. The colours represent speed of motion under the dynamic (lighter is faster, darker is slower). Parameters: $p = 1$; $q = 3$; $\theta = 1$.

$(x_P + x_D)$ becomes very low, the existence of DTs catalyses the conversion towards cooperation (see Figure 5.2b). By increasing the efficiency in the case of threat signalling, we found that the range of compositions which lead to all defectors is reduced even further. The ratio q/p also favours DTs over Ds, which can provide another avenue towards cooperation that does not exist in the absence of signalling.

In Figure 5.2d, we show that the results remain robust when we introduce C players, and in fact that the model is more resilient to compositions with high x_C compared to traditional social punishment models, in which C outperforms P (when p is high).

Analytically, we confirm the results from our replicator dynamics analysis by computing the rest points for each system of equations. For P, D, C, we found the following stable rest point on the PC edge which we can see mirrored in Figure 5.2a (at $x_P = \frac{1}{2}$):

$$\{x_P = \frac{P-S+p}{R-S-T+P+p+q},\ x_D = 1 - x_P,\ x_C = 0\}.$$

Studying the case of PT, D, DT hints at the interesting dynamics seen in Figure 5.2b. We observe the following edge rest points:

$$\begin{aligned}
&\{x_{PT} = \frac{P-S+p+\theta}{R-S-T+P+p+q+\theta},\\
&\qquad x_D = \frac{R-T+q}{R-S-T+P+p+q+\theta}, x_{DT} = 0\},\\
&\{x_{PT} = 0,\ x_D = \frac{R-P}{R-S+p+\theta},\ x_{DT} = \frac{P-S+p+\theta}{R-S+p+\theta}\},\\
&\{x_{PT} = 0,\ x_{DT} = 1 - x_D\}.
\end{aligned}$$

We add that for the parameter values seen in Figure 5.2, we can observe a slight evolutionary advantage of PT against standard defectors, compared to traditional social punishers. For the same values of p and q, while also having to pay a cost of signalling $\theta = p$, we observe the points $x_P = \frac{1}{2}$ without signalling and $x_{PT} = \frac{3}{5}$ after signalling. This shows that signalling social punishers have higher chances of survival against wrongdoers, even if the aggregated costs they expend towards punishment are larger than the ones paid by standard punishers. Note that we only discuss edge rest points for

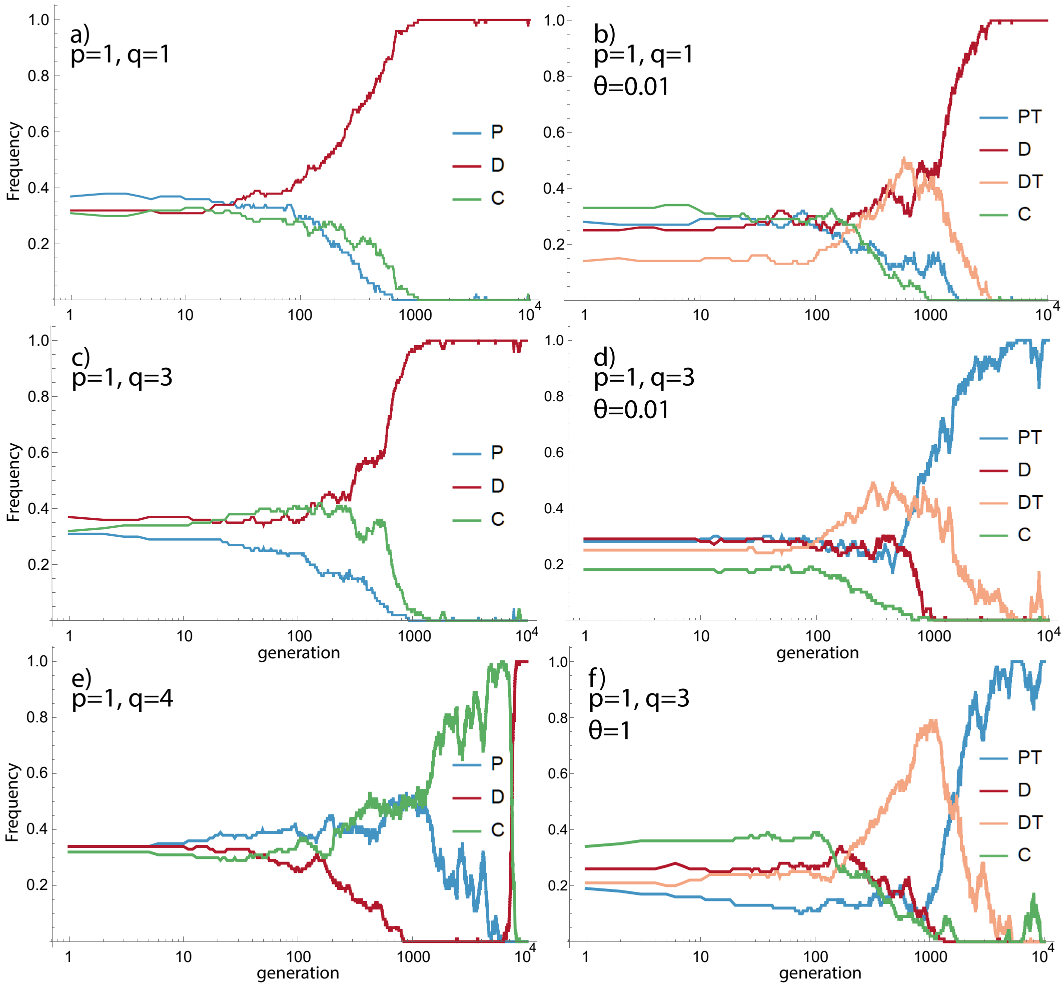

Fig. 5.3 **Time evolution of strategies before and after signalling the threat of punishment.** Typical evolution of frequencies over time before (left column) and after (right column) introducing signalling of threat (social punishment). Note that we only address the typical cases for clear comparisons.

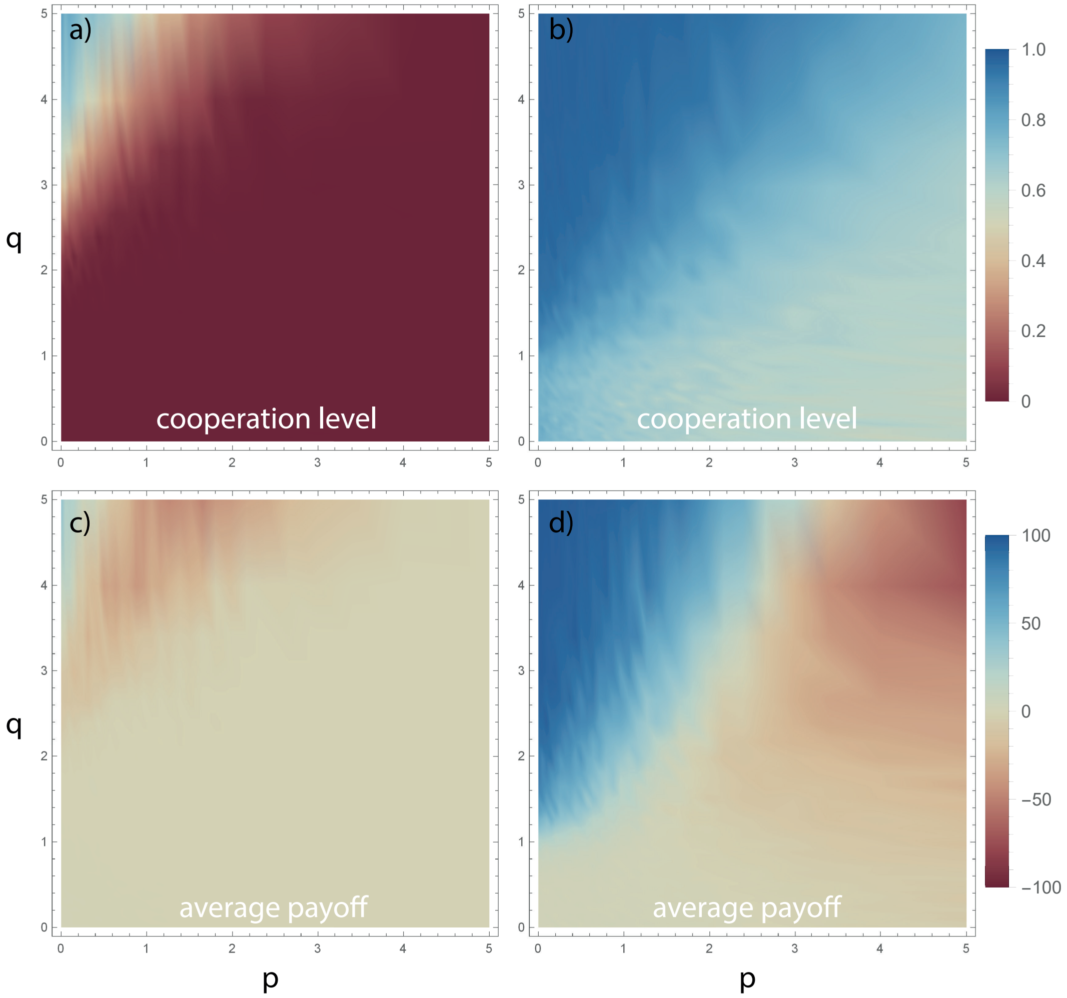

Fig. 5.4 **Effects of the efficiency ratio of punishment on the frequency of cooperation** (punishers and cooperators alike) and average population payoff, before (left column) and after (right column) introducing inexpensive signalling (social punishment). Parameters: $\theta = 0.01$.

clear and concise presentation. Vertex rest points do exist in most scenarios, and they can be clearly seen in Figure 5.2.

Agent-based Simulations for Finite Populations

The initial results from the agent-based simulations confirm the trend we found using replicator equations (see Figure 5.2). For typical runs, DT outperforms D which leads to the rapid extinction of D players and invites a booming growth of PTs (see Figure 5.3). When q/p is low enough, we observed some exceptions to the norm, in which the initial conditions led to a population of predominantly DTs, which allowed PTs to flourish which, in turn, led to further defection. This type of cyclic behaviour only happened when punishing an act of defection and/or signalling it were extremely costly. We note that the results are robust even for costly signalling.

When signalling is not costly, we found that cooperation is greatly increased across virtually all ranges of p and q. In the case of simple social punishment, punishment is only effective at increasing cooperation when it is also efficient ($q/p > 3$). On the other hand, even when punishment is very inefficient, we find a high propensity for pro-social behaviour in the presence of threat. Even when punishing is extremely costly ($q/p < 1$), we observe close to 50% cooperation. At moderate efficiency ($q/p > 1.5$), the frequency of cooperation is very high ($\approx 70\%$), comparable to highly efficient punishment without threat. We confirm the trends for higher values of θ (See Figures 5.3f and 5.5a), but also observe one interesting outcome of signalling that may suggest a direct benefit towards the evolutionary promotion of punishing behaviour. Even at much higher ratios of efficiency of punishment (q/p), it is usually cooperators who prevail over the population, as seen in Figure 5.3e. Conversely, fearful defectors only cooperate with punishers, in the case of signalling, which naturally promotes the evolutionary advantage of punishers (see Figure 5.3f). One valuable side effect of this phenomenon is that the system is more resilient to mutation and exploration, a large population of punishers is more able to deal with defectors when compared to a large population of cooperators, where a single defector acts akin to a wolf among lambs.

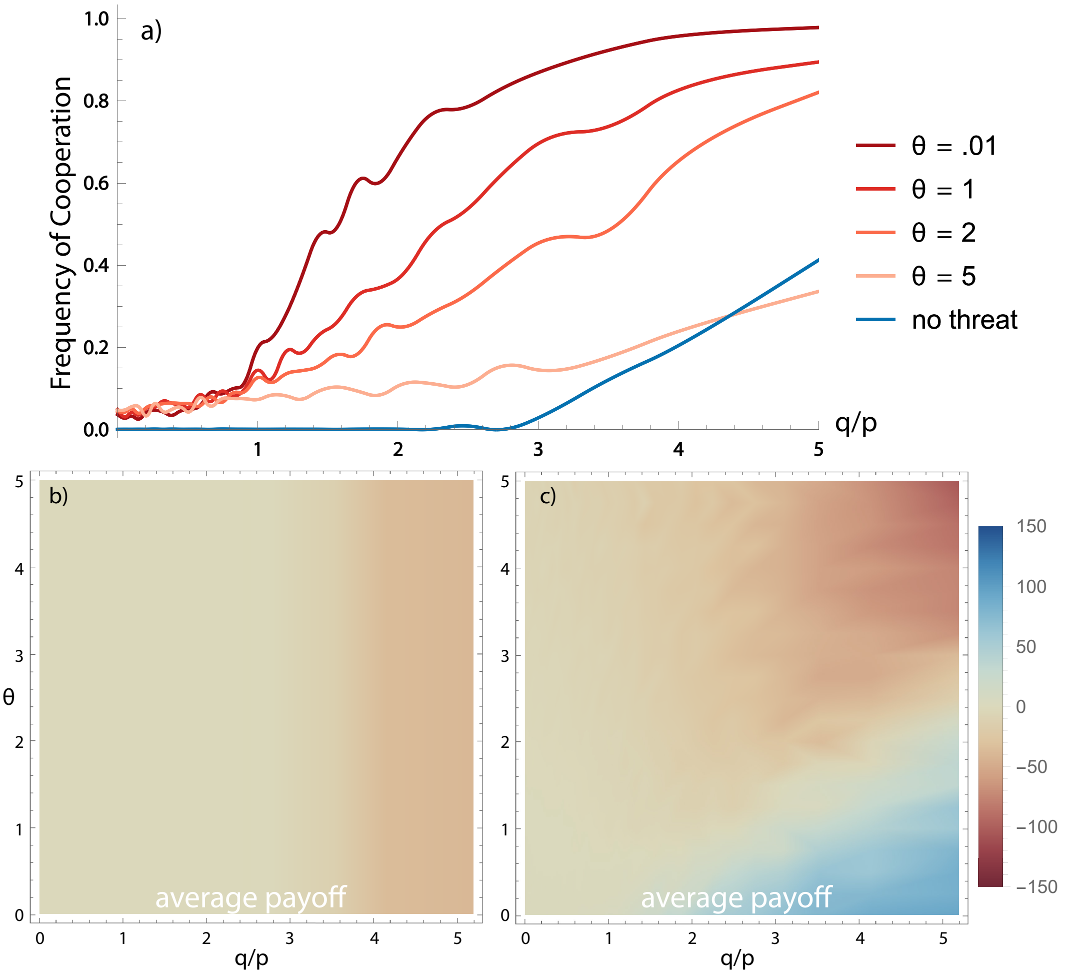

Fig. 5.5 **How costly signalling influences cooperation for (PT, D, DT, C) versus no threat (P, D, C)**. Panel a: frequency of all cooperative strategies for various values of θ and for the standard case. Panels b and c: average payoffs as a function of efficiency of punishment and cost of signalling for no threat (left) and threat (right). Parameters: $p = 1$.

Welfare in the population (i.e. average population payoff) was overall much higher in our extended social punishment model (see Figures 5.4c, 5.4d, 5.5b and 5.5c). Of note, when punishment was very expensive ($p > 2$), average payoff decreases dramatically in the extended model. Intuitively, this happens because PTs survive by cooperating with Cs and DTs, but also incur great costs in order to punish defectors and to deter DTs from following that trend. For higher values of q/p, we see another stark difference. With the exception of a very small region of values, welfare decreases ($\bar{\Pi} \approx -25$) as cooperation goes up, whereas the opposite is true for signalling ($\bar{\Pi} \approx 75$). The *power* difference between Ps and Ds proves damaging to social welfare - Ds do not cooperate, which causes Ps to lose fitness and in return, they pay a further cost, compounding the losses, causing Ds to incur even more loss. While this behaviour fosters cooperation, it greatly decreases social welfare. Inversely, a single act of punishment is enough to convert the entire population of DTs to cooperation, which is a qualifying factor in the growth of average payoff.

Our comprehensive study of the external factors under which cooperation emerges, in regards to efficiency of punishment and the cost of signalling shows that fear of punishment enhances cooperation for almost all configurations (with the notable exception of highly efficient punishment coupled with expensive signalling). The results suggest that the transparency of social punishment, specifically the awareness agents have regarding acts of retribution, coupled with the ease of advertising said acts, behaves as a fulcrum towards cooperation. Fear of punishment, therefore, is most effective when awareness of who is or is not a punisher is high. On the other hand, the more deleterious an act of punishment is, the more likely it becomes for standard costly punishment to lead towards satisfactory outcomes.

We also show that social welfare increases when signalling is not very costly, irrespective of the punishment efficiency. Interestingly, at higher θ values, social welfare lowers even as cooperation increases. We show that when punishment is effective, signalling can lead to lower levels of welfare. Intuitively, this suggests that advertising an ineffective act of punishment is productive even when signalling is expensive.

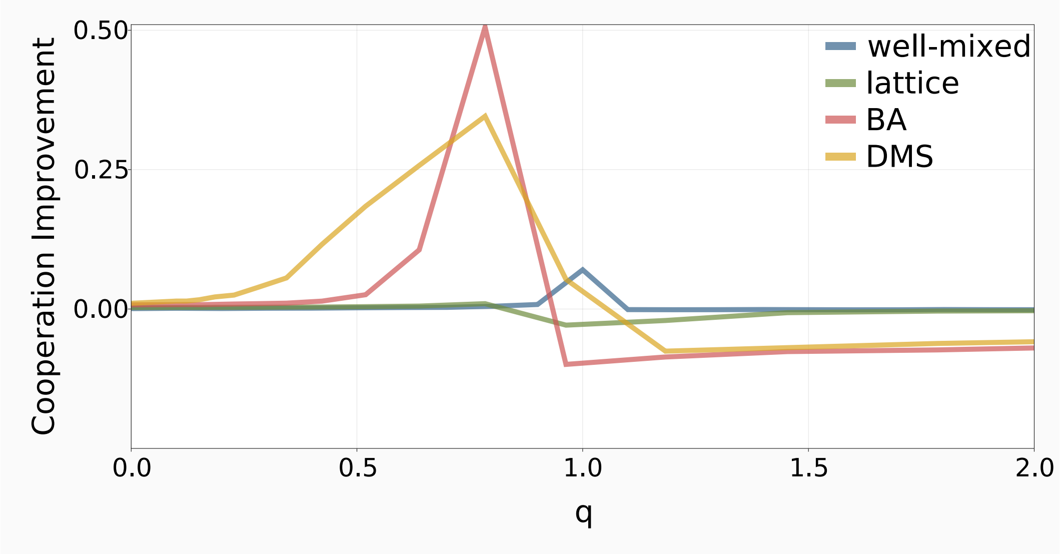

Fig. 5.6 **Improvement of cooperative acts relative to baselines in the presence of institutional threat with indirect observations.** Depicted are the differences between the fraction of cooperative acts in the presence of threat, and the fraction of cooperation in the baseline case, for different network structures. Parameters: $\rho = 1$.

5.4.2 Preliminary Results with Institutional Punishment

In the previous section, we have shown the potential for fearful defectors to emerge from social punishment. Here, we consider whether they can emerge under institutional punishment, and study the ramifications of their existence, both in the cooperative outcome and in the financial costs for the sanctioning institution.

Fear of institutional punishment improves cooperative outcomes in socially diverse populations

In the presence of social diversity, the threat of institutional punishment can significantly improve perceived acts of cooperation when punishment is not deleterious (i.e. small q). Indeed, for both lowly clustered and highly clustered scale-free networks (i.e. BA and DMS, respectively), there is a marked improvement in cooperative acts compared to the baselines (see Figure 5.6). In contrast, for intermediate and high sanctions, we observe a slight decrease in cooperative outcomes. In other words, when sanctions are not exceedingly injurious to defection, fearful defectors can pave the way for cooperation, and are better equipped to coexisting with unconditional defectors. If the strength of punishment is high enough to prevent the proliferation of unconditional defectors, fearful defectors can nevertheless coexist with cooperators. These findings suggest that DT strategists can better coexist with both D and C players in socially diverse (heterogeneous) settings.

In homogeneous populations (i.e. well-mixed or lattice networks), signalling the institutional threat of punishment does not significantly alter or improve cooperative outcomes compared to punishment by itself (see Figure 5.6). We observe a slight increase in cooperation for intermediary sanctions (when $q \approx R$) for WM populations, with a negligible decrease in cooperation for SL populations for the same parameter values. When sanctions are low, or if they are enough to ensure the success of cooperators, then we see no effect of signalling the threat of punishment in terms of cooperation, for homogeneous populations. We note that these findings hold true for direct observations, although the effects are diminished slightly (see Figure B.1 in Appendix B).

Signalling threat can reduce the financial burden on sanctioning institutions

In line with the previously discussed findings on cooperative outcomes, we observe that signalling the threat of institutional punishment can greatly reduce the financial burden placed on sanctioning institutions, especially in heterogeneous populations (see Figure 5.7). Social diversity and signalling the threat of punishment produce a synergistic effect, allowing fearful defectors to avoid excessive acts of punishment, thereby cooperating with one another, instead. In highly clustered scale-free networks, especially, we measure a significant reduction in total costs almost irrespective of the strength of punishment q.

Particularly, we suggest that further investigations are needed to determine whether signalling punitive acts in homogeneous populations would be an effective way to reduce sanctioning and improve social welfare. For instance, we found that varying the defectors' response to fear (i.e. the sensi-

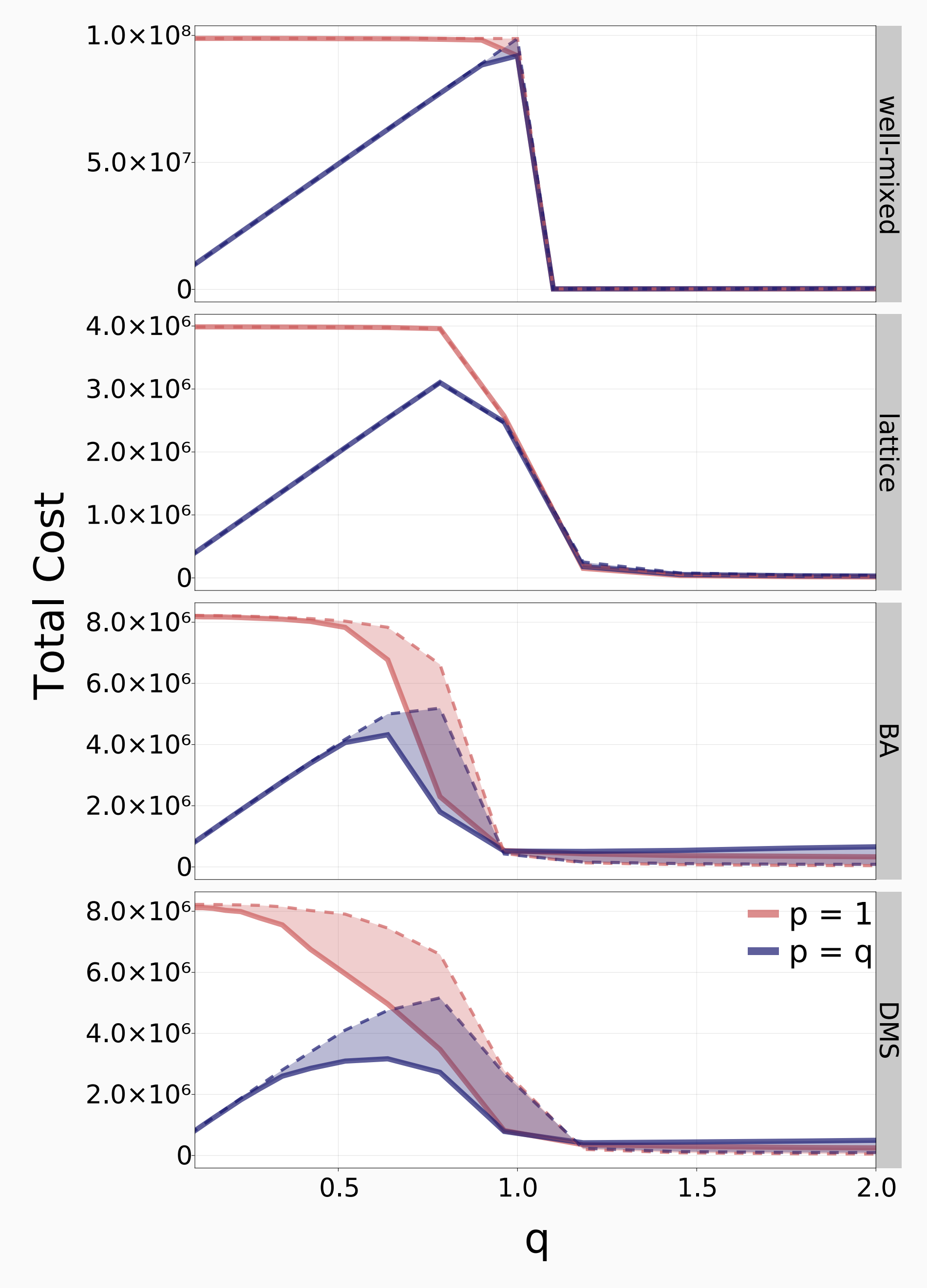

Fig. 5.7 **Threat of institutional punishment reduces the total cost of punishment.** Depicted are the accumulated costs of punishment, for different network types with indirect observations. The solid lines depict the accumulated costs when p is normalised to 1 and when it is equal to the sanction imposed on defection. Dashed lines depict this cost for the baseline case (in the absence of signalling). The coloured areas highlight the difference between these two lines. Parameters: $\rho = 1$.

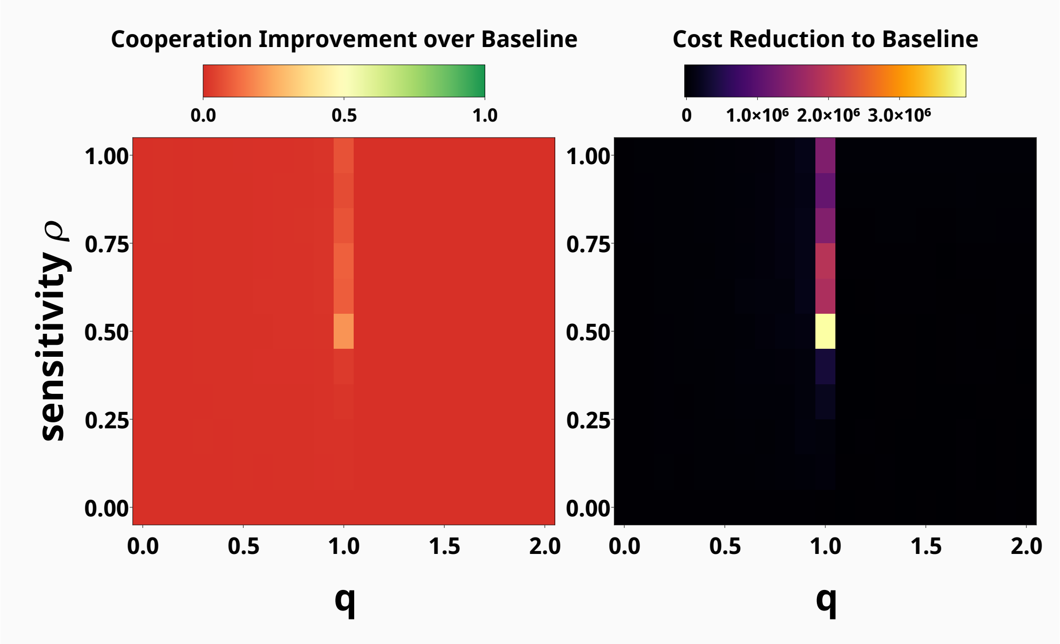

Fig. 5.8 **Fear of punishment in WM populations reduces the cost to institutions.** Panels show the improvement in cooperative acts and the reduction in accumulated cost, compared to the baseline, with varying levels of sensitivity to the signal (ρ), using indirect observations. Parameters: $p = 1$.

tivity to the signal ρ) can sometimes produce better cooperative outcomes and reduce costs in well-mixed populations (see Figure 5.8). In very specific settings, fearful defectors can pave the way towards cooperation even in homogeneous populations, but these effects are restricted to a very narrow parameter space (i.e. intermediate sanctions $q \approx R$).

In the absence of social diversity, DT players cannot adequately benefit from signalling the threat of punishment. We can further see this by studying the frequency of strategies in homogeneous populations (see Figure 5.9). When DT strategists fail to respond to the signal sufficiently (i.e. $\rho \to 0$), they are equivalent to D strategists. On the other hand, if they always respond to the signal (i.e. $\rho = 1$), then little differentiates them from unconditional cooperators. Thus, in homogeneous populations, fearful defectors can rarely exist as something other than one of the other two strategies, as the signal is almost always available to them. Moreover, spatiality by itself is not enough to synergise with signalling (i.e. in SL populations), and heterogeneity is

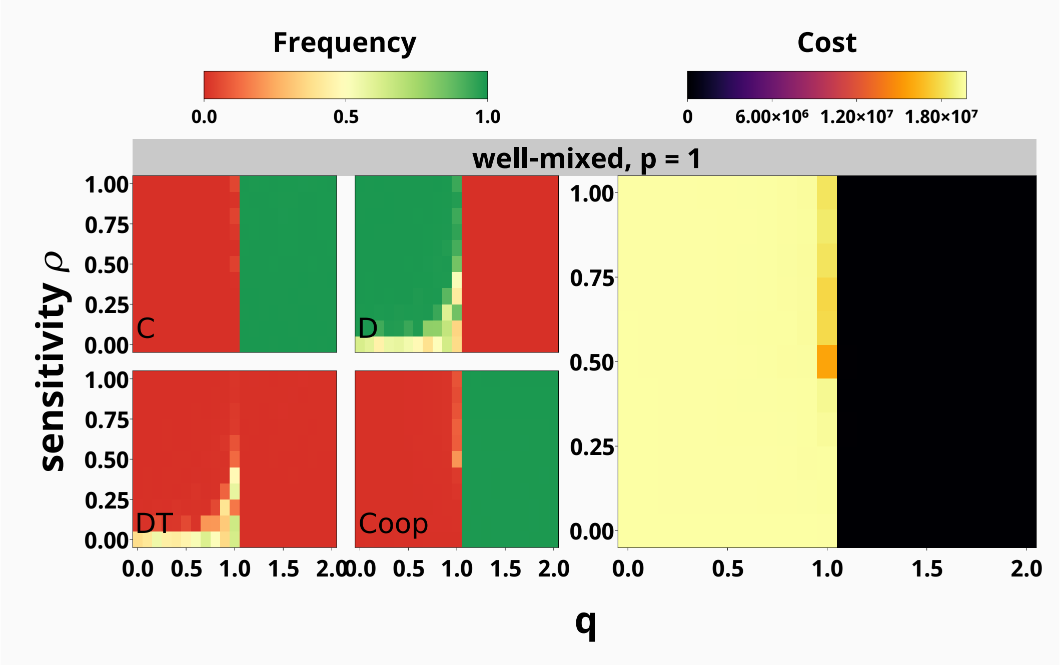

Fig. 5.9 **Fear of institutional punishment changes the outcome of evolutionary dynamics in well-mixed populations.** Heatmaps show the fraction of each strategy and overall cooperative acts, as well as the total accumulated cost when signalling the threat of institutional punishment with indirect observations. Parameters: $p = 1$.

needed to produce the marked benefits we observe by signalling the threat of punishment.

Indirect observations favour fearful defectors, who pave the way to cooperation

Indirect observations allow fearful defectors to coexist with cooperators by avoiding unnecessary punitive acts (see Figure 5.10). Previously (see again Figures 5.6 and 5.7), we suggested that the presence of these conditional defectors can pave the way towards cooperation, all the while serving to reduce unnecessary sanctioning.

Assuming instead that defectors would only respond to first-hand (i.e. direct) observations, then this typically leads to C strategists being more frequent in the population (see Figure 5.11). Interestingly, in such a case, the synergistic effects of the threat of punishment and social diversity are diminished (shown in detail in Figures B.1 and B.2 in Appendix B). While indirect

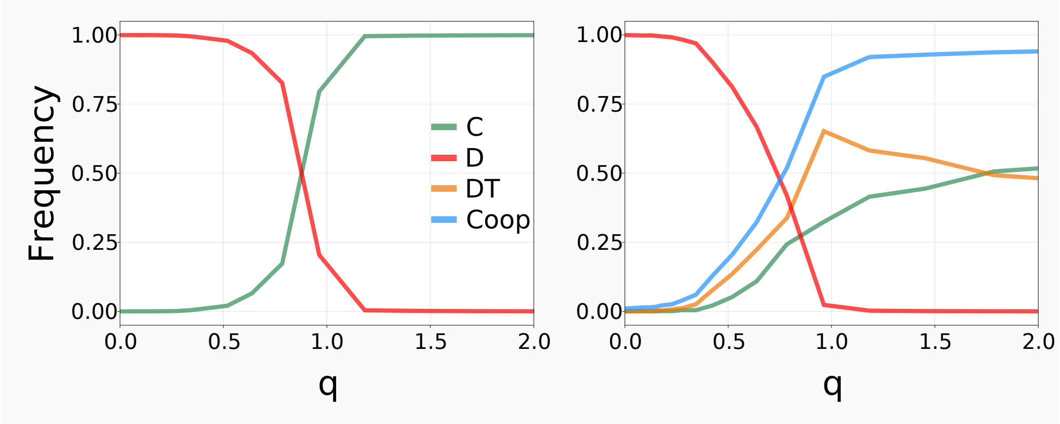

Fig. 5.10 **Indirect observations of the threat of punishment allow for the coexistence of fearful defectors and cooperators.** Panels show the frequency of each strategy in DMS populations, as well as the fraction of cooperative acts in the presence of threat. The left panel shows the baseline (i.e. in the absence of threat), while the right panel shows threat of punishment with indirect observations. Parameters: $\rho = 1$.

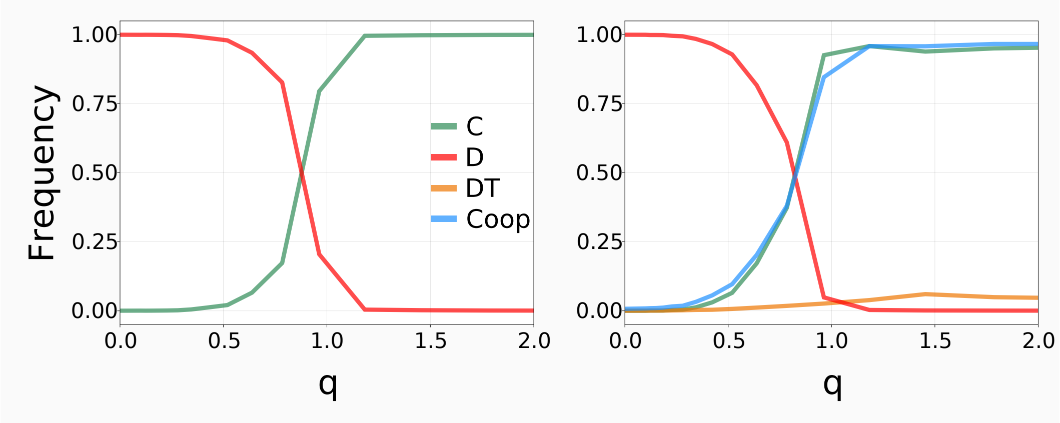

Fig. 5.11 **Direct observations of the threat of punishment promote unconditional cooperators.** Panels show the frequency of each strategy in DMS populations, as well as the fraction of cooperative acts in the presence of threat. The left panel shows the baseline (i.e. in the absence of threat), while the right panel shows threat of punishment with direct observations. Parameters: $\rho = 1$.

observations enable the coexistence of fearful defectors and cooperators, this would also suggest that if punitive acts were to cease, then pro-sociality would be at risk of collapsing. On the other hand, direct observations, which generally lead to a higher frequency of unconditional cooperation, are associated with a decreased improvement in overall cooperation and higher costs than second-hand experience of the signal. We note that these findings are robust across the whole range of network topologies (see Appendix B).

5.5 Discussion

Punishment used as a deterring mechanism to prevent further damaging actions against the punisher or their peers appears to be a commonly found behaviour in human society and even in some animal hierarchies [Clutton-Brock and Parker, 1995]. Much of recent literature has concluded, however, that punishment may have evolved for reasons other than the promotion of cooperation, because significant benefits to punishers could typically not be found in the context of game theory [Dreber et al., 2008; Rankin et al., 2009]. Indeed, it may be the case that even if punishing defectors incurs an immediate cost, it discourages observers from repeating said action, as long as the accumulated costs of punishment are outweighed by the additional acts of cooperation evoked over long runs [dos Santos et al., 2011]. Our models suggest that this does not only happen in repeated interactions and that punishment can evolve through advertising the acts of punishment.

We show that signalling acts as a catalyst for the emergence of cooperation when defectors are fearful of the punishers who advertise themselves as such. Furthermore, we argue that exhibiting deeds of punishment can explain the success of punishers, when traditional social punishment mechanisms would otherwise fail due to external factors, such as lowly efficient acts of punishment. Indeed, it seems to be the case that fearing punishment can discourage future defectors even more than the evolutionary dynamics associated with inexpensive, deleterious deeds of retribution. Moreover, we show how the traditionally damaging effects of social punishment upon social welfare can be mitigated by way of threat. Because signalling punishers cooperate indiscriminately, they outperform fearful defectors who are always vying for higher status at the expense of others, including themselves.

The prosperity of the population observed under threat of punishment speaks for the preventive nature of advertising acts of justice. Undeniably, it is a beneficial outcome for wicked ventures not to occur in the first place, but contexts such as the development of AI or climate change provide us with unparalleled incentive to prevent potentially disastrous consequences. Given the importance of intrinsic factors that guide the decisions of researchers and policy makers in the field [Baum, 2017], we aim to explore further how the concept of threat, and the self-preservation associated with it, could help guide the current literature on this issue. Additionally, implementing this type of signal response could improve safety conditions in MA systems such as artificial societies [Balke and Villatoro, 2011; Villatoro et al., 2011], especially in cases where the transparency of interactions is reduced.

Moreover, punishment has been suggested to be relatively uncommon in nature, often existing as a way to equalise payoffs rather than as a deterrent to prompt behaviour change [Raihani and Bshary, 2015, 2019; Raihani et al., 2012]. Thus, punishment by itself often prompts retaliatory behaviour if enacted by a peer, something which can be avoided with an institutional sanctioning scheme. Importantly, we have attempted to move away from purely negative incentives and looked instead for ways in which punishment can be avoided. We note also that signalling might play a key role in the distribution of positive incentives, as well. While the mathematical formulations of the promise of reward and the threat of punishment are similar, much of the psychological literature (see e.g. [Raihani et al., 2012]) indicates that human agents would respond very differently to the two. Certainly, one could envision a whole array of behavioural experiments solely designed to answer whether the promise of reward could be more effective at promoting cooperation than the threat of punishment in human subjects.

Moreover, we set the foundational basis for signalling the threat of institutional sanctions. In this setting, we particularly focus on defectors with varying degrees of fear or aversion to being punished, and on the effect of network structures on institutional punishment. We show that in socially diverse settings, when defectors are fearful of the signal, sanctions can be reduced, positively affecting cooperative outcomes and total costs. Heterogeneous structures of interaction can lead to the emergence of cooperation even when sanctions are small, reducing the financial burden on sanctioning institutions. Furthermore, we argue that more investigation is needed to explore the viability of signalling institutional acts of punishment in homogeneous populations. Overall, we argue that advertising punitive acts can be an effective way to avoid the pitfalls of traditional punishment, leading to higher social welfare and pro-social behaviour.

6 | AI Safety in Heterogeneous Settings

Our posturings, our imagined self-importance, the delusion that we have some privileged position in the Universe, are challenged by this point of pale light. Our planet is a lonely speck in the great enveloping cosmic dark. In our obscurity, in all this vastness, there is no hint that help will come from elsewhere to save us from ourselves.

—Carl Sagan, *Pale Blue Dot*

In which we explore the dynamics of safety behaviour in an AI race for supremacy. Here[1], we study the effects of social diversity on safety outcomes in AI development. Our findings indicate that, when participants portray a strong diversity in terms of connections and peer-influence, the tension that exists in homogeneous settings is significantly reduced, thereby lessening the need for regulatory actions. Furthermore, our results suggest that the design and implementation of meticulous interventions on a minority of participants can influence an entire population towards an ethical and sustainable use of advanced technology.

6.1 Introduction

Researchers and stakeholders alike have urged for due diligence in regard to AI development on the basis of several concerns. Not least among them is that AI systems could easily be applied to nefarious purposes, such as espionage or cyberterrorism [Taddeo and Floridi, 2018]. Moreover, the desire to be at the foreground of the state-of-the-art or the pressure imposed by upper management might tempt developers to ignore safety procedures or ethical consequences [Armstrong et al., 2016; Cave and ÓhÉigeartaigh, 2018]. Indeed, such concerns have been expressed in many forms, from letters of scientists against the use of AI in military applications [Future of Life Institute, 2015, 2019], to blogs of AI experts requesting careful communications [Brooks, 2017], and proclamations on the ethical use of AI [Declaration, 2018; Jobin et al., 2019; Russell et al., 2015; Steels and de Mantaras, 2018].

Regulation and governance of advanced technologies such as Artificial Intelligence (AI) has become increasingly more important given their potential implications, such as associated risks and ethical issues [Declaration, 2018; European Commission, 2020; Future of Life Institute, 2015, 2019; Jobin et al., 2019; Perc et al., 2019; Russell et al., 2015; Steels and de Mantaras, 2018]. With the great benefits promised from being first able to supply such technologies, stake-holders might cut corners on safety precautions in order to ensure a rapid deployment, in a race towards AI market supremacy (AIS)

[1] The model and results presented in this chapter are also reported in Cimpeanu et al. [2022].

[Armstrong et al., 2016; Cave and ÓhÉigeartaigh, 2018]. One does not need to look very far to find potentially disastrous scenarios associated with AI [Armstrong et al., 2016; O'neil, 2016; Pamlin and Armstrong, 2015; Sotala and Yampolskiy, 2014], but accurately predicting outcomes and accounting for these risks is exceedingly difficult in the face of uncertainty [Armstrong et al., 2014]. As part of the double-bind problem put forward by the Collingridge Dilemma, the impact of a new technology is difficult to predict before it has been already extensively developed and widely adopted, and also difficult to control or change after it has become entrenched [Collingridge, 1980]. Given the lack of available data and the inherent unpredictability involved in this new field of technology, a modelling approach is therefore desirable to provide a better grasp of any expectations with regard to a race for AIS. Such modelling allows for dynamic descriptions of several key features of the AI race (or its parts), providing an understanding of possible outcomes, considering external factors and conditions, and the ramifications of any policies that aim to regulate such race.

With this aim in mind, a baseline model of an innovation race has been recently proposed [Han et al., 2020], in which innovation dynamics are pictured through the lens of Evolutionary Game Theory (EGT) and all race participants are equally well-connected in the system (well-mixed populations). The baseline results showed the importance of accounting for different time-scales of development, and also exposed the dilemmas that arise when what is individually preferred by developers differs from what is globally beneficial. When domain supremacy could be achieved in the short-term, unsafe development required culling for to promote the welfare of society, and the opposite was true for the very long term, to prevent excessive regulation at the start of exploration. However, real-world stakeholders and their interactions are far from homogeneous. Some individuals are more influential than others, or play different roles in the unfolding of new technologies. Technology races are shaped by complex networks of exchange, influence, and competition where diversity abounds. It has been shown that particular networks of contacts can promote the evolution of positive behaviours in various settings, including cooperation [Chen et al., 2015; Ohtsuki et al., 2006; Perc and Szolnoki, 2010; Perc et al., 2017; Santos et al., 2006a, 2008], fairness [Cimpeanu et al., 2021a; Page et al., 2000; Santos et al., 2017; Szolnoki et al., 2012; Wu et al., 2013] and trust [Kumar et al., 2020]. In this chapter, we take inspiration from the disconnect between the previous line of research and the heterogeneity observed in real-world interactions, and ask whether network topology can influence the adoption of safety measures in innovation dynamics, and shape the tensions of the AI development race. Thus, in this chapter we study the dynamics of the previously proposed innovation race [Han et al., 2020] on multiple networks of interaction, but also explore several novel ways in which safety zealots could be exploited to reduce the existential risks of swift AI development.

The impact of network topology is particularly important in the context of technology regulation and governance. Technology innovation and collaboration networks (e.g. among firms, stakeholders and AI researchers) are highly heterogeneous [Newman, 2004; Schilling and Phelps, 2007]. Developers or development teams interact more frequently within their groups than without, forming alliances and networks of followers and collaborators [Ahuja, 2000; Barabasi, 2014]. Many companies compete in several markets while others compete in only a few, and their positions in inter-organisational networks strongly influence their behaviour (such as resource sharing) and innovation outcome [Ahuja, 2000; Shipilov and Gawer, 2020]. It is important to understand how diversity in the network of contacts influences race dynamics and the conditions under which regulatory actions are needed. Therefore, we depart from a minimal AI race model [Han et al., 2020], examining instead how network structures influence safety decision making within an AI development race.

In a structured population, players are competing with co-players in their network neighbourhoods. Firms interact or directly compete through complex ties of competition, such that some players may play a pivotal role in a global outcome. Here we abstract these relationships as a graph or a network. We compare different forms of network structures, from homogeneous ones — such as complete graphs (equivalent to well-mixed populations), and square lattices — to different types of scale-free networks [Barabási and Albert, 1999] (see Methods), representing different levels of diversity in the number of co-player races a player can compete in. Our results show that when race participants are distributed in a heterogeneous network, the conflicting tensions arising in the well-mixed case are significantly reduced, thereby softening the need for regulatory actions. This is, however, not the case when the network is not accompanied by some degree of relational heterogeneity, even in different types of spatial lattice networks.

In the following sections, we describe the models in detail, then present our results.

6.2 Models and Methods

We first define the AI race game [Han et al., 2020] and recall relevant results from previous works in the well-mixed populations setting.

6.2.1 AI race model definition

Assuming that winning the race towards supremacy is the goal of the development teams (or players) and that a number of development steps (or advancements/rounds) are required, the players have two strategic options in each step: to follow safety precautions (denoted by strategy SAFE) or to ignore them (denoted by strategy UNSAFE) [Han et al., 2020]. As it takes more time and effort to comply with the precautionary requirements, playing SAFE is not only costlier, but also implies a slower development speed, compared to playing UNSAFE. Let us also assume that to play SAFE, players need to pay a cost c, whereas the opposite strategy is free. The increase in speed when playing UNSAFE is given by a free parameter $s > 1$, while the speed when playing SAFE is normalised to 1. The interactions are iterated until one or more teams achieve a designated objective, after having completed W development steps. As a result, the players obtain a large benefit B, shared among those who reach the target objective at the same time. However, a setback or disaster can happen with some probability, which is assumed to increase with the number of times the safety requirements have been omitted by the winning team(s). Although many potential AI disaster scenarios have been sketched [Armstrong et al., 2016; Pamlin and Armstrong, 2015], the uncertainties in accurately predicting these outcomes are high. When such a disaster occurs, risk-taking participants lose all their benefits. We denote by

p_r the risk probability of such a disaster occurring when no safety precaution is followed at all.

We model an AI development race as a repeated two-player game, consisting of W development rounds. In each round, the players can collect benefits from their intermediate AI products, depending on whether they choose to play SAFE or UNSAFE. Assuming a fixed benefit b, from the AI market, teams share this benefit proportionally to their development speed. Moreover, we assume that with some probability p_{fo} those playing UNSAFE might be found out, wherein their disregard for safety precautions is exposed, leading to their products not being adopted due to safety concerns, thus receiving 0 benefit. Thus, in each round of the race, we can write the payoff matrix as follows (with respect to the row player)

$$\Pi = \begin{array}{c} \\ \mathit{SAFE} \\ \mathit{UNSAFE} \end{array} \begin{array}{c} \begin{array}{cc} \mathit{SAFE} & \mathit{UNSAFE} \end{array} \\ \begin{pmatrix} -c + \frac{b}{2} & -c + (1 - p_{fo})\frac{b}{s+1} + p_{fo}b \\ (1 - p_{fo})\frac{sb}{s+1} & (1 - p_{fo}^2)\frac{b}{2} \end{pmatrix} \end{array}. \tag{6.1}$$

For instance, when two SAFE players interact, each needs to pay the cost c and they share the benefit b. When a SAFE player interacts with an UNSAFE one, the SAFE player pays a cost c and obtains (with probability p_{fo}) the full benefit b in case the UNSAFE co-player is found out, and obtains (with probability $1 - p_{fo}$) a small part of the benefit $b/(s+1)$ otherwise, dependent on the co-player's speed of development s. When playing with a SAFE player, the UNSAFE one does not have to pay any cost and obtains a larger share $bs/(s+1)$ when not found out. Finally, when an UNSAFE player interacts with another one, it obtains the shared benefit $b/2$ when both are not found out, but the full benefit b when it is not found out while the co-player is found out, and 0 otherwise. The corresponding average payoff is thus: $(1 - p_{fo})\Big[(1 - p_{fo})(b/2) + p_{fo}b\Big] = (1 - p_{fo}^2)\frac{b}{2}$.

In the AI development process, players repeatedly interact (or compete) with each other using the *innovation* game described above. In order to clearly examine the effect of population structures on the overall outcomes of the AI race, in line with previous network reciprocity analyses (e.g. in social dilemma games [Santos et al., 2006a, 2008; Szabó and Fáth, 2007]), we focus in this chapter on two unconditional strategies [Han et al., 2020]:

- AS (always complies with safety precautions)
- AU (never complies with safety precautions)

Denoting by Π_{ij} $(i, j \in \{1, 2\})$ the entries of the matrix Π above, the payoff matrix defining the averaged payoffs for AU vs AS reads

$$\begin{array}{c c} & \begin{array}{cc} AS & AU \end{array} \\ \begin{array}{c} AS \\ AU \end{array} & \left(\begin{array}{cc} \frac{B}{2W} + \Pi_{11} & \Pi_{12} \\ (1-p_r)\left(\frac{sB}{W} + \Pi_{21}\right) & (1-p_r)\left(\frac{sB}{2W} + \Pi_{22}\right) \end{array}\right). \end{array} \tag{6.2}$$

As described in Equation 6.2, we encounter the following scenarios. When only two safe players interact, they complete the race simultaneously after an average of W development rounds, thereby obtaining the averaged split of the full prize $\frac{B}{2W}$ per round; furthermore, the safe players also obtain the intermediate benefit per round (π_{11}, see Equation 6.1). When a safe player only encounters an unsafe player, the only benefit obtained by the safe player is the intermediary benefit in each round, whereas the unsafe player receives the full prize B; moreover, the unsafe player completes the race in $\frac{W}{s}$ development rounds, so it receives an extra average of $\frac{sB}{W}$ of the full prize per round. Furthermore, the unsafe behaviour attracts the possibility of a disaster occurring, causing them to lose all gains, with probability p_r, which is reflected in the payoff matrix (consider π_{22} in Equation 6.1). Similarly, we can extract the average payoffs for solely two unsafe players interacting, by considering that they finish the race at the same time and get the appropriate intermediate benefit π_{22} (See Equation 6.1).

6.2.2 Summary of previous results in well-mixed settings

In order to clearly present the contribution of the present work, we recall the analytical conditions derived in [Han et al., 2020] and how these will be used to inform the analysis that follows. Our analysis will differentiate between two development regimes: an early/short-term regime and a late/long-term one. The difference in time-scale between the two regimes plays a key role in identifying which regulatory actions are needed and when. This distinction is in line with previous works adopting analytical approaches

Table 6.1 Model parameters and parameter space analysed

Parameter	Symbol	Range Analysed
Population size	N	{100, 1000, 1024}
Intensity of selection	β	{1}
Average connectivity of a scale-free network	z	{4}
Number of new edges for each new node in SF networks	m	{2}
Probability of being found out when playing unsafe	p_{fo}	{0, 0.05, 0.1, ..., 1}
Probability of disaster occurring due to unsafe development	p_r	{0, 0.05, 0.1, ..., 1}
Benefit of winning the race (reaching AI supremacy)	B	$\{10^4\}$
Benefit of intermediate AI advancements	b	{4}
Cost of adhering to safety standards	c	{1}
Speed of development (due to disregarding safety)	s	{1, 1.25, 1.5, ..., 5}
Number of development rounds until AI supremacy is reached	W	$\{100, 10^6\}$

using stochastic population dynamics. The early regime is underpinned by the race participants' ability to readily reach the ultimate prize B in the shortest time frame available. In other words, winning the ultimate prize in W rounds is much more important than any benefits achieved in single rounds until then, i.e. $B/W >> b$. Contrarily, a late regime is defined by a desire to do well in each development round, as technological supremacy cannot be achieved in the foreseeable future. That is, singular gains b, even when accounting for the safety cost c, become more tempting than aiming towards winning the ultimate prize, i.e. $B/W << b$. For a reminder of the meanings of the parameters described above, see Table 6.1.

We have also made use of the previous analytical results [Han et al., 2020] which identify the risk-dominant boundaries of the AI race game for both early and late development regimes in well-mixed populations. These are useful as a baseline or reference model, determining the regions in which regulatory actions are needed or otherwise, and moreover, if needed, which behaviour should be promoted. In the early regime, the two dotted lines mark region (II) within the boundaries $p_r \in [1 - 1/s,\ 1 - 1/(3s)]$ for which safety development is the preferred collective outcome, but where unsafe development is selected for by social dynamics (see e.g. Figure 6.1, first row). Thus, in this region (II), regulation is required to improve safety compliance. Outside of these boundaries, safe (in region I) and unsafe (in region III), respectively, are both the preferred collective outcomes and the ones selected for by social dynamics, hence requiring no regulatory actions. For the late AI race (e.g. Figure 6.1, bottom row), the solid black line marks the boundary above which safety is the preferred collective outcome, where $p_r < 1 - \frac{b-2c}{b(1-p_{fo}^2)}$, whereas the blue line indicates where AS becomes risk-dominant against AU, where $p_r < \frac{4c(s+1)+2b(s-1)}{b(1+3s)}$. Again, in this regime three regions can be distinguished, with (I) and (III) having similar meanings to those in the early regime. However, differently from the early regime, in region (II) regulatory actions are needed to improve (unsafe) innovation instead of safety compliance, due to the low risk. These regions are derived from the analytical conditions described in [Han et al., 2020], where these are explained in further detail.

6.2.3 Population Dynamics

We consider a population of agents distributed on a network (see below for different network types), who are randomly assigned a strategy AS or AU. Below, we recall the evolutionary process presented in Sections 2.2 and 2.3, in Chapter 2. At each time step or generation, each agent plays the game with its immediate neighbours. The success of each agent (i.e., its fitness) is the sum of the payoffs in all these encounters. In Appendix C, we also discuss the limit where scores are normalised by the number of interactions (i.e., the connection *degree* of a node) [Santos and Pacheco, 2006]. Each individual fitness, as detailed below, defines the time-evolution of strategies, as successful choices will tend to be imitated by their peers.

At the end of each generation, a randomly selected agent A with a fitness f_A chooses to copy the strategy of a randomly selected neighbour, agent B, with fitness f_B with a probability p that increases with their fitness difference. Here we adopt the well-studied Fermi update or pairwise comparison rule, where [Santos et al., 2012a; Traulsen et al., 2006]:

$$p = (1 + e^{\beta(f_A - f_B)})^{-1}. \tag{6.3}$$

In this case, β conveniently describes the selection intensity — i.e., the importance of individual success in the imitations process: $\beta = 0$ represents neutral drift while $\beta \to \infty$ represents increasingly deterministic imitation

[Traulsen et al., 2006]. Varying β allows capturing a wide range of update rules and levels of stochasticity, including those used by humans, as measured in lab experiments [Grujić and Lenaerts, 2020; Rand et al., 2013; Zisis et al., 2015]. In line with previous works and lab experiments, we set $\beta = 1$ in our simulations, ensuring a high intensity of selection [Pinheiro et al., 2012b]. This update rule implicitly assumes an asynchronous update rule, where at most one imitation occurs at each time-step. We have nonetheless confirmed that similar results are obtained with a synchronous update rule.

6.2.4 Network Topologies

Links in the network describe a relationship of proximity both in the interactional sense (whom the agents can interact with), but also observationally (whom the agents can imitate). Ergo, the network of interactions coincides with the imitation network [Ohtsuki et al., 2007]. As each network type converges at different rates and naturally presents with various degrees of heterogeneity, we choose different population sizes and maximum numbers of runs in the various experiments to account for this while optimising runtime. For a detailed discussion of the studied network topologies, please see Section 2.4 in Chapter 2.

Specifically, to study the effect of network structures on the safety outcome, we will analyse the following types of networks, from simple to more complex:

1. Well-mixed population (WM) (complete graph): each agent interacts with all other agents in a population,

2. Square lattice (SL) of size $N = L \times L$ with periodic boundary conditions—a widely adopted population structure in population dynamics and evolutionary games (for a survey, see [Szabó and Fáth, 2007]). Each agent can only interact with its four immediate edge neighbours. We also study the 8-neighbour lattice for confirmation (see Appendix C),

3. Scale-free (SF) networks [Barabási and Albert, 1999; Dorogovtsev, 2010; Newman, 2003], generated through two growing network models — the widely-adopted Barabási-Albert (BA) model [Albert and Barabási, 2002;

Barabási and Albert, 1999] and the Dorogovtsev-Mendes-Samukhin (DMS) model [Dorogovtsev, 2010; Dorogovtsev et al., 2001], the latter of which allows us to assess the role of a large number of triangular motifs (i.e. high clustering coefficient). Both BA and DMS models portray a power-law degree distribution $P(k) \propto k^{-\gamma}$ with the same exponent $\gamma = 3$. In the BA model, graphs are generated via the combined mechanisms of growth and preferential attachment where new nodes preferentially attach to m existing nodes with a probability that is proportional to their already existing number of connections [Barabási and Albert, 1999]. In the case of the DMS model, new connections are chosen based on an edge lottery: each new vertex attaches to both ends of randomly chosen edges, also connecting to m existing nodes. As such, we favour the the creation of triangular motifs, thereby enhancing the clustering coefficient of the graph. In both cases, the average connectivity is $z = 2m$.

Overall, WM populations offer a convenient baseline scenario, where interaction structure is absent. With the SL we introduce a network structure, yet one where all nodes can be seen as equivalent. Finally, the two SF models allow us to address the role of heterogeneous structures with low (BA) and high (DMS) clustering coefficients. The SF networks portray a heterogeneity which mimics the power-law distribution of wealth (and opportunities) of real-world settings.

6.2.5 Computer Simulations

For well-mixed populations and lattice networks, we chose populations of $N = 100$ agents and $N = 32 \times 32$ agents, respectively. In contrast, for scale-free networks, we chose $N = 1000$, while also pre-seeding with agents 10 different networks (of each type) on which to run all the experiments in an effort to minimise the effect of network topology and the initial, stochastic distributions of players. We chose an average connectivity of $z = 4$ for our SF networks, to coincide with the regular average connectivity in square lattices for the sake of comparison.

We simulated the evolutionary process for 10^4 generations (a generation corresponds to N time-steps) in the case of scale-free networks and 10^3 generations otherwise. The equilibrium frequencies of each strategy were obtained by averaging over the final 10^3 steps. Each data point shown below was obtained from averaging over 25 independent realisations, for each of the 10 different instances used in each network topology.

6.3 Results

Based on extensive computer simulations, our analysis identifies the prevalence of individuals adopting unsafe procedures after reaching a stationary state and infer the most likely behavioural trends and patterns associated with the agents taking part in the AI race game for distinct network topologies. The main findings from this work are described in this section, whereby each subsection will provide a key insight followed by the results and intuitions which motivate each claim.

6.3.1 Heterogeneous interactions reduce unsafe development viability in both short and long-term races

We examine the impact of different network structures, homogeneous and heterogeneous, on the safety outcome of the evolutionary dynamics for the two different development regimes descried above. To commence our analysis, we first study the role of degree-homogeneous graphs (here illustrated by structural spatiality) in the evolution of strategies in the AI race game. Firstly, we simulated the AI race game in well-mixed populations (see Figure 6.1, first column). We then explored the same game on a square lattice, where each agent can interact with its four edge neighbours, in Figure 6.1 (second column). We show that the trends remain the same when compared with well-mixed populations, with very slight differences in numerical values between the two. Specifically, towards the top of area (Region **II**), at the risk-dominant boundary between AS and AU players in the case of an early AI race, we see some safe developmental activity where previously there was

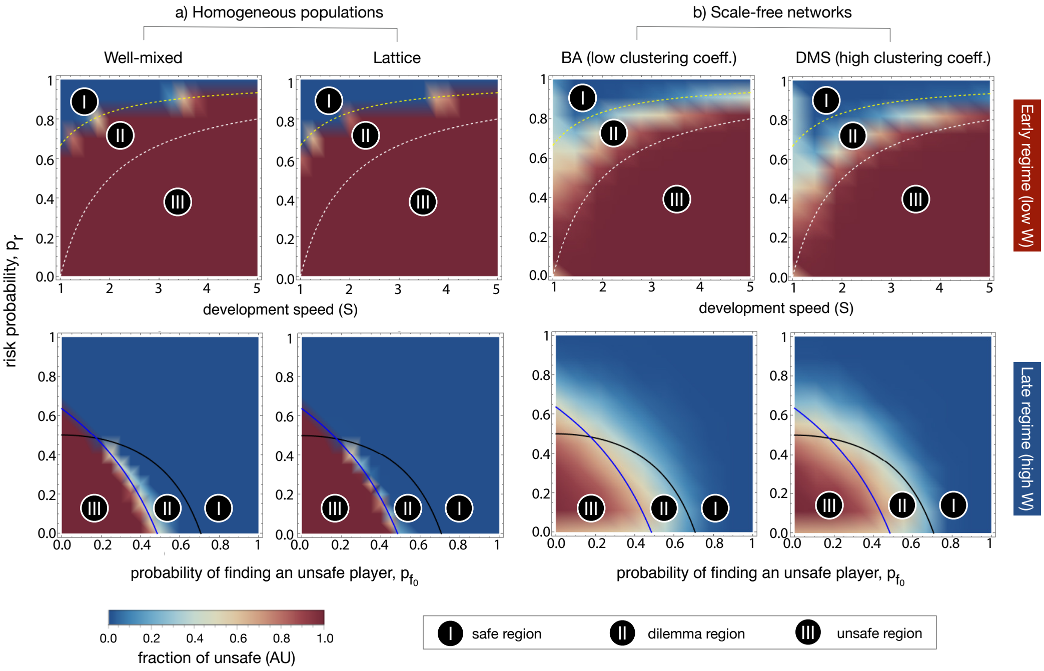

Fig. 6.1 **Safety dynamics in the AI race with different networks of interaction**. Dotted and full lines indicate the phase diagram obtained analytically [Han et al., 2020]. In the early regime (upper panels), region II indicates the parameters in which safe AI development is the preferred collective outcome, but unsafe development is expected to emerge and regulation may be needed — thus the dilemma. In regions I and III, safe and unsafe AI development, respectively, are both the preferred collective outcomes and the ones expected to emerge from self-organization, hence not requiring regulation. In the late regime (lower panels), the solid black line marks the boundary above which safety is the preferred outcome, whereas the blue line indicates the boundary above which safety becomes risk dominant against unsafe development. The results obtained for well-mixed populations and lattices (a) suggest that, for both early and late regimes, the nature of the dilemma, as represented by the analytical phase diagram, remains unaltered. Moreover, homogeneous interaction structures cannot reduce the need for regulation in the early regime. Differently, we show that heterogeneous interaction structures (scale-free networks, (b)) are able to significantly reduce the prevalence of unsafe behaviors for almost all parameter regions, including both late and early regimes. This effect is enhanced whenever scale-free networks are combined with high clustering coefficient. Other parameters: $p_{fo} = 0.5$, and $W = 100$ (top panels); $s = 1.5$ and $W = 10^6$ (bottom panels).

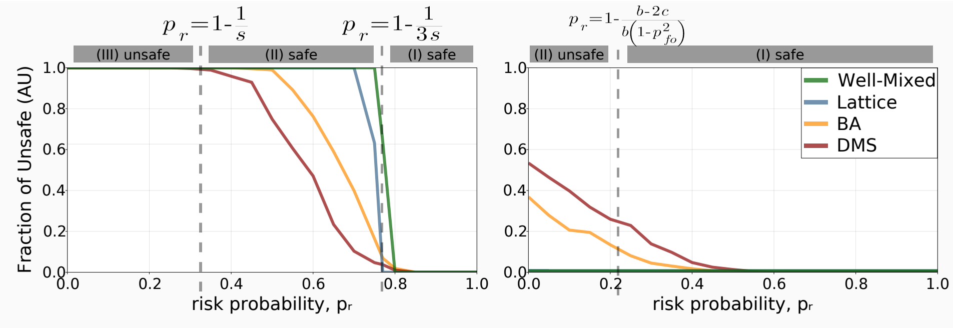

Fig. 6.2 **Heterogeneous networks moderate the need for regulation**, shown by measuring frequency of unsafe developments across a range of different risk probabilities. The boundaries between zones are indicated with blue dashed lines, whereas the grey-highlighted texts on top of the figures indicate the collectively desired behaviour in each zone. The left panel reports the results for the early regime ($p_{fo} = 0.5$, $W = 100$), while the right panel does so for the late regime ($p_{fo} = 0.6$, $W = 10^6$) (parameter values are chosen for a clear illustration). Parameters: $s = 1.5$.

none. In practice, this shifts the boundary very slightly towards an optimal conclusion.

Thus, except for minute atypical situations, we may argue that homogeneous spatial variation is not enough to influence safe technological development, with minimal improvement when compared with a well-mixed population (complete network). To further increase our confidence that such structures have very small effects on the AI race game, we confirm that 8-neighbour lattices (where agents can also interact with corner neighbours) yield very similar trends, with negligible differences when compared to either the regular square lattice or well-mixed populations (see Appendix C, Figure C.2).

As a means of investigating beyond simple homogeneous structures and their roles in the evolution of appropriate developmental practices in the AI race, we make use of the previously defined BA and DMS network models. Contrary to the findings on homogeneous networks, scale-free interaction structures produce marked improvements in almost all parameter regions of the AI race game (see Figure 6.1).

Previously, it has been suggested that different approaches to regulation were required, subject to the time-line and risk region in which the AI development race is placed, after inferring the preferences developers would have towards safety compliance [Han et al., 2020]. Given that innovation in the field of AI (or more broadly, technological advancement as a whole), should be profitable (and robust) to developers, shareholders and society altogether, we must therefore discuss the analytical loci where these objectives can be fulfilled. Assuredly, we see that diversity in players introduces two marked improvements in both early and late safety regimes. Firstly and most importantly, we note that very little regulation is required in the case of a late AI race (large W), principally concerning the existing observations in homogeneous settings (e.g., well-mixed populations and lattices). Intuitively, this suggests that there is little encouragement needed to promote risk-taking in late AIS regimes: Diversity enables beneficial innovation. Secondly, the region for early AIS regimes in which regulation must be enforced is diminished, but not completely eliminated. Consequently, governance should still be prescribed when developers are racing towards an early or otherwise unidentified AI regime (based on the number of development steps or risk of disaster). It stands to reason that insight into what regime type the AI race operates in, is therefore paramount to the success of any potential regulatory actions. The following sections will attempt to look further into assessing these observations.

Figure 6.1 (top panels) presents a fine-grained glimpse into the early regime. In region (**II**), the safety dilemma zone, social welfare is once more conspicuously improved by heterogeneity. Concerted safe behaviour is favoured, even in the face of being disregarded by social dynamics in the analytical sense. We discern the clear improvements discussed earlier, but also echo the messages put forward in [Han et al., 2020]. We contend that it is vital for regulators to intervene in these conditions, for encouraging prosocial, safe conduct, and in doing so avert conceivably dangerous outcomes. Heterogeneity lessens the burden on policy makers, allowing for greater freedom in the errors and oversights that could occur in governing towards the goal of safe AI development.

While the difference between heterogeneous and homogeneous networks is evident, there also exists a distinction between the different types of heterogeneous networks. In this chapter we discuss the BA and DMS models, and also their normalised counterparts, in which individuals' payoffs are divided by the number of neighbours. In such scenarios one could assume that there is an inherent cost to maintaining a link to another agent. In this sense, there exists some levelling of the payoffs, seemingly increasing fairness and reducing wealth inequality. But we confirm that normalising the network leads to similar dynamics as observed in homogeneous populations (see Figure C.3 in Appendix C), with only very slight differences.

In order to accurately depict the measured differences between the different types of networks, we varied the risk probability (p_r) for both the early and late regime. We report the results of this analysis in Figure 6.2, where we also show the preferred collective outcome, using the different regions described earlier in this section. These figures help expose the effect of heterogeneity on the frequency of unsafe behaviour in the different dilemma zones. In particular, we notice a mediating effect in the requirements for regulation, for both regimes and types of scale-free networks.

Specifically, in the case of the early regime (see Figure 6.2, left column), we observe the presence of safety for a much broader range of risk probability values, than in the case of either well-mixed or structured populations. In the late regime (see Figure 6.2, right column), however, we also highlight an increase in unsafe behaviour even beyond the boundary for which safety would have been the preferred collective outcome. In this case, heterogeneity has its drawbacks. On the one hand, innovative behaviour sees some improvement when it is in the interest of the common good for it to be so, but the same is true, albeit rarely, when it is not. We also note that the effects described above are amplified in the case of DMS networks, in comparison to their BA counterparts. Observing a high degree of inter-group interactions (clustering) may play a key role in determining if intervention is required in the AI race. Moreover, we confirm these findings by producing typical runs showing the time evolution of unsafe behaviour for each network type (please see Figure C.1 in Appendix C).

6.3.2 Hubs and their role in decelerating the race

Highly connected individuals (hubs) typically play a key role in many real-world networks of contacts and change the dynamics observed in heteroge-

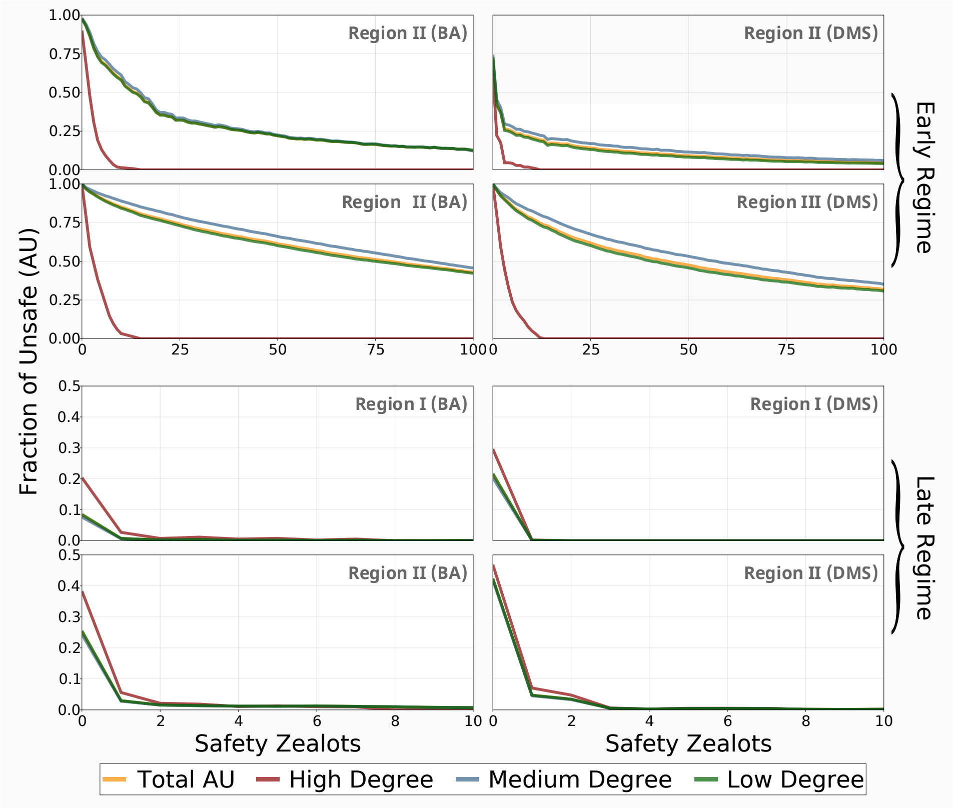

Fig. 6.3 **Hubs prefer slower, thus safer developments in the early race**, and this can be further exploited by progressively introducing safety zealots in highly connected nodes. We show the results for both regimes, as well as the appropriate regions where safety (early region II and late region I), and conversely where innovation (early region III and late region II) are the preferred collective outcomes. The top four panels report the results for the early regime ($p_{fo} = 0.5,\ W = 100$ with $p_r = 0.5$ for region II and $p_r = 0.1$ for region III), and the bottom four do so for the late regime ($p_{fo} = 0.6,\ W = 10^6$ with $p_r = 0.3$ for region I and $p_r = 0.1$ and region II). We show a subset of the results in the late regime for clear representation; see Figure C.8 in Appendix C for a comprehensive view. Other parameters: $s = 1.5$.

neous populations [Pastor-Satorras and Vespignani, 2001; Perc and Szolnoki, 2010; Santos et al., 2006a, 2008]. In order to study the role that hubs play in the AI race, in the context of scale-free networks, we classify nodes into three separate connectivity classes [Santos et al., 2008]. We obtain three classes of individuals, based on their number of contacts (links) k_i and the average network connectivity z:

1. Low degree, whenever $k_i < z$,
2. Medium degree, whenever $z \leq k_i < \frac{k_{max}}{3}$ and
3. High degree (hubs), whenever $\frac{k_{max}}{3} \leq k_i \leq k_{max}$.

Dedicated minorities are often identified as major drivers in the emergence of collective behaviours in social, physical and biological systems, see [Cardillo and Masuda, 2020; Pacheco and Santos, 2011; Paiva et al., 2018; Santos et al., 2019]. Given the relative importance of hubs in other systems, we explore whether highly connected, committed individuals are prime targets for safety regulation in the AI race. By introducing individuals with pathologically safe tendencies (fixed behaviours) [Santos et al., 2019]—these are sometimes referred to as zealots, see [Cardillo and Masuda, 2020; Kumar et al., 2020; Pacheco and Santos, 2011; Santos et al., 2019]—in the network, we can better understand the power of influential devotees in the safe development of a general AI.

We progressively introduce pathological safe players based on their degree centrality (i.e. number of connections). In other words, the most connected nodes will be the first to be targeted. The benefits of this approach are twofold, as they allow us to study the relative differences between the three classes of individuals, but also the effect of regulating the key developers in the AI race. For a full analysis of the differences between high, medium and low degree individuals in the baseline case, please see Figure C.5 in Appendix C.

Hubs prefer slower, safer developments in the early AI race, and this can be further exploited by introducing safety zealots in key locations in the network (see Figure 6.3). When safety is the preferred collective outcome, hubs can drive the population away from unsafe development, and this effect is even more apparent in the case of highly clustered scale-free networks (Figure 6.3, right column). Following the sharp increase in global safety after the conversion of high degree players to zealotry, we also observe a similar, but not as pronounced influence as the most connected medium degree individuals follow suit, an effect which plateaus shortly thereafter. We further confirm these results by selecting the same targets (the top 10% of individuals based on degree centrality), but introducing them in reverse order (i.e. starting with the highest connected medium degree individuals and ending with the most connected high degree ones; see Figure C.9 in Appendix C).

Whereas the capacity of hubs to drive the population towards safety is evident in region II of the early regime (when safety is the collective preferred outcome), the opposite is true for region III. High degree individuals are more capable at influencing the overall population than medium degree individuals, even the most highly connected ones, but we see a much more gradual decrease in innovation as the most connected nodes are steadily converted to zealotry. Even in the presence of great uncertainty, a small percentage of very well connected developers can ensure safety with very little negative impact on innovation.

Conversely, we show that highly connected individuals prefer innovation in the late regime, irrespective of the preferred collective outcome. But even the introduction of one pathological safe player (0.1% of the population) in the largest hub is enough to ensure that the entire population converges to safe development in most instances. In the case when safety is socially preferred (third row of Figure 6.3), the successful regulation of the AI race requires a very small minority of individuals to dedicate themselves to safety, but in cases of uncertainty, innovation is very easy to stifle in the late regime, even when it would be beneficial not to do so (region II).

6.3.3 A small minority of highly connected individuals can help mitigate race tensions under uncertainty

Uncertainty can limit the options of regulatory agencies in the quest towards the development of safe AI, and narrow solutions to regulation could have

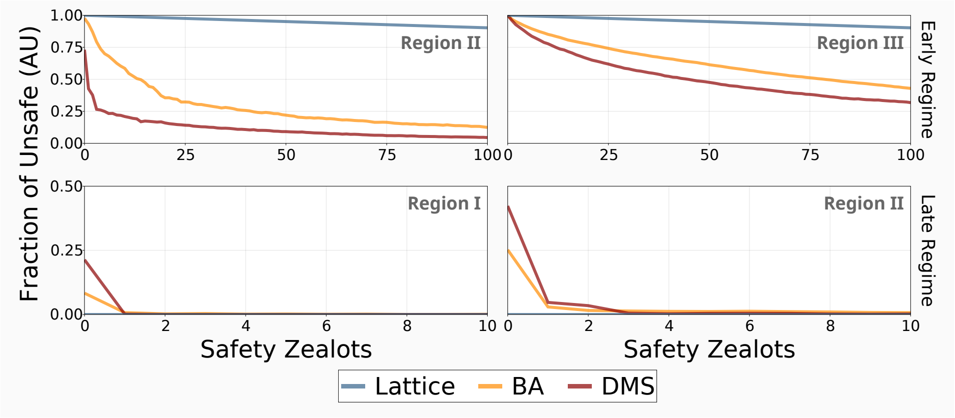

Fig. 6.4 **Introducing a small number of safety zealots can mitigate race tensions under uncertainty**. We show the results for both regimes, as well as the appropriate regions where safety (early region II and late region I), and conversely where innovation (early region III and late region II) are the preferred collective outcomes. The top panels report the results for the early regime ($p_{fo} = 0.5$, $W = 100$ with $p_r = 0.5$ for region II and $p_r = 0.1$ for region III), and the bottom do so for the late regime ($p_{fo} = 0.6$, $W = 10^6$ with $p_r = 0.3$ for region I and $p_r = 0.1$ and region II). We note that these values were chosen for clear representation. Other parameters: $s = 1.5$.

potentially disastrous consequences, given the existential risk that general AI poses to humanity [Scherer, 2015]. Moreover, the promised benefit of such technology is great enough that stifling innovation could be nearly as harmful as the catastrophic consequences themselves, given the potential solutions that such technology could provide to problems in the context of existential risk, healthcare, politics, and many other fields (e.g. [McKinney et al., 2020; Rolnick et al., 2019]).

To provide general solutions to the problem of regulating the AI race, we explore the impact of safety zealots (as discussed in the previous section) across the whole range of possible scenarios. We cannot be sure of the nature of the network of contacts that governs real-world AI developers, nor the actual timeline of the race. We show that by enforcing safety for a very select minority of highly connected individuals, race tensions can be mitigated in nearly all cases (see Figure 6.4). We provide a full analysis of the effect of zealots in well-mixed networks in Figure C.4 (in Appendix C), and note that the lack of heterogeneity produces nearly identical results to lattice networks.

Slowing key individuals in the early regime can dramatically reduce existential risk in the case of heterogeneous interactions. For both regions, hubs in DMS networks can drive the other nodes towards safety (see Figure 6.4, top panels), but the reduction in unsafe developments in region II is significantly higher than in region III for low numbers of safety zealots. Outside of the few individuals that are converted to zealots, other nodes maintain their speed and continually innovate in region III, which suggests that this approach could be fundamental to the governance of developmental races. We note that if the proportion of safety zealots is not high enough, this effect cannot be reproduced, even in the presence of additional interference (such as artificially funding zealots or accelerating their development); For a more thorough analysis, please see Figures C.6 and C.7, in Appendix C.

Given a drawn-out race, this small minority of zealots can negatively impact innovation in the late regime (region II), where the relative increase that heterogeneous interactions provided rapidly disappears as pathological players are introduced. On the other hand, conditional strategies have been shown to further diminish the need to promote innovation in these conditions [Han et al., 2020], and the introduction of these advanced strategies in this model could eliminate the negative effects of safety zealots in this region.

6.4 Discussion

In this chapter, we have considered the implications of network dynamics on a technological race for supremacy in the field of AI, with its implied risks and hazardous consequences [Armstrong et al., 2016; Pamlin and Armstrong, 2015; Sotala and Yampolskiy, 2014]. We make use of a previously proposed evolutionary game theoretic model [Han et al., 2020] and study how the tension and temptation resulting from the race can can be mediated, for both early and late development regimes.

Network reciprocity has been shown to promote the evolution of various positive outcomes in many settings [Ranjbar-Sahraei et al., 2014; Santos et al., 2008; Szolnoki et al., 2012; Wu et al., 2013] and, given the high levels of heterogeneity identified in the networks of firms, stakeholders and AI researchers [Newman, 2004; Schilling and Phelps, 2007], it is important to understand the effects of reciprocity and how it shapes the dynamics and global outcome of the development race. It is just as important to ensure that appropriate context-dependent regulatory actions are provided. This modelling approach and the associated results are applicable to other technologies and competitions, such as patent races or the development of biotechnology, pharmaceuticals, and climate change mitigation technology, where there is a significant advantage to be achieved by reaching some target first [Abbott et al., 2009; Burrell and Kelly, 2020; Campart and Pfister, 2014; Denicolò and Franzoni, 2010; Lemley, 2012]. Given a sufficiently tempting potential gain, individuals are more likely to invest in high-risk technology [Andrews et al., 2018], which suggests that these insights could be applicable to many similar fields in which risk and innovation must be constantly balanced.

It is noteworthy that, despite a number of proposals and debates on how to prevent, regulate, or resolve an AI race [Askell et al., 2019; Baum, 2017; Cave and ÓhÉigeartaigh, 2018; Geist, 2016; Han et al., 2019; Shulman and Armstrong, 2009; Taddeo and Floridi, 2018; Vinuesa et al., 2020], only a few formal modelling studies have been proposed [Armstrong et al., 2016; Han et al., 2020, 2022]. These works focus on homogeneous populations, where there are no inherent structures indicating the network of contacts among developing teams. Innovation dynamics (including AI) emerge from complex systems marked by a strong diversity in influence and companies' power. Firms create intricate networks of concurrent development, in which some develop a higher number of products, influencing and competing with a significant number of others. Our work advances this line of research, revealing the impact of these network structures among race participants, on the dynamics and global outcome of the development race.

We began by validating the analytical results obtained as a baseline in a completely homogeneous population [Han et al., 2020], using extensive agent-based simulations. We then adopted a similar methodology to analyse the effects of gradually increasing network heterogeneity, equivalent to diversifying the connectivity and influence of the race participants. This was accomplished by studying square lattices, and later two types of scale-free networks with varying degrees of clustering, with and without normalised payoffs (i.e. wealth inequality). Our findings suggest that the race tensions previously found in homogeneous networks are lowered, but that this effect only occurs in the presence of a certain degree of relational heterogeneity. In other words, spatial complexity by itself is not sufficient for the expectation of tempering the need for regulatory actions. Amongst all the network types studied, we found that scale-free networks with high clustering are the least demanding in terms of regulatory need, closely followed by regular scale-free networks.

The questions of how network structures and diversity influence the outcomes of behavioural dynamics, or the roles of network reciprocity, have been studied extensively in many fields, including Computer Science, Physics, Evolutionary Biology and Economics [Ahuja, 2000; Han et al., 2018; Perc, 2019; Perc et al., 2013, 2017; Raghunandan and Subramanian, 2012; Ranjbar-Sahraei et al., 2014; Santos et al., 2006a, 2008; Szabó and Fáth, 2007]. Network reciprocity can promote the evolution of positive behaviours in various settings including cooperation dilemmas [Perc et al., 2017; Ranjbar-Sahraei et al., 2014; Santos et al., 2006a, 2008], fairness [Page et al., 2000; Szolnoki et al., 2012; Wu et al., 2013] and trust [Kumar et al., 2020]. Their applications are diverse, ranging from healthcare [Newman, 2004], to network interference and influence maximization [Bloembergen et al., 2014; Cimpeanu et al., 2019; Wilder et al., 2018a], and to climate change [Santos and Pacheco, 2011]. The present work contributes new insights to this literature by studying the role of network reciprocity in the context of a technology development race. This strategy scenario is more intricate than the above-mentioned game theoretical scenarios (i.e., cooperation, trust and fairness) because, on the one hand, whether a social dilemma arises (where a collectively desired behaviour is not selected by evolutionary dynamics) depends on external factors (e.g., risk probability p_r in the early regime and monitoring probability p_{fo} in the late regime) [Han et al., 2020]. On the other hand, the collectively desired behaviour in the arisen social dilemma is different depending on the time-scale in which the race occurs. Interestingly, regardless of this more complex nature of the scenario, the different desirable behaviours can always be promoted in heterogeneous networks.

As an avenue of exploring the role of prominent players in the development race, we make use of a previously proposed model of studying the influence of nodes based on their degrees of connectivity [Santos et al., 2008]. These highly connected individuals have a tendency towards safety compliance in comparison to their counterparts. In an attempt to exploit this effect, as well as to better understand the impact of such seemingly significant nodes, we introduced several pathological players [Cardillo and Masuda, 2020; Kumar et al., 2020; Pacheco and Santos, 2011] in key locations of the network (highly connected nodes). We showed the role of hubs in slowing development and promoting safety, and argue that a small minority of influential developers can drastically reduce race tensions in almost all cases. The addition of pathological participants in these important locations can play a key role in the emergence of safety, without sacrificing innovation, and this effect is robust under uncertain race conditions. Our contribution explains the effects of heterogeneity in the networks that underlie the interactions between developers and teams of developers. We contend that there exist several ways in which this type of network heterogeneity could be promoted by relevant decision-makers, but argue that such mechanisms merit a dedicated body of research. Some examples of this could include dynamical linking [Pacheco et al., 2006b], whereby the relationship between two nodes could be altered by an outside decision-maker or the parties involved, or modifying the stakeholders' access to information, thereby amplifying selection dynamics [Tkadlec et al., 2021].

We note that our analyses focus on the binary extremes of developer behaviour, safe or unsafe development, in an effort to focus an already expansive problem into a manageable discussion. The addition of conditional, mixed, or random strategies could provide the basis for a novel piece of work. As observed with conditionally safe players in the well-mixed scenario [Han et al., 2020], we envisage that these additions would show little to no effect in the early regime, with the opposite being true for the late regime, at least in homogeneous settings.

In short, our results have shown that heterogeneous networks can significantly mediate the tensions observed in a well-mixed world, in both early and late development regimes [Han et al., 2020], thereby reducing the need for regulatory actions. Since a real-world network of contacts among technological firms and developers/researchers appears to be highly non-homogeneous, our findings provide important insights for the design of technological regulation and governance frameworks (such as the one proposed in the EU White Paper [European Commission, 2020]). Namely, the underlying structure of the relevant network (among developers and teams) needs to be carefully investigated to avoid for example unnecessary actions (i.e. regulating when that is not needed, as would have been otherwise suggested in homogeneous world models). Moreover, our findings suggest to increase heterogeneity or diversity in the network as a way to escape tensions arisen from a race for technological supremacy.

7 | Conclusions and General Discussion

Don’t adventures ever have an end? I suppose not.
Someone else always has to carry on the story.
—JRR Tolkien, *The Fellowship of the Ring*

In which we wrap up the contributions of this thesis, present a general discussion, and suggest topics for future work.

7.1 Summary of Conclusions

In the previous chapters, we extended the current models of external interference in various settings, explored the use of costly signals of the threat of punishment, and studied safety dynamics in the development of AI. In particular, we focused on how network characteristics influence the evolution of certain behaviours in social systems. Through large-scale computer simulations, we systematically analysed the emergent phenomena of these complex systems. We argue (as detailed in the following sections) that these findings provide important insights that could advise institutional policy in a wide range of settings. Keeping to the open-ended questions posed in Section 2.6, we summarise herewith the following conclusions:

- Social diversity, employed using heterogeneous networks of interaction, substantially influences the choice of investment approaches available to institutions. Investment is not trivial in these settings, contrary to previous findings in well-mixed and lattice populations. Counterintuitively, incentivising positive behaviour can lead to the exploitation of cooperators, harming pro-sociality in lieu of fostering it. Highly clustered scale-free networks make it easy to select the most effective candidates for receiving endowments. (Chapter 3)

- Global observations are typically less likely to yield optimal solutions to investment. In this sense, extensive information gathering, local observations, and stricter investment policies are often needed to reduce spending without sacrificing pro-social outcomes. Whether to promote fairness or cooperation, these findings remain robust across either social dilemma. (Chapters 3 and 4)

- Centrality in the network, which measures the influence of a node, does not usually serve as a reliable pathway towards promoting cooperation. In the context of fairness, influential individuals can be leveraged to reduce spending, but their reach is often not enough to ensure unanimously fair outcomes. (Chapters 3 and 4)

- In the Ultimatum Game, the opportunity to target multiple roles in the interactions further complicates decision-making and the distribution of endowments. Strictly targeting individuals who are fair in the role of proposers and of responders is remarkably conducive to ensuring fairness. Relaxing these standards requires extensive information gathering, whereby targeting fair proposers becomes a viable alternative. Social diversity simplifies decision-making, revealing novel approaches available to investors wishing to promote fairness. (Chapter 4)
- A higher propensity towards behavioural exploration (i.e. mutation rate) serves as an equaliser between the different roles in the Ultimatum Game. Thus, less specific investment schemes, targeting either fair responders or fair proposers, are often more cost-effective than strict approaches, which lead to over-spending. But importantly, no single approach is wholly robust across several mutation rates, highlighting their significance in the decision-making process of institutions seeking fairness. (Chapter 4)
- Proposing a novel mechanism of costly signalling, we show that fearful defectors can emerge through evolutionary dynamics when social punishment by itself would be ineffective. The signal acts as a deterrent to defection, creating a pathway towards cooperation and increasing social welfare. Furthermore, we extend this model to include the threat of institutional sanctions but find that the existence of fearful defectors hinges on social diversity. (Chapter 5)

In chapter 6, we proposed a timely application domain, that of safe AI, which readily ties in with the aforementioned models and findings. While still preliminary in scope, we have shown the positive effect that social diversity has in the speedy development of AI. When developers portray a strong diversity in terms of influence and connections, the tensions which exist in homogeneous populations are considerably diminished, thereby attenuating the need for regulatory action. Furthermore, our results evince that the design and implementation of meticulous interventions on a minority of participants can influence an entire population towards an ethical and sustainable use of advanced technology.

We believe that these findings, and the computational methods which enabled them, are important contributions for the reasons detailed below.

7.2 Usefulness

There is a common aphorism stating that *all models are wrong, but some are useful* [Box, 1976]. Indeed, models exist only as abstract representations of reality, and scrutinising their unbroken validity is perhaps not the appropriate line of inquiry. Then, why are our models – and by extension, this thesis – useful? We believe that there exist several such reasons, and we outline some of them below:

1. **Shed new light on ways to engineer pro-social behaviour**: The first reason for which we believe our models to be useful relates not to the evolutionary origins of cooperation but rather to the contexts where it has not emerged naturally. We have already seen that cooperation pervades at all scales of biological life but often the most valuable lessons can be learned when cooperation is lacking. Let us consider the very timely ongoing issues of global pandemics and climate change. The failures of institutional policies can cascade through society to produce immeasurable ill effects, fuelling disasters as opposed to mitigating them. Now more than ever, we are faced with catastrophic consequences arising from these mistakes – utterly inappropriate responses at the beginning of the COVID-19 pandemic, leading to mass contagion rampaging throughout the western world [Roberts, 2020]; and July of 2021 being the hottest month that the world has ever seen since the records began in 1880 [NOAA, 2021]. At the time of writing this thesis, these are ongoing issues, with no apparent end in sight. Thus, we believe our findings can provide useful insights into these very relevant problems, and serve as conceptual building blocks in the ongoing literature of designing appropriate institutional incentives.

2. **Provide a novel account of the implications of social diversity**: In this thesis, we place great emphasis on the underlying structures of interaction between individuals. Real-world networks of individuals are highly diverse in terms of connections and influence, which is not consistently mirrored in the literature. Many of these features can be captured by external decision-makers; we can exploit the world's patent heterogeneity in creative ways in the design of effective interventions used to leverage cooperation, fairness or safety. Beyond the obvious implications of our results, we also believe they are useful to understand which individuals are most able to steer their peerage towards a desirable outcome. Thus, inadvertently addressing not only which individuals would benefit the most from endowments but also which are in a position to influence others, simultaneously. This relationship is often complex and non-trivial, and helping some individuals can lead to the collapse of existing reciprocal structures, as we have seen in Chapter 3.

3. **Wherever possible, we aim to target preventive as opposed to punitive measures**: One relevant feature found throughout this thesis, encompassing all previous examples, is that we shy away from an explicit mechanism of punishment. There are several reasons for this: firstly, explicit punishment has been proven effective only in very specific settings [Dreber et al., 2008; Ohtsuki et al., 2009; Wu et al., 2009]. Secondly, punitive acts are inherently harmful, leading to a decrease in social welfare, and even counterproductive in certain settings, such as the aforementioned examples of pandemics and climate change. In fact, incentive schemes are tending away from explicit sanctions. To name a few specific examples: the pledges mechanism in the Paris agreement, the relatively recent obligation of reporting payment differences between men and women employees in the UK [The Economist, 2018], or the use of energy-efficiency labels (see also [Encarnação et al., 2016]). In all these cases, punitive measures are instead replaced with indirect sanctions (e.g. by consumers, voters or investors) based on public information, or by positive incentive schemes. Finally, in some cases sanctions are inherently unethical. For instance, we might look at the example of incentives to aid pregnant women in quitting smoking

[Adams et al., 2014; Bauld et al., 2017; Chamberlain et al., 2017]. While the serious long-term implications of tobacco consumption during pregnancy are a considerable motivator, this remains a very sensitive topic, and policy makers are restricted to the use of positive incentives to encourage certain health behaviours.

7.3 Applicability

As stated previously, one of the main reasons why we consider our work to be of interest is that it contributes to the literature on the evolution of cooperative behaviour among humans. While theoretical in nature, this domain is a fundamental question, irrespective of any endeavour of applying these findings to more specific scenarios beyond extending our understanding of societies and collective behaviour. In Chapter 6, we delve deeply into one such domain, studying safety dynamics in AI development, but below we list several other examples which could profit from our insights into institutional incentives:

- **Pro-social computing**: With the advent of autonomous technology, it is crucial to determine how best to engineer pro-social behaviour in hybrid societies of humans and machines [Akata et al., 2020; Paiva et al., 2018]. Engineering pro-sociality in such a context might hinge on exogenous agents or institutions able to engage in the distribution of incentives. For instance, we can envisage machines specifically designed to reward positive behaviour in machines with unrelated goals or even autonomous machines that can foster pro-sociality within human groups. This approach is of particular interest in large-scale populations with realistic networks of interaction, especially when a minority of meticulously engineered artificial agents may produce regime shifts towards desirable outcomes.
- **Biodiversity**: Several wildlife management organisations (e.g., the WWF) aim to maintain biodiversity in regions where anthropogenic factors lead to the extinction of species crucial to the ecosystem. In this context, the organisation is external to the ecosystems and has to

make careful decisions on how best to interfere. For instance, they might consider modifying the species composition of a community, deciding when and to what degree to interfere [Levin, 2000]. Since a more impactful intervention typically implies larger costs in terms of human resources and equipment, the organisation has to achieve a balance between cogent wildlife management and working within budget constraints. Moreover, due to the evolutionary dynamics of the ecosystem (e.g., frequency and structure dependence) [Hofbauer and Sigmund, 1998; Maynard Smith, 1982; Santos et al., 2006a], undesired behaviours can reoccur over time. Given this, the organisation also has to take into account the fact that it will have to repeatedly interfere in the ecosystem to sustain levels of biodiversity over time.

- **Risky innovation**: In Chapter 6, we study the fine balance between safe (thus slow) and unsafe (i.e. fast, innovative) technological development in a race towards AI supremacy. Notwithstanding, these findings apply broadly to other technologies and competitions, such as patent races or the development of biotechnology, pharmaceuticals, and climate change mitigation technology, where there is a significant advantage to be achieved by reaching some target first [Abbott et al., 2009; Burrell and Kelly, 2020; Campart and Pfister, 2014; Denicolò and Franzoni, 2010; Lemley, 2012]. Given a sufficiently tempting potential gain, individuals are more likely to invest in high-risk technology [Andrews et al., 2018], which suggests that these insights are robust across many similar fields in which risk and innovation must be constantly balanced.

- **Health behaviour**: There is a growing body of literature on the effectiveness of financial interventions used to encourage certain health behaviours, such as attending vaccinations or taking part in regular physical activity [Adams et al., 2014]. Closely related to this thesis is the example of smoking cessation during pregnancy [Bauld et al., 2017; Chamberlain et al., 2017]. Tobacco smoking remains one of the prevalent preventable factors associated with complications during and after pregnancy, for both the mother and their baby. As we stated previously, this is a particular example where negative incentives are unethical and counter-productive. What is more, the role of significant others has been identified as one of the key factors which could influence the success of such interventions [Bauld et al., 2017]. Thus, we argue that our findings would be of interest in this domain, given our focus on positive incentives and the effects of the networks which underlie these interactions.

7.4 Future Work

The models presented here can serve as a basis for new models of external interference in complex systems. The theoretical findings invite novel experimental works. For instance, we could test whether network dynamics have similar effects when studying real interactions between humans, either in the laboratory or in online platforms. Furthermore, we propose several research avenues for future work:

- **Machine learning heuristics**: The mechanisms we have proposed thus far have advanced the literature on external interference, but our contribution in this regard remains incipient. Using machine learning techniques, we could solve the bi-objective optimisation problem posed by cost-effective interference, therefore identifying solutions that are robust across a wider range of social dilemmas. Moreover, machine learning and other data-driven approaches could seamlessly integrate with current evolutionary game theoretic models. For instance, we could mine data on real-world networks and their features, thus being able to answer questions on how specific topology would impact the behaviour of agents in a particular application domain. One main problem of purely data-driven approaches is that they usually lack explainability, whereas game theoretic models are infinitely explainable but almost never very accurate representations of real data. Thus, merging the two fields together becomes crucially important for future work.

- **Episodic investments**: Often, we find that the continuous process of distributing endowments can lead to overspending. Certain approaches are initially promising but too demanding to perpetuate. By splitting evolutionary time further, e.g. by defining epochs, one could start asking further questions related to the timing of implementing certain schemes. One approach might be more suitable at the onset of a regime shift, while another might succeed in reducing costs once the population has stabilised to a desirable state. This approach could also be used to design adaptive interference schemes, which not only start and stop appropriately but are dynamic in terms of epochs or perceived benefits.

- **Costly signalling**: We have shown that costly signals can evolve alongside social punishment, and to some extent institutional punishment, serving as a deterrent to defection. Nevertheless, many questions remain, shrouding this fundamental mechanism that could undoubtedly explain the prevalence of punishment in human societies. In peer punishment, we have not yet explored the coexistence of punishers and signalling punishers nor fearful defectors with varying levels of fear; moreover, we have not yet changed the underlying network of interactions between individuals in this setting. Signalling the institutional threat of punishment appears not to measurably improve outcomes in homogeneous populations, thus more investigations are required to consolidate these observations. To conclude this point, signalling the promise of reward could be a potential avenue towards deterring defection, as well.

- **Governance in AI**: We have already proposed that hubs serve a key role in safety adoption within a race towards transformative AI. One key area that has not yet been systematically explored is an analysis of different incentive mechanisms going beyond simple reward and punishment schemes [Han et al., 2021], leveraging methods from network influence maximization [Bloembergen et al., 2014; Cimpeanu et al., 2019; Wilder et al., 2018b], taxation [Endriss et al., 2011], as well as other novel mechanisms that take into account inherent wealth inequality and spatial complexity. Another critical matter is identifying which types of networks govern the real-world interactions between AI developers, as well as the behavioural tendencies that guide the actions of these players. Moreover, in this work we have not considered conditional safety behaviour as explored in the original AI race model [Han et al., 2020], nor more complex strategies that could naturally attract fine-grained approaches to regulation. A natural continuation would be to determine how various incentive mechanisms can be used efficiently in the treatment of unsafe AI development, to mine and analyse real-world data to calibrate the aforementioned models, and finally, to perform behavioural experiments studying how human participants behave when presented with an AI racing scenario.

References

Frederick M Abbott, Maurice Nelson Graham Dukes, and Graham Dukes. *Global pharmaceutical policy: ensuring medicines for tomorrow's world*. Edward Elgar Publishing, 2009.

Jean Adams, Emma L. Giles, Elaine McColl, and Falko F. Sniehotta. Carrots, sticks and health behaviours: a framework for documenting the complexity of financial incentive interventions to change health behaviours. *Health psychology review*, 8(3):286–295, 2014. ISSN 1743-7202. doi: 10.1080/17437199.2013.848410. URL https://pubmed.ncbi.nlm.nih.gov/25053215/.

Gautam Ahuja. Collaboration networks, structural holes, and innovation: A longitudinal study. *Administrative science quarterly*, 45(3):425–455, 2000.

George Ainslie. *Breakdown of will*. Cambridge University Press, 2001. ISBN 9780521596947.

Zeynep Akata, Dan Balliet, Maarten De Rijke, Frank Dignum, Virginia Dignum, Guszti Eiben, Antske Fokkens, Davide Grossi, Koen Hindriks, Holger Hoos, Hayley Hung, Catholijn Jonker, Christof Monz, Mark Neerincx, Frans Oliehoek, Henry Prakken, Stefan Schlobach, Linda Van Der Gaag, Frank Van Harmelen, Herke Van Hoof, Birna Van Riemsdijk, Aimee Van Wynsberghe, Rineke Verbrugge, Bart Verheij, Piek Vossen, and Max Welling. A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence. *Computer*, 53(08):18–28, aug 2020. ISSN 0018-9162. doi: 10.1109/MC.2020.2996587.

Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47, 2002.

Réka Albert, Hawoong Jeong, and Albert-László Barabási. Diameter of the World-Wide Web. *Nature*, 401(6749):130–131, 1999. ISSN 1476-4687. doi: 10.1038/43601. URL https://doi.org/10.1038/43601.

Benjamin Allen, Gabor Lippner, Yu Ting Chen, Babak Fotouhi, Naghmeh Momeni, Shing Tung Yau, and Martin A. Nowak. Evolutionary dynamics on any population structure. *Nature 2017 544:7649*, 544(7649):227–230, mar 2017. ISSN 1476-4687. doi: 10.1038/nature21723. URL https://www.nature.com/articles/nature21723.

Talbot M Andrews, Andrew W Delton, and Reuben Kline. High-risk high-reward investments to mitigate climate change. *Nature Climate Change 2018 8:10*, 8(10):890–894, 2018. ISSN 1758-6798. doi: 10.1038/s41558-018-0266-y. URL https://www.nature.com/articles/s41558-018-0266-y.

Tibor Antal, A Traulsen, H Ohtsuki, C E Tarnita, and M A Nowak. Mutation-selection equilibrium in games with multiple strategies. *J. Theor. Biol.*, 258: 614–622, 2009.

Masahiko Aoki. *Toward a comparative institutional analysis*. MIT Press, 2001. ISBN 9780262011877. URL https://mitpress.mit.edu/books/toward-comparative-institutional-analysis.

Stuart Armstrong, Kaj Sotala, and Seán S Ó hÉigeartaigh. The errors, insights and lessons of famous AI predictions–and what they mean for the future. *Journal of Experimental \& Theoretical Artificial Intelligence*, 26(3):317–342, 2014.

Stuart Armstrong, Nick Bostrom, and Carl Shulman. Racing to the precipice: a model of artificial intelligence development. *AI & SOCIETY*, 31(2): 201–206, 2016. doi: 10.1007/s00146-015-0590-y.

Amanda Askell, Miles Brundage, and Gillian Hadfield. The Role of Cooperation in Responsible AI Development. *arXiv preprint arXiv:1907.04534*, 2019.

Robert Axelrod. Effective Choice in the Prisoner's Dilemma. *Journal of Conflict Resolution*, 24(1):3–25, mar 1980. ISSN 0022-0027. doi: 10.1177/002200278002400101. URL https://doi.org/10.1177/002200278002400101.

Robert Axelrod. *The Evolution of Cooperation*. Basic Books, ISBN 0-465-02122-2, 1984.

Y Bachrach, E Elkind, R Meir, D Pasechnik, M Zuckerman, J Rothe, and J Rosenschein. The Cost of Stability in Coalitional Games. In *Algorithmic Game Theory*, volume 5814 of *LNCS*, pages 122–134. Springer Berlin Heidelberg, 2009. ISBN 978-3-642-04644-5. doi: 10.1007/978-3-642-04645-2_12.

Tina Balke and Daniel Villatoro. Operationalization of the Sanctioning Process in Utilitarian Artificial Societies. In *Proceedings of the 2011 International Conference on Coordination, Organizations, Institutions, and Norms in Agent System VII*, COIN@AAMAS/WI-IAT'11, pages 167–185, Berlin, Heidelberg, 2011. Springer-Verlag. ISBN 9783642355448.

doi: 10.1007/978-3-642-35545-5_10. URL https://doi.org/10.1007/978-3-642-35545-5{_}10.

Albert-László Barabási. Scale-Free Networks: A Decade and Beyond. *Science*, 325(5939):412–413, jul 2009. doi: 10.1126/science.1173299. URL https://doi.org/10.1126/science.1173299.

Albert-Laszlo Barabasi. *Linked-how Everything is Connected to Everything Else and what it Means F*. Perseus Books Group, 2014.

Albert-László Barabási. *Network Science*. Cambridge University Press, 2016. ISBN 1107076269. URL https://books.google.com/books/about/Network{_}Science.html?id=iLtGDQAAQBAJ.

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

Alain Barrat and Romualdo Pastor-Satorras. Rate equation approach for correlations in growing network models. *Physical Review E*, 71(3):36127, 2005.

Scott Barrett and Astrid Dannenberg. Climate negotiations under scientific uncertainty. *Proceedings of the National Academy of Sciences of the United States of America*, 109(43):17372–17376, oct 2012. ISSN 00278424. doi: 10.1073/PNAS.1208417109/-/DCSUPPLEMENTAL. URL https://www.pnas.org/content/109/43/17372https://www.pnas.org/content/109/43/17372.abstract.

Chris T. Bauch and David J.D. Earn. Vaccination and the theory of games. *Proceedings of the National Academy of Sciences*, 101(36):13391–13394, sep 2004. ISSN 0027-8424. doi: 10.1073/PNAS.0403823101. URL https://www.pnas.org/content/101/36/13391https://www.pnas.org/content/101/36/13391.abstract.

Linda Bauld, Hilary Graham, Lesley Sinclair, Kate Flemming, Felix Naughton, Allison Ford, Jennifer McKell, Dorothy McCaughan, Sarah Hopewell, Kathryn Angus, Douglas Eadie, and David Tappin. Barriers to and facilitators of smoking cessation in pregnancy and following childbirth: literature review and qualitative study. *Health technology assessment (Winchester, England)*, 21(36):V–158, 2017. ISSN 2046-4924. doi: 10.3310/HTA21360. URL https://pubmed.ncbi.nlm.nih.gov/28661375/.

Seth D Baum. On the promotion of safe and socially beneficial artificial intelligence. *AI \& Society*, 32(4):543–551, 2017. doi: 10.1007/s00146-016-0677-0.

Monica Bianchini, Marco Gori, and Franco Scarselli. Inside pagerank. *ACM Transactions on Internet Technology (TOIT)*, 5(1):92–128, 2005.

Francis Bloch, Matthew O Jackson, and Pietro Tebaldi. Centrality measures in networks. *Available at SSRN 2749124*, 2019.

Daan Bloembergen, Bijan Ranjbar Sahraei, Haitham Bou-Ammar, Karl Tuyls, and Gerhard Weiss. Influencing Social Networks: An Optimal Control Study. In *ECAI*, volume 14, pages 105–110, 2014.

Jérémy Boes and Frédéric Migeon. Self-organizing multi-agent systems for the control of complex systems. *Journal of Systems and Software*, 134:12–28, 2017. ISSN 0164-1212. doi: https://doi.org/10.1016/j.jss.2017.08.038. URL http://www.sciencedirect.com/science/article/pii/S0164121217301838.

Paolo Boldi and Sebastiano Vigna. Axioms for centrality. *Internet Mathematics*, 10(3-4):222–262, 2014.

Phillip Bonacich. Some unique properties of eigenvector centrality. *Social networks*, 29(4):555–564, 2007.

Nick Bostrom. Strategic Implications of Openness in AI Development. *Global Policy*, 2017. doi: 10.1111/1758-5899.12403.

George E.P. Box. Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799, 1976. ISSN 1537274X. doi: 10.1080/01621459.1976.10480949.

Robert Boyd and Peter J Richerson. *Culture and the evolutionary process.* University of Chicago Press, Chicago, IL, US, 1985. ISBN 0-226-06931-1 (Hardcover); 0-226-06933-8 (Paperback).

Robert Boyd, Herbert Gintis, Samuel Bowles, and Peter J. Richerson. The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences*, 100(6):3531–3535, mar 2003. ISSN 00278424. doi: 10.1073/pnas.0630443100. URL http://dx.doi.org/10.1073/pnas.0630443100.

Anna D. Broido and Aaron Clauset. Scale-free networks are rare. *Nature Communications*, 10(1), dec 2019. ISSN 20411723. doi: 10.1038/S41467-019-08746-5. URL /pmc/articles/PMC6399239//pmc/articles/PMC6399239/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6399239/.

Rodney Brooks. The Seven Deadly Sins of Predicting the Future of AI, 2017. URL https://rodneybrooks.com/the-seven-deadly-sins-of-predicting-the-future-of-ai/.

Robert Burrell and Catherine Kelly. The COVID-19 pandemic and the challenge for innovation policy. *Available at SSRN 3576481*, 2020.

Sandy Campart and Etienne Pfister. Technological races and stock market value: evidence from the pharmaceutical industry. *Economics of Innovation and New Technology*, 23(3):215–238, 2014.

Alessio Cardillo and Naoki Masuda. Critical mass effect in evolutionary games triggered by zealots. *Physical Review Research*, 2(2), jun 2020. ISSN 2643-1564. doi: 10.1103/physrevresearch.2.023305. URL http://dx.doi.org/10.1103/PhysRevResearch.2.023305.

David Catteeuw, Bernard Manderick, and The Anh Han. Evolutionary Stability of Honest Signaling in Finite Populations. In *Proceedings of IEEE Congress on Evolutionary Computation*, pages 2864–2870. IEEE Computer Society, 2013. ISBN 9781479904549. doi: 10.1109/CEC.2013.6557917.

Stephen Cave and Seán S ÓhÉigeartaigh. An AI Race for Strategic Advantage: Rhetoric and Risks. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 36–40, 2018. ISBN 978-1-4503-6012-8. doi: 10.1145/3278721.3278780. URL http://doi.acm.org/10.1145/3278721.3278780.

Catherine Chamberlain, Alison O'Mara-Eves, Jessie Porter, Tim Coleman, Susan M. Perlen, James Thomas, and Joanne E. Mckenzie. Psychosocial interventions for supporting women to stop smoking in pregnancy. *Cochrane Database of Systematic Reviews*, 2017(2), feb 2017. ISSN 1469493X. doi: 10.1002/14651858.CD001055.PUB5/MEDIA/CDSR/CD001055/IMAGE_N/NCD001055-CMP-008-01.PNG. URL https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD001055.pub5/fullhttps://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD001055.pub5/abstract.

Xiaojie Chen and Matjaž Perc. Optimal distribution of incentives for public cooperation in heterogeneous interaction environments. *Frontiers in behavioral neuroscience*, 8:248, 2014.

Xiaojie Chen, Tatsuya Sasaki, Åke Brännström, and Ulf Dieckmann. First carrot, then stick: how the adaptive hybridization of incentives promotes cooperation. *Journal of the royal society interface*, 12(102):20140935, 2015.

Yann Chevaleyre, Paul E Dunne, Ulle Endriss, Jérôme Jerome Lang, Michel Lemaitre, Nicolas Maudet, Julian Padget, Steve Phelps, Juan A Rodr\'\igues-Aguilar, Paulo Sousa, Juan A Rodrguez-Aguilar, and Paulo Sousa. Issues in Multiagent Resource Allocation. *Informatica*, 30:3–31, 2006. ISSN 0350-5596.

T Cimpeanu and T A Han. Making an Example: Signalling Threat in the Evolution of Cooperation. In *2020 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8, 2020a. ISBN VO -. doi: 10.1109/CEC48606.2020.9185749.

Theodor Cimpeanu and The Anh Han. Fear of Punishment Promotes the Emergence of Cooperation and Enhanced Social Welfare in Social Dilemmas. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '20, pages 1819–1821, Richland, SC, 2020b. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450375184.

Theodor Cimpeanu, The Anh Han, and Francisco C Santos. Exogenous Rewards for Promoting Cooperation in Scale-Free Networks. In *ALIFE 2019*, pages 316–323. MIT Press, 2019.

Theodor Cimpeanu, Cedric Perret, and The Anh Han. Cost-efficient interventions for promoting fairness in the ultimatum game. *Knowledge-Based Systems*, 233:107545, 2021a.

Theodor Cimpeanu, Cedric Perret, and The Anh Han. Promoting Fair Proposers, Fair Responders or Both? Cost-Efficient Interference in the Spatial Ultimatum Game. In *In Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021)*, pages 1480–1482, 2021b.

Theodor Cimpeanu, Francisco C Santos, Luís Moniz Pereira, Tom Lenaerts, and The Anh Han. Artificial intelligence development races in heterogeneous settings. *Scientific Reports*, 12(1):1723, 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-05729-3. URL https://doi.org/10.1038/s41598-022-05729-3.

T H Clutton-Brock and Geoff Parker. Punishment in Animal Societies. *Nature*, 373:209–216, 1995. doi: 10.1038/373209a0.

David Collingridge. *The social control of technology*. New York : St. Martin's Press, 1980.

Ross Cressman, Jie Wen Song, Bo Yu Zhang, and Yi Tao. Cooperation and evolutionary dynamics in the public goods game with institutional incentives. *Journal of Theoretical Biology*, 299:144–151, apr 2012. ISSN 0022-5193. doi: 10.1016/J.JTBI.2011.07.030.

Ross Cressman, Jia-Jia Wu, Cong Li, and Yi Tao. Game Experiments on Cooperation Through Reward and Punishment. *Biological Theory 2013 8:2*, 8 (2):158–166, apr 2013. ISSN 1555-5550. doi: 10.1007/S13752-013-0106-2. URL https://link.springer.com/article/10.1007/s13752-013-0106-2.

J. A. Cuesta, R. Jiménez, H. Lugo, and A. Sánchez. The shared reward dilemma. *Journal of Theoretical Biology*, 251(2):253–263, mar 2008. ISSN 0022-5193. doi: 10.1016/J.JTBI.2007.11.022.

Luca Dall'Asta, Andrea Baronchelli, Alain Barrat, and Vittorio Loreto. Nonequilibrium dynamics of language games on complex networks. *Physical Review E*, 74(3):36105, 2006.

Charles Darwin. *The Origin Of Species*. Murray, 1911. ISBN 0451529065.

George Datseris, Ali R Vahdati, and Timothy C DuBois. Agents.jl: a performant and feature-full agent-based modeling software of minimal code complexity. *SIMULATION*, page 00375497211068820, jan 2022. ISSN 0037-5497. doi: 10.1177/00375497211068820. URL https://doi.org/10.1177/00375497211068820.

R M Dawes. Social Dilemmas. *http://dx.doi.org/10.1146/annurev.ps.31.020180.001125*, 31(1):169–193, nov 1980. ISSN 0066-4308. doi: 10.1146/ANNUREV.PS.31.020180.001125. URL https://www.annualreviews.org/doi/abs/10.1146/annurev.ps.31.020180.001125.

Richard Dawkins. *The Selfish Gene*. Oxford University Press, 1976. ISBN 9780191537554.

Dominique J F De Quervain, Urs Fischbacher, Valerie Treyer, Melanie Schellhammer, and Others. The neural basis of altruistic punishment. *Science*, 305(5688):1254, 2004.

Alexis de Tocqueville. *De la démocratie en Amérique*. Saunders and Otley, Paris, 1st edition, 1835.

Montreal Declaration. The Montreal Declaration for the Responsible Development of Artificial Intelligence Launched. https://www.canasean.com/the-montreal-declaration-for-the-responsible-development-of-artificial-intelligence-launched/, 2018.

Vincenzo Denicolò and Luigi A Franzoni. On the winner-take-all principle in innovation races. *Journal of the European Economic Association*, 8(5):1133–1158, 2010.

F Dercole, F Della Rossa, and C Piccardi. Direct reciprocity and model-predictive rationality explain network reciprocity over social ties. *Scientific reports*, 9(1):5367, 2019.

Alessandro Di Stefano, Marialisa Scatà, Aurelio La Corte, Pietro Liò, Emanuele Catania, Ermanno Guardo, and Salvatore Pagano. Quantifying the role of homophily in human cooperation using multiplex evolutionary game theory. *PloS one*, 10(10):e0140646, 2015.

Alessandro Di Stefano, Marialisa Scatà, Barbara Attanasio, Aurelio La Corte, Pietro Lió, and Sajal K Das. A Novel Methodology for designing Policies in Mobile Crowdsensing Systems. *Pervasive and Mobile Computing*, 67:101230, 2020.

Elias Fernández Domingos, Jelena Grujić, Juan C. Burguillo, Georg Kirchsteiger, Francisco C. Santos, and Tom Lenaerts. Timing Uncertainty in Collective Risk Dilemmas Encourages Group Reciprocation and Polarization. *iScience*, 23(12):101752, dec 2020. ISSN 2589-0042. doi: 10.1016/J.ISCI.2020.101752.

S Dorogovtsev. *Complex networks*. Oxford: Oxford University Press, 2010.

Sergey N Dorogovtsev, José Fernando F Mendes, and Alexander N Samukhin. Structure of growing networks with preferential linking. *Physical review letters*, 85(21):4633, 2000.

Sergey N Dorogovtsev, Jos F F Mendes, and Alexander N Samukhin. Size-dependent degree distribution of a scale-free growing network. *Physical Review E*, 63(6):62101, 2001.

Maria R. D'Orsogna and Matjaž Perc. Statistical physics of crime: A review. *Physics of Life Reviews*, 12:1–21, mar 2015. ISSN 1571-0645. doi: 10.1016/J.PLREV.2014.11.001.

Miguel dos Santos, Daniel Rankin, and Claus Wedekind. The evolution of punishment through reputation. *Proceedings. Biological sciences / The Royal Society*, 278:371–377, 2011. doi: 10.1098/rspb.2010.1275.

Anna Dreber, David G Rand, Drew Fudenberg, and Martin A Nowak. Winners don't punish. *Nature*, 452(7185):348–351, 2008.

Manh Hong Duong and The Anh Han. On Equilibrium Properties of the Replicator–Mutator Equation in Deterministic and Random Games. *Dynamic Games and Applications*, pages 1–23, 2019.

Manh Hong Duong and The Anh Han. Statistics of the number of equilibria in random social dilemma evolutionary games with mutation. *European Physical Journal B*, 2021a.

Manh Hong Duong and The Anh Han. Cost efficiency of institutional incentives in finite populations. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 2021b.

Émile Durkheim. *The Rules of Sociological Method. Edited with an introduction by Steven Lukes; translated by W. D. Halls.* The Free Press, 1895. ISBN 978-0-02-907940-9.

Sara Encarnação, Fernando P. Santos, Francisco C. Santos, Vered Blass, Jorge M. Pacheco, and Juval Portugali. Paradigm shifts and the interplay between state, business and civil sectors. *Royal Society Open Science*, 3(12), dec 2016. ISSN 20545703. doi: 10.1098/RSOS.160753. URL https://royalsocietypublishing.org/doi/full/10.1098/rsos.160753.

U Endriss, S Kraus, J Lang, and M Wooldridge. Incentive Engineering for Boolean Games. *IJCAI '11*, pages 2602–2607, 2011.

European Commission. White paper on Artificial Intelligence – An European approach to excellence and trust. Technical report, European Commission, 2020. URL AccessedMay26,https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020{_}en.pdf.

Yinhai Fang, Tina P. Benko, Matjaž Perc, Haiyan Xu, and Qingmei Tan. Synergistic third-party rewarding and punishment in the public goods game. *Proceedings of the Royal Society A*, 475(2227), jul 2019. ISSN 14712946. doi: 10.1098/RSPA.2019.0349. URL https://royalsocietypublishing.org/doi/abs/10.1098/rspa.2019.0349.

Ernst Fehr and Urs Fischbacher. The Nature of Human Altruism. *Nature*, 425:785–791, 2003. doi: 10.1038/nature02043.

Ernst Fehr and Simon Gachter. Cooperation and Punishment in Public Goods Experiments. *American Economic Review*, 90(4):980–994, 2000. URL http://econpapers.repec.org/RePEc:aea:aecrev:v:90:y:2000:i:4:p:980-994.

Ernst Fehr and Simon Gachter. Altruistic punishment in humans. *Nature*, 415:137–140, 2002.

Ernst Fehr and Klaus M Schmidt. A theory of fairness, competition, and cooperation. *The quarterly journal of economics*, 114(3):817–868, 1999.

Michal Feldman and John Chuang. Overcoming free-riding behavior in peer-to-peer systems. *ACM SIGecom Exchanges*, 5(4):41–50, jul 2005. ISSN 1551-9031. doi: 10.1145/1120717.1120723. URL https://dl.acm.org/doi/abs/10.1145/1120717.1120723.

D Floreano, S Mitri, A Perez-Uribe, and L Keller. Evolution of altruistic robots. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5050 LNCS: 232–248, 2008.

Hans Föllmer. Random economies with many interacting agents. *Journal of Mathematical Economics*, 1(1):51–62, mar 1974. ISSN 0304-4068. doi: 10.1016/0304-4068(74)90035-4.

David M. Frank and Sahotra Sarkar. Group Decisions in Biodiversity Conservation: Implications from Game Theory. *PLOS ONE*, 5(5):e10688, 2010. ISSN 1932-6203. doi: 10.1371/JOURNAL.PONE.0010688. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0010688.

H Franks, N Griffiths, and A Jhumka. Manipulating convention emergence using influencer agents. *Journal of Autonomous Agents and Multi-Agent Systems (JAAMAS)*, 26(3):315–353, 2013.

H Franks, N Griffiths, and S S Anand. Learning Agent Influence in MAS with Complex Social Networks. *Journal of Autonomous Agents and Multi-Agent Systems (JAAMAS)*, 28(5):836–866, 2014.

Feng Fu, Daniel I. Rosenbloom, Long Wang, and Martin A. Nowak. Imitation dynamics of vaccination behaviour on social networks. *Proceedings of the Royal Society B: Biological Sciences*, 278(1702):42–49, jan 2011. ISSN 14712970. doi: 10.1098/RSPB.2010.1107. URL https://royalsocietypublishing.org/doi/abs/10.1098/rspb.2010.1107.

D Fudenberg and L A Imhof. Imitation processes with small mutations. *Journal of Economic Theory*, 131:251–262, 2005.

Drew Fudenberg and David K Levine. *The Theory of Learning in Games*, volume 1 of *MIT Press Books*. The MIT Press, dec 1998.

Future of Life Institute. Autonomous Weapons: An Open Letter from AI \& Robotics Researchers. Technical report, Future of Life Institute, Cambridge, MA, 2015.

Future of Life Institute. Lethal Autonomous Weapons Pledge. https://futureoflife.org/lethal-autonomous-weapons-pledge/, 2019.

Julián Garcia and Arne Traulsen. Evolution of coordinated punishment to enforce cooperation from an unbiased strategy space. *Journal of the Royal Society Interface*, 16(156):20190127, 2019.

Edward Moore Geist. It's already too late to stop the AI arms race: We must manage it instead. *Bulletin of the Atomic Scientists*, 72(5):318–321, 2016.

António R Góis, Fernando P Santos, Jorge M Pacheco, and Francisco C Santos. Reward and punishment in climate change dilemmas. *Sci. Rep.*, 9(1):1–9, 2019.

Avner Greif and David D Laitin. A Theory of Endogenous Institutional Change. *The American Political Science Review*, 98(4):633–652, jan 2004. ISSN 00030554, 15375943. URL http://www.jstor.org/stable/4145329.

Jelena Grujić and Tom Lenaerts. Do people imitate when making decisions? Evidence from a spatial Prisoner?s Dilemma experiment. *Royal Society open science*, 7(7):200618, 2020.

Özgür Gürerk, Bernd Irlenbusch, and Bettina Rockenbach. The competitive advantage of sanctioning institutions. *Science*, 312(5770):108–111, apr 2006. ISSN 00368075. doi: 10.1126/SCIENCE.1123633/SUPPL_FILE/GURERK.SOM.PDF. URL https://www.science.org/doi/abs/10.1126/science.1123633.

Werner Güth, Rolf Schmittberger, and Bernd Schwarze. An experimental analysis of ultimatum bargaining. *Journal of economic behavior \& organization*, 3(4):367–388, 1982.

T A Han, L M Pereira, and F C Santos. The emergence of commitments and cooperation. In *AAMAS'2012*, pages 559–566, 2012.

T A Han, L M Pereira, F C Santos, and T Lenaerts. Good Agreements Make Good Friends. *Scientific reports*, 3(2695), 2013.

The Anh Han. *Intention Recognition, Commitments and Their Roles in the Evolution of Cooperation: From Artificial Intelligence Techniques to Evolutionary Game Theory Models*, volume 9. Springer SAPERE series, 2013. ISBN 978-3-642-37511-8.

The Anh Han. Emergence of Social Punishment and Cooperation through Prior Commitments. In *Proceedings of the Conference of the American Association of Artificial Intelligence (AAAI'2016)*, pages 2494–2500, Phoenix, Arizona, USA, 2016.

The Anh Han and Long Tran-Thanh. Cost-effective external interference for promoting the evolution of cooperation. *Scientific reports*, 8(15997), 2018.

The Anh Han, Simon Lynch, Long Tran-Thanh, and Francisco C Santos. Fostering Cooperation in Structured Populations Through Local and Global Interference Strategies. In *IJCAI-ECAI'2018*, pages 289–295. AAAI Press, 2018.

The Anh Han, Lu\'\is Moniz Pereira, and Tom Lenaerts. Modelling and Influencing the AI Bidding War: A Research Agenda. In *Proceedings of the AAAI/ACM conference AI, Ethics and Society*, pages 5–11, 2019.

The Anh Han, Luis Moniz Pereira, Francisco C Santos, and Tom Lenaerts. To Regulate or Not: A Social Dynamics Analysis of an Idealised AI Race. *Journal of Artificial Intelligence Research*, 69:881–921, 2020.

The Anh Han, Lu\'\is Moniz Lu\'\is Moniz Luis Moniz Pereira, Tom Lenaerts, and Francisco C Santos. Mediating artificial intelligence developments through negative and positive incentives. *PloS one*, 16(1):e0244592, 2021. doi: 10.1371/journal.pone.0244592.

The Anh Han, Tom Lenaerts, Francisco C Santos, and Luis Moniz Pereira. Voluntary safety commitments provide an escape from over-regulation in AI development. *Technology in Society (In Press)*, 2022.

G Hardin. The tragedy of the commons. *Science*, 162:1243–1248, 1968.

John C Harsanyi. On the Rationality Postulates Underlying the Theory of Cooperative Games. *The Journal of Conflict Resolution*, 5(2):179–196, feb 1961. ISSN 00220027, 15528766. URL http://www.jstor.org/stable/172785.

C Hauert, A Traulsen, H Brandt, M A Nowak, and K Sigmund. Via freedom to coercion: The emergence of costly punishment. *Science*, 316:1905–1907, 2007.

Ch. Hauert. Fundamental Clusters in Spatial 2 x 2 Games. *Proceedings: Biological Sciences*, 268(1468):761–769, feb 2001. ISSN 09628452. URL http://www.jstor.org/stable/3067624.

Christoph Hauert, Silvia De Monte, Josef Hofbauer, and Karl Sigmund. Volunteering as Red Queen Mechanism for Cooperation in Public Goods Games. *Science*, 296(5570):1129, 2002.

Joseph Henrich, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, Richard McElreath, Michael Alvard, Abigail Barr, Jean Ensminger, Natalie Smith Henrich, Kim Hill, Francisco Gil-White, Michael Gurven, Frank W. Marlowe, John Q. Patton, and David Tracer. "Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences*, 28(6):795–815, dec 2005. ISSN 0140-525X. doi: 10.1017/S0140525X05000142. URL https://www.cambridge.org/core/product/identifier/S0140525X05000142/type/journal{_}article.

Joan M. Herbers. Darwin's one special difficulty: celebrating Darwin 200. *Biology Letters*, 5(2):214–217, apr 2009. ISSN 1744957X. doi: 10.1098/RSBL.2009.0014. URL https://royalsocietypublishing.org/doi/abs/10.1098/rsbl.2009.0014.

Benedikt Herrmann, Christian Thöni, and Simon Gächter. Antisocial punishment across societies. *Science*, 319(5868):1362–1367, 2008. ISSN 00368075. doi: 10.1126/science.1153808.

Christian Hilbe and Arne Traulsen. Emergence of responsible sanctions without second order free riders, antisocial punishment or spite. *Scientific reports*, 2(458), 2012.

Chien-Ju Ho, Yu Zhang, Jennifer Wortman Vaughan, and Mihaela van der Schaar. Towards Social Norm Design for Crowdsourcing Markets. *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, jul 2012. URL https://www.aaai.org/ocs/index.php/WS/AAAIW12/paper/view/5295.

J Hofbauer and K Sigmund. *Evolutionary Games and Population Dynamics*. Cambridge University Press, 1998.

Gonglin Hou, Fei Wang, Jieyan Shi, Weijiang Chen, and Jie Yu. Which is the ideal sanction for cooperation? An experimental study on different types of third-party sanctions. *PsyCh Journal*, 8(2):212–231, jun 2019. ISSN 2046-0260. doi: 10.1002/PCHJ.259. URL https://onlinelibrary.wiley.com/doi/full/10.1002/pchj.259https://onlinelibrary.wiley.com/doi/abs/10.1002/pchj.259https://onlinelibrary.wiley.com/doi/10.1002/pchj.259.

Simon Huttegger, Brian Skyrms, Pierre Tarrès, and Elliott Wagner. Some dynamics of signaling games. *Proc Natl Acad Sci U S A.*, 111(3):10873–10880, 2014. ISSN 0027-8424. doi: 10.1073/pnas.1400838111.

Rasmus Ibsen-Jensen, Krishnendu Chatterjee, and Martin A. Nowak. Computational complexity of ecological and evolutionary spatial dynamics. *Proceedings of the National Academy of Sciences of the United States of America*, 112(51):15636–15641, dec 2015. ISSN 10916490. doi: 10.1073/PNAS.1511366112/-/DCSUPPLEMENTAL. URL https://www.pnas.org/content/112/51/15636https://www.pnas.org/content/112/51/15636.abstract.

Nicholas R Jennings, Peyman Faratin, Alessio R Lomuscio, Simon Parsons, Michael J Wooldridge, and Carles Sierra. Automated negotiation: prospects, methods and challenges. *Group decision and negotiation*, 10(2): 199–215, 2001.

Raúl Jiménez, Haydee Lugo, José A. Cuesta, and Angel Sánchez. Emergence and resilience of cooperation in the spatial prisoner's dilemma via a reward mechanism. *Journal of Theoretical Biology*, 250(3):475–483, feb 2008. ISSN 0022-5193. doi: 10.1016/J.JTBI.2007.10.010.

Raúl Jiménez, José A. Cuesta, Haydée Lugo, and Angel Sánchez. The shared reward dilemma on structured populations. *Journal of Economic Interaction and Coordination 2009 4:2*, 4(2):183–193, apr 2009. ISSN 1860-7128. doi: 10.1007/S11403-009-0053-Y. URL https://link.springer.com/article/10.1007/s11403-009-0053-y.

Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, pages 1–11, 2019.

Takafumi Kanazawa, Yasuhiko Fukumoto, Toshimitsu Ushio, and Takurou Misaka. Replicator dynamics with Pigovian subsidy and capitation tax. *Nonlinear Analysis: Theory, Methods & Applications*, 71(12):e818–e826, dec 2009. ISSN 0362-546X. doi: 10.1016/J.NA.2008.11.072.

J. Kiley Hamlin, Karen Wynn, Paul Bloom, and Neha Mahajan. How infants and toddlers react to antisocial others. *Proceedings of the National Academy of Sciences of the United States of America*, 108(50):19931–19936, dec 2011. ISSN 00278424. doi: 10.1073/PNAS.1110306108/-/DCSUPPLEMENTAL. URL https://www.pnas.org/content/108/50/19931https://www.pnas.org/content/108/50/19931.abstract.

Jack Knight. *Institutions and social conflict*. Cambridge University Press, Cambridge [England] ;;New York N.Y., 1992. ISBN 9780511528170.

Harold William Kuhn and Albert William Tucker. *Contributions to the Theory of Games (AM-28), Volume II*. Princeton University Press, dec 1953. doi: 10.1515/9781400881970/HTML.

Aanjaneya Kumar, Valerio Capraro, and Matjaž Perc. The evolution of trust and trustworthiness. *Journal of the Royal Society Interface*, 17(169): 20200491, 2020.

Robert Kurzban, Peter DeScioli, and Erin M O'Brien. Audience Effects on Moralistic Punishment. In *Evolution and Human Behavior*, volume 28, pages 75–84. Elsevier, mar 2007. doi: 10.1016/J.EVOLHUMBEHAV.2006.06.001.

Joung Hun Lee, Karl Sigmund, Ulf Dieckmann, and Yoh Iwasa. Games of corruption: How to suppress illegal logging. *Journal of Theoretical Biology*, 367:1–13, feb 2015. ISSN 10958541. doi: 10.1016/J.JTBI.2014.10.037.

Mark A Lemley. The myth of the sole inventor. *Michigan Law Review*, pages 709–760, 2012.

J-S. Jean-Sébastien Lerat, The Anh Han, and Tom Lenaerts. Evolution of Common-Pool Resources and Social Welfare in Structured Populations. In *Proceedings of the 23nd international joint conference on Artificial intelligence (IJCAI'2013)*, pages 2848–2854. AAAI Press, AAAI Press, 2013.

Simon A Levin. Multiple scales and the maintenance of biodiversity. *Ecosystems*, 3(6):498–506, 2000.

V Levit, T Grinshpoun, A Meisels, and A L C Bazzan. Taxation search in boolean games. In *AAMAS '13, Saint Paul, MN, USA, May 6-10, 2013*, pages 183–190, 2013.

Kristian Lindgren and Mats G Nordahl. Evolutionary dynamics of spatial games. *Physica D: Nonlinear Phenomena*, 75(1-3):292–309, 1994. ISSN 0167-2789. doi: DOI:10.1016/0167-2789(94)90289-5. URL http://www.sciencedirect.com/science/article/B6TVK-46TY4NH-P/2/227286bbabbfe4c920ded5850d87f133.

Sean Luke, Claudio Cioffi-Revilla, Liviu Panait, Keith Sullivan, and Gabriel Balan. MASON: A Multiagent Simulation Environment. *SIMULATION*, 81 (7):517–527, jul 2005. ISSN 0037-5497. doi: 10.1177/0037549705058073. URL https://doi.org/10.1177/0037549705058073.

M W Macy and A Flache. Learning dynamics in social dilemmas. *Proceedings of the National Academy of Sciences of the United States of America*, 99:7229–7236, 2002.

Stephen J. Majeski. Arms races as iterated prisoner's dilemma games. *Mathematical Social Sciences*, 7(3):253–266, jun 1984. ISSN 0165-4896. doi: 10.1016/0165-4896(84)90022-2.

Julia Marton-Lafevre and Others. *Peace parks: conservation and conflict resolution*. Mit Press, 2007.

David Masad and Jacqueline Kazil. Mesa: An Agent-Based Modeling Framework. In *Proceedings of the 14th Python in Science Conference*, pages 51–58. SciPy, 2015. doi: 10.25080/MAJORA-7B98E3ED-009. URL https://www.youtube.com/watch?v=lcySLoprPMc.

J Maynard Smith. *Evolution and the Theory of Games*. Cambridge University Press, Cambridge, 1982.

J. Maynard Smith and G. R. Price. The Logic of Animal Conflict. *Nature 1973 246:5427*, 246(5427):15–18, 1973. ISSN 1476-4687. doi: 10.1038/246015a0. URL https://www.nature.com/articles/246015a0.

Alex McAvoy and Christoph Hauert. Asymmetric evolutionary games. *PLoS Comput Biol*, 11(8):e1004349, 2015.

Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S Corrado, Ara Darzi, Mozziyar Etemadi, Florencia Garcia-Vicente, Fiona J Gilbert, Mark Halling-Brown, Demis Hassabis, Sunny Jansen, Alan Karthikesalingam, Christopher J Kelly, Dominic King, Joseph R Ledsam, David Melnick, Hormuz Mostofi, Lily Peng, Joshua Jay Reicher, Bernardino Romera-Paredes, Richard Sidebottom, Mustafa Suleyman, Daniel Tse, Kenneth C Young, Jeffrey De Fauw, and Shravya Shetty. International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788): 89–94, 2020. ISSN 1476-4687. doi: 10.1038/s41586-019-1799-6. URL https://doi.org/10.1038/s41586-019-1799-6.

Tomasz Michalak, Joanna Tyrowicz, Peter McBurney, and Michael Wooldridge. Exogenous coalition formation in the e-marketplace based on geographical proximity. *Electronic Commerce Research and Applications*, 8 (4):203–223, 2009.

Richard E. Michod. Evolution of individuality during the transition from unicellular to multicellular life. *Proceedings of the National Academy of Sciences*, 104(suppl 1):8613–8618, may 2007. ISSN 0027-8424. doi: 10.1073/PNAS.0701489104. URL https://www.pnas.org/content/104/suppl{_}1/8613https://www.pnas.org/content/104/suppl{_}1/8613.abstract.

John Miller and Scott Page. *Complex Adaptive Systems: An Introduction to Computational Models of Social Life*. Princeton University Press, 2009. URL https://press.princeton.edu/books/paperback/9780691127026/complex-adaptive-systems.

Melanie Mitchell. *Complexity: A Guided Tour* . Oxford University Press, USA, 2009. ISBN 0195124413. URL https://books.google.com/books/about/Complexity.html?id=sSgzHayrDBsC.

Roger B. Myerson. *Game Theory, Rationality, and Intelligence*. Harvard University Press, 1997. ISBN 0674341163.

John F. Nash. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, 36(1):48–49, jan 1950. ISSN 0027-8424. doi: 10.1073/PNAS.36.1.48. URL https://www.pnas.org/content/36/1/48https://www.pnas.org/content/36/1/48.abstract.

Randolf M Nesse. Natural selection and the capacity for subjective commitment. In Randolf M Nesse, editor, *Evolution and the capacity for commitment*, pages 1–44. New York: Russell Sage, 2001.

Mark Newman. *Networks, 2nd edition,*. Oxford university press, 2018.

Mark E J Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

Mark E J Newman. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the national academy of sciences*, 101(suppl 1): 5200–5205, 2004.

Mark E J Newman. The mathematics of networks. *The new palgrave encyclopedia of economics*, 2(2008):1–12, 2008.

Walter. Nicholson. *Intermediate microeconomics and its application*. Dryden Press, 2000. ISBN 0030259169. URL https://books.google.com/books/about/Intermediate{_}Microeconomics{_}and{_}Its{_}Appl.html?id=cxe3AAAAIAAJ.

NOAA. State of the Climate: Global Climate Report for Annual 2021. Technical report, NOAA National Centers for Environmental Information, 2021. URL https://www.ncdc.noaa.gov/sotc/global/202113.

Douglass C. North. Institutions, Institutional Change and Economic Performance. *Institutions, Institutional Change and Economic Performance*, oct 1990. doi: 10.1017/CBO9780511808678.

Douglass C North. Institutions. *Journal of Economic Perspectives*, 5(1):97–112, mar 1991. doi: 10.1257/jep.5.1.97. URL https://www.aeaweb.org/articles?id=10.1257/jep.5.1.97.

M A Nowak. *Evolutionary Dynamics: Exploring the Equations of Life*. Harvard University Press, Cambridge, MA, 2006a.

M A Nowak and K Sigmund. Tit for tat in heterogeneous populations. *Nature*, 355:250–253, 1992.

M. A. Nowak and K. Sigmund. Evolution of indirect reciprocity by image scoring. *Nature 1998 393:6685*, 393(6685):573–577, jun 1998. ISSN 1476-4687. doi: 10.1038/31225. URL https://www.nature.com/articles/31225.

M A Nowak, A Sasaki, C Taylor, and D Fudenberg. Emergence of cooperation and evolutionary stability in finite populations. *Nature*, 428:646–650, 2004.

Martin A Nowak. Five Rules for the Evolution of Cooperation. *Science*, 314 (5805):1560, 2006b.

Martin A Nowak. Evolving cooperation. *Journal of Theoretical Biology*, 299: 1–8, 2012. ISSN 00225193. doi: 10.1016/j.jtbi.2012.01.014. URL http://dx.doi.org/10.1016/j.jtbi.2012.01.014.

Martin A Nowak and Robert M May. Evolutionary games and spatial chaos. *Nature*, 359(6398):826–829, 1992.

Martin A. Nowak, Sebastian Bonhoeffer, and Robert M. May. Spatial games and the maintenance of cooperation. *Proceedings of the National Academy of Sciences of the United States of America*, 91(11):4877–4881, may 1994. ISSN 0027-8424. doi: 10.1073/PNAS.91.11.4877. URL https://pubmed.ncbi.nlm.nih.gov/8197150/.

Martin A Nowak, Karen M Page, and Karl Sigmund. Fairness versus reason in the ultimatum game. *Science*, 289(5485):1773–1775, 2000.

Ndidi Bianca Ogbo, Aiman Elgarig, and The Anh Han. Evolution of Coordination in Pairwise and Multi-player Interactions via Prior Commitments. *Adaptive Behavior (In Press)*, 2021.

Hisashi Ohtsuki, Christoph Hauert, Erez Lieberman, and Martin A Nowak. A simple rule for the evolution of cooperation on graphs and social networks. *Nature*, 441(7092):502–505, 2006. ISSN 0028-0836. doi: 10.1038/nature04605. URL http://www.nature.com/doifinder/10.1038/nature04605.

Hisashi Ohtsuki, Martin A Nowak, and Jorge M Pacheco. Breaking the symmetry between interaction and replacement in evolutionary dynamics on graphs. *Physical review letters*, 98(10):108106, 2007.

Hisashi Ohtsuki, Yoh Iwasa, and Martin A Nowak. Indirect reciprocity provides only a narrow margin of efficiency for costly punishment. *Nature*, 457(7601):79–82, 2009.

Cathy O'neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, 2016.

Toby Ord. *The Precipice: Existential Risk and the Future of Humanity*. Bloomsbury Publishing, 2020. ISBN 526600218.

Martin J Osborne. *An introduction to game theory*. Oxford University Press, 2004.

Elinor Ostrom. *Governing the commons: The evolution of institutions for collective action*. Cambridge university press, 1990.

Elinor Ostrom. Polycentric systems for coping with collective action and global environmental change. *Global Environmental Change*, 20(4):550–557, oct 2010. ISSN 09593780. doi: 10.1016/j.gloenvcha.2010.07.004.

J M Pacheco, F C Santos, and F A C C Chalub. Stern-Judging: A Simple, Successful Norm Which Promotes Cooperation under Indirect Reciprocity. *PLoS Comput. Biol.*, 2(12):e178, 2006a.

Jorge M Pacheco and Francisco C Santos. The messianic effect of pathological altruism. *Pathological altruism.*, page 300, 2011.

Jorge M Pacheco, Arne Traulsen, and Martin A Nowak. Coevolution of Strategy and Structure in Complex Networks with Dynamical Linking. *Phys. Rev. Lett.*, 97(25):258103, dec 2006b. doi: 10.1103/PhysRevLett.97.258103. URL https://link.aps.org/doi/10.1103/PhysRevLett.97.258103.

Jorge M. Pacheco, Vítor V. Vasconcelos, and Francisco C. Santos. Climate change governance, cooperation and self-organization. *Physics of Life Reviews*, 11(4):573–586, dec 2014. ISSN 1571-0645. doi: 10.1016/J.PLREV.2014.02.003.

Jorge M Pacheco, V\'\itor V Vasconcelos, Francisco C Santos, and Brian Skyrms. Co-evolutionary Dynamics of Collective Action with Signaling for a Quorum. *PLOS Computational Biology*, 11(2):1–12, 2015. doi: 10.1371/journal.pcbi.1004101. URL https://doi.org/10.1371/journal.pcbi.1004101.

Karen M Page, Martin A Nowak, and Karl Sigmund. The spatial ultimatum game. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 267(1458):2177–2182, 2000.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

Ana Paiva, Fernando P Francisco C Santos, and Fernando P Francisco C Santos. Engineering pro-sociality with autonomous agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pages 7994–7999, 2018.

Dennis Pamlin and Stuart Armstrong. Global challenges: 12 risks that threaten human civilization. *Global Challenges Foundation, Stockholm*, 2015.

Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):3200, 2001.

Alexandra S Penn, Richard A Watson, Alexander Kraaijeveld, and Jeremy Webb. Systems Aikido-A Novel Approach to Managing Natural Systems. In *in Proc. of the ALIFE XII Conference*, pages 577–580. MIT press, 2010.

Elizabeth Pennisi. How did cooperative behavior evolve. *Science*, 309 (5731):93, jul 2005. ISSN 00368075. doi: 10.1126/SCIENCE.309.5731.93/ASSET/F0DAEBDC-D718-4F85-AE7B-C9F4A2F36325/ASSETS/

SCIENCE.309.5731.93.FP.PNG. URL https://www.science.org/doi/abs/10.1126/science.309.5731.93.

Matjaž Perc. The social physics collective. *Scientific reports*, 2019.

Matjaž Perc and Attila Szolnoki. Coevolutionary games—a mini review. *BioSystems*, 99(2):109–125, 2010.

Matjaž Perc, Jesús Gómez-Gardenes, Attila Szolnoki, Luis M Flor\'\ia, and Yamir Moreno. Evolutionary dynamics of group interactions on structured populations: a review. *Journal of the royal society interface*, 10(80):20120997, 2013.

Matjaž Perc, Jillian J Jordan, David G Rand, Zhen Wang, Stefano Boccaletti, and Attila Szolnoki. Statistical physics of human cooperation. *Physics Reports*, 687:1–51, 2017.

Matjaž Perc, Mahmut Ozer, and Janja Hojnik. Social and juristic challenges of artificial intelligence. *Palgrave Communications*, 5(1):1–7, 2019.

Nicola Perra and Santo Fortunato. Spectral centrality measures in complex networks. *Physical Review E*, 78(3):36107, 2008.

Flávio L. Pinheiro, Jorge M. Pacheco, and Francisco C. Santos. From Local to Global Dilemmas in Social Networks. *PLOS ONE*, 7(2):e32114, feb 2012a. ISSN 1932-6203. doi: 10.1371/JOURNAL.PONE.0032114. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0032114.

Flavio L Pinheiro, Francisco C Santos, and Jorge M Pacheco. How selection pressure changes the nature of social dilemmas in structured populations. *New Journal of Physics*, 14(7):73035, 2012b.

Flávio L. Pinheiro, Francisco C. Santos, and Jorge M. Pacheco. Linking Individual and Collective Behavior in Adaptive Social Networks. *Physical review letters*, 116(12), mar 2016. ISSN 1079-7114. doi: 10.1103/PHYSREVLETT.116.128702. URL https://pubmed.ncbi.nlm.nih.gov/27058108/.

Julia Poncela, Jesús Gómez-Gardenes, L M Flor\'\ia, and Yamir Moreno. Robustness of cooperation in the evolutionary prisoner's dilemma on complex networks. *New Journal of Physics*, 9(6):184, 2007.

William Poundstone. *Prisoner's dilemma*. Doubleday, 1992. ISBN 0385415672.

Simon T Powers, Daniel J Taylor, and Joanna J Bryson. Punishment can promote defection in group-structured populations. *Journal of theoretical biology*, 311:107–116, 2012.

Amy R Pritchett and Antoine Genton. Negotiated decentralized aircraft conflict resolution. *IEEE transactions on intelligent transportation systems*, 19(1):81–91, 2017.

M A Raghunandan and C A Subramanian. Sustaining cooperation on networks: an analytical study based on evolutionary game theory. In *AAMAS'12*, volume 12, pages 913–920. Citeseer, 2012.

Nichola. Raihani. *The Social Instinct : How Cooperation Shaped the World*. Jonathan Cape, 2021. ISBN 9781787332041.

Nichola Raihani and David Aitken. Uncertainty, rationality and cooperation in the context of climate change. *Climatic Change*, 108(1):47–55, 2011.

Nichola J Raihani and Redouan Bshary. The reputation of punishers. *Trends in ecology & evolution*, 30(2):98–103, 2015.

Nichola J. Raihani and Redouan Bshary. Punishment: one tool, many uses. *Evolutionary Human Sciences*, 1, 2019. ISSN 2513-843X. doi: 10.1017/EHS.2019.12. URL https://www.cambridge.org/core/journals/evolutionary-human-sciences/article/punishment-one-tool-many-uses/FD1940BB4D5A39D017A09D4C162B4D28.

Nichola J. Raihani, Alex Thornton, and Redouan Bshary. Punishment and cooperation in nature. *Trends in Ecology & Evolution*, 27(5):288–295, may 2012. ISSN 0169-5347. doi: 10.1016/J.TREE.2011.12.004.

David G Rand and Martin A Nowak. Human cooperation. *Trends in Cognitive Sciences*, 17(8):413–425, 2013. ISSN 1364-6613. doi: https://doi.org/10.1016/j.tics.2013.06.003. URL https://www.sciencedirect.com/science/article/pii/S1364661313001216.

David G Rand, Joseph J Armao IV, Mayuko Nakamaru, and Hisashi Ohtsuki. Anti-social punishment can prevent the co-evolution of punishment and cooperation. *Journal of Theoretical Biology*, 265(4):624–632, 2010.

David G Rand, Corina E Tarnita, Hisashi Ohtsuki, and Martin A Nowak. Evolution of fairness in the one-shot anonymous Ultimatum Game. *Proceedings of the National Academy of Sciences*, 110(7):2581–2586, 2013.

David G Rand, Martin A Nowak, James H Fowler, and Nicholas A Christakis. Static network structure can stabilize human cooperation. *Proc Natl Acad Sci USA*, 111(48):17093–17098, 2014.

Bijan Ranjbar-Sahraei, Haitham Bou Ammar, Daan Bloembergen, Karl Tuyls, Gerhard Weiss, B R Sahraei, H Bou-Ammar, Daan Bloembergen, Karl Tuyls, Gerhard Weiss, Bijan Ranjbar-Sahraei, Haitham Bou Ammar, Daan Bloembergen, Karl Tuyls, Gerhard Weiss, B R Sahraei, H Bou-Ammar, Daan Bloembergen, Karl Tuyls, and Gerhard Weiss. Evolution of cooperation in arbitrary complex networks. In *AAMAS '14, Paris, France, May 5-9,* *2014*, pages 677–684. International Foundation for Autonomous Agents and Multiagent Systems, 2014.

Daniel J Rankin, Miguel dos Santos, and Claus Wedekind. The evolutionary significance of costly punishment is still to be demonstrated. *Proceedings of the National Academy of Sciences*, 106(50):E135 LP – E135, dec 2009. ISSN 1091-6490. doi: 10.1073/pnas.0911990107. URL http://www.pnas.org/content/106/50/E135.abstracthttp://www.ncbi.nlm.nih.gov/pubmed/19995971http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2795507.

Paul Resnick, Richard Zeckhauser, John Swanson, and Kate Lockwood. The value of reputation on eBay: A controlled experiment. *Experimental Economics 2006 9:2*, 9(2):79–101, jun 2006. ISSN 1573-6938. doi: 10.1007/S10683-006-4309-2. URL https://link.springer.com/article/10.1007/s10683-006-4309-2.

Ernesto Reuben and Arno Riedl. Enforcement of contribution norms in public good games with heterogeneous populations. *Games and Economic Behavior*, 77(1):122–137, 2013. ISSN 0899-8256. doi: https://doi.org/10.1016/j.geb.2012.10.001. URL http://www.sciencedirect.com/science/article/pii/S0899825612001492.

Les Roberts. Institutional Failures in COVID-19 | Think Global Health, 2020. URL https://www.thinkglobalhealth.org/article/institutional-failures-covid-19.

David Rolnick, Priya L Donti, Lynn H Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, Alexandra Luccioni, Tegan Maharaj, Evan D Sherwin, S Karthik Mukkavilli, Konrad P Kording, Carla Gomes, Andrew Y Ng, Demis Hassabis, John C Platt, Felix Creutzig, Jennifer Chayes, and Yoshua Bengio. Tackling Climate Change with Machine Learning, 2019.

Stuart Russell, S Hauert, R Altman, and M Veloso. Ethics of artificial intelligence. *Nature*, 521(7553):415–416, 2015.

Stuart J. Russell. *Human Compatible : Artificial Intelligence and the Problem of Control*. Viking, 2019. ISBN 978-0-525-55861-3.

Alan G. Sanfey, James K. Rilling, Jessica A. Aronson, Leigh E. Nystrom, and Jonathan D. Cohen. The neural basis of economic decision-making in the Ultimatum Game. *Science (New York, N.Y.)*, 300(5626):1755–1758, jun 2003. ISSN 1095-9203. doi: 10.1126/SCIENCE.1082976. URL https://pubmed.ncbi.nlm.nih.gov/12805551/.

F C Santos and J M Pacheco. Scale-free networks provide a unifying framework for the emergence of cooperation. *Phys. Rev. Lett.*, 95:98104, 2005. ISSN 0031-9007.

F. C. Santos and J. M. Pacheco. A new route to the evolution of cooperation. *Journal of Evolutionary Biology*, 19(3):726–733, may 2006. ISSN 1010061X. doi: 10.1111/j.1420-9101.2005.01063.x.

F C Santos, J M Pacheco, and T Lenaerts. Evolutionary Dynamics of Social Dilemmas in Structured Heterogeneous Populations. *Proceedings of the National Academy of Sciences of the United States of America*, 103:3490–3494, 2006a. ISSN 0027-8424.

F C Santos, M D Santos, and J M Pacheco. Social diversity promotes the emergence of cooperation in public goods games. *Nature*, 454:214–216, 2008.

Fernando P Santos, Jorge M Pacheco, Ana Paiva, and Francisco C Santos. Structural power and the evolution of collective fairness in social networks. *PLoS one*, 12(4):e0175687, 2017.

Fernando P Santos, Jorge M Pacheco, and Francisco C Santos. Indirect Reciprocity and Costly Assessment in Multiagent Systems. In *Thirty-Second AAAI Conference on Artificial Intelligence*, pages 4727–4734, 2018a.

Fernando P Santos, Francisco C Santos, and Jorge M Pacheco. Social norm complexity and past reputations in the evolution of cooperation. *Nature*, 555(7695):242–245, 2018b.

Fernando P Francisco C Fernando P Santos, Jorge M Pacheco, Ana Paiva, and Fernando P Francisco C Fernando P Santos. Evolution of collective fairness in hybrid populations of humans and agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6146–6153, 2019.

Francisco C Santos and Jorge M Pacheco. Risk of collective failure provides an escape from the tragedy of the commons. *Proceedings of the National Academy of Sciences of the United States of America*, 108(26):10421–10425, 2011.

Francisco C. Santos, Jorge M. Pacheco, and Tom Lenaerts. Cooperation Prevails When Individuals Adjust Their Social Ties. *PLOS Computational Biology*, 2(10):e140, oct 2006b. ISSN 1553-7358. doi: 10.1371/JOURNAL.PCBI.0020140. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.0020140.

Francisco C. Santos, Jorge M. Pacheco, and Brian Skyrms. Co-evolution of pre-play signaling and cooperation. *Journal of Theoretical Biology*, 274(1): 30–35, apr 2011. ISSN 00225193. doi: 10.1016/j.jtbi.2011.01.004.

Francisco C Santos, Flavio L Pinheiro, Tom Lenaerts, and Jorge M Pacheco. The role of diversity in the evolution of cooperation. *Journal of theoretical biology*, 299:88–96, 2012a.

Francisco C. Santos, Vítor V. Vasconcelos, Marta D. Santos, P. N.B. Neves, and Jorge M. Pacheco. EVOLUTIONARY DYNAMICS OF CLIMATE CHANGE UNDER COLLECTIVE-RISK DILEMMAS. *https://doi.org/10.1142/S0218202511400045*, 22(SUPPL.1), apr 2012b. ISSN 02182025. doi: 10.1142/S0218202511400045.

Tatsuya Sasaki and Satoshi Uchida. Rewards and the evolution of cooperation in public good games. *Biology Letters*, 10(1):20130903, 2014. ISSN 1744957X. doi: 10.1098/RSBL.2013.0903. URL https://royalsocietypublishing.org/doi/abs/10.1098/rsbl.2013.0903.

Tatsuya Sasaki, Åke Brännström, Ulf Dieckmann, and Karl Sigmund. The take-it-or-leave-it option allows small penalties to overcome social dilemmas. *Proceedings of the National Academy of Sciences*, 109(4):1165–1169, 2012.

Bastin Tony Roy Savarimuthu, Maryam Purvis, Martin Purvis, and Stephen Cranefield. Social norm emergence in virtual agent societies. In *Declarative Agent Languages and Technologies VI*, pages 18–28. Springer, 2009.

Akrati Saxena and Sudarshan Iyengar. Centrality measures in complex networks: A survey. *arXiv preprint arXiv:2011.07190*, 2020.

Marialisa Scatà, Alessandro Di Stefano, Aurelio La Corte, Pietro Liò, Emanuele Catania, Ermanno Guardo, and Salvatore Pagano. Combining evolutionary game theory and network theory to analyze human cooperation patterns. *Chaos, Solitons \& Fractals*, 91:17–24, 2016.

Thomas C Schelling. *Micro Motives and Macro Behavior*. Norton, 1978. ISBN 0393090094. URL https://books.google.com/books/about/Micromotives{_}and{_}Macrobehavior.html?id=4C5mQgAACAAJ.

Matthew U Scherer. Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies. *SSRN Electronic Journal*, 2015. ISSN 0897-3393. doi: 10.2139/SSRN.2609777. URL https://papers.ssrn.com/abstract=2609777.

Melissa A Schilling and Corey C Phelps. Interfirm collaboration networks: The impact of large-scale network structure on firm innovation. *Management science*, 53(7):1113–1126, 2007.

Bruce Schneier. *Liars and Outliers: Enabling the Trust that Society Needs to Thrive*. Wiley, 2012.

Sarah Schoenmakers, Christian Hilbe, Bernd Blasius, and Arne Traulsen. Sanctions as honest signals–The evolution of pool punishment by public sanctioning institutions. *Journal of theoretical biology*, 2014.

R Selten. A note on evolutionarily stable strategies in asymmetric animal conflicts. *Journal of theoretical biology*, 84(1):93–101, may 1980. ISSN 0022-5193 (Print). doi: 10.1016/s0022-5193(80)81038-1.

Andrew Shipilov and Annabelle Gawer. Integrating research on interorganizational networks and ecosystems. *Academy of Management Annals*, 14(1): 92–121, 2020.

M. B. Short, M. R. D'Orsogna, V. B. Pasour, G. E. Tita, P. J. Brantingham, A. L. Bertozzi, and L. B. Chayes. A STATISTICAL MODEL OF CRIMINAL BEHAVIOR. *https://doi.org/10.1142/S0218202508003029*, 18(SUPPL.):1249–1267, nov 2011. ISSN 02182025. doi: 10.1142/S0218202508003029.

Carl Shulman and Stuart Armstrong. Arms control and intelligence explosions. In *7th European Conference on Computing and Philosophy (ECAP), Bellaterra, Spain, July*, pages 2–4, 2009.

K Sigmund, C Hauert, and M Nowak. Reward and punishment. *Proceedings of the National Academy of Sciences*, 98(19):10757–10762, 2001.

K Sigmund, H De Silva, A Traulsen, and C Hauert. Social learning promotes institutions for governing the commons. *Nature*, 466:7308, 2010.

Karl Sigmund. *The Calculus of Selfishness*. Princeton University Press, 2010.

R Sinatra, J Iranzo, J Gómez-Gardeñes, L M Floría, V Latora, and Y Moreno. The Ultimatum Game in complex networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(09):P09012, 2009. ISSN 1742-5468. doi: 10.1088/1742-5468/2009/09/p09012. URL http://dx.doi.org/10.1088/1742-5468/2009/09/P09012.

Brian Skyrms. *Signals: Evolution, Learning, and Information*. Oxford University Press, 2010.

Hannah M Smidt. United Nations peacekeeping locally: enabling conflict resolution, reducing communal violence. *Journal of Conflict Resolution*, 64 (2-3):344–372, 2020.

John Smith and Eors Szathmary. *The Major Transitions in Evolution*. Oxford University Press, 1995. ISBN 978-0-19-850294-4.

Kaj Sotala and Roman V Yampolskiy. Responses to catastrophic AGI risk: a survey. *Physica Scripta*, 90(1):18001, 2014.

Luc Steels and Ramon de Mantaras. The Barcelona declaration for the proper development and usage of artificial intelligence in Europe. *AI Communications*, pages 1–10, 2018.

John E Stewart. Evolutionary possibilities: Can a society be constrained so that “the good” self-organizes? *World Futures*, 74(1):1–35, oct 2017. ISSN 1556-1844. doi: 10.1080/02604027.2017.1357985. URL http://dx.doi.org/10.1080/02604027.2017.1357985.

Jessica Su, Aneesh Sharma, and Sharad Goel. The effect of recommendations on network structure. *25th International World Wide Web Conference, WWW 2016*, pages 1157–1167, 2016. doi: 10.1145/2872427.2883040.

Ussif Rashid Sumaila. A review of game-theoretic models of fishing. *Marine Policy*, 23(1):1–10, jan 1999. ISSN 0308-597X. doi: 10.1016/S0308-597X(97)00045-6.

György Szabó and Gábor Fáth. Evolutionary games on graphs. *Physics Reports*, 446(4-6):97–216, 2007. ISSN 03701573. doi: 10.1016/j.physrep.2007.04.004.

Attila Szolnoki, Matjaž Perc, and György Szabó. Defense mechanisms of empathetic players in the spatial ultimatum game. *Physical review letters*, 109(7):78701, 2012.

Mariarosaria Taddeo and Luciano Floridi. Regulate artificial intelligence to avert cyber arms race. *Nature*, 556(7701):296–298, 2018.

Alessandro Tavoni and Simon Levin. Managing the climate commons at the nexus of ecology, behaviour and economics. *Nature Climate Change 2014 4:12*, 4(12):1057–1063, nov 2014. ISSN 1758-6798. doi: 10.1038/nclimate2375. URL https://www.nature.com/articles/nclimate2375.

Andreia Sofia Teixeira, Fernando P Francisco C Fernando P Santos, Alexandre P Francisco, and Fernando P Francisco C Fernando P Santos. Eliciting Fairness in N-Player Network Games through Degree-Based Role Assignment. *Complexity*, 2021, 2021.

The Economist. Forcing employers to reveal their gender pay gaps is making them think | The Economist. *The Economist*, 2018. URL https://www.economist.com/britain/2018/04/07/forcing-employers-to-reveal-their-gender-pay-gaps-is-making-them-think.

Josef Tkadlec, Andreas Pavlogiannis, Krishnendu Chatterjee, and Martin A Nowak. Fast and strong amplifiers of natural selection. *Nature Communications 2021 12:1*, 12(1):1–6, 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-24271-w. URL https://www.nature.com/articles/s41467-021-24271-w.

A Traulsen, M A Nowak, and J M Pacheco. Stochastic Dynamics of Invasion and Fixation. *Phys. Rev. E*, 74:11909, 2006.

Arne Traulsen, Christoph Hauert, Hannelore De Silva, Martin A Nowak, and Karl Sigmund. Exploration dynamics in evolutionary games. *Proc. Natl. Acad. Sci. USA*, 106(3):709–712, 2009.

R L Trivers. The evolution of reciprocal altruism. *Quaterly Review of Biology*, 46:35–57, 1971.

Karl Tuyls, Julien Perolat, Marc Lanctot, Georg Ostrovski, Rahul Savani, Joel Z Leibo, Toby Ord, Thore Graepel, and Shane Legg. Symmetric decomposition of asymmetric games. *Scientific Reports*, 8(1):1–20, 2018.

Eric van Damme, Kenneth G. Binmore, Alvin E. Roth, Larry Samuelson, Eyal Winter, Gary E. Bolton, Axel Ockenfels, Martin Dufwenberg, Georg Kirchsteiger, Uri Gneezy, Martin G. Kocher, Matthias Sutter, Alan G. Sanfey, Hartmut Kliemt, Reinhard Selten, Rosemarie Nagel, and Ofer H. Azar. How Werner Güth's ultimatum game shaped our understanding of social behavior. *Journal of Economic Behavior & Organization*, 108:292–318, dec 2014. ISSN 0167-2681. doi: 10.1016/J.JEBO.2014.10.014.

Vitor V Vasconcelos, Francisco C Santos, and Jorge M Pacheco. A bottom-up institutional approach to cooperative governance of risky commons. *Nature Climate Change*, 3(9):797–801, 2013.

Daniel Villatoro, Giulia Andrighetto, Jordi Sabater-Mir, and Rosaria Conte. Dynamic sanctioning for robust and cost-efficient norm compliance. In *IJCAI*, volume 11, pages 414–419, 2011.

Ricardo Vinuesa, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone Langhans, Max Tegmark, and Francesco Fuso Nerini. The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications*, 11 (233), 2020.

John Von Neumann and Oskar Morgenstern. *Theory of games and economic behavior*. Princeton University Press, 1944. ISBN 978-0691130613.

Markus Waibel, Dario Floreano, and Laurent Keller. A Quantitative Test of Hamilton's Rule for the Evolution of Altruism. *PLOS Biology*, 9(5): e1000615, may 2011. ISSN 1545-7885. doi: 10.1371/JOURNAL.PBIO.1000615. URL https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1000615.

Shengxian Wang, Xiaojie Chen, and Attila Szolnoki. Exploring optimal institutional incentives for public cooperation. *Communications in Nonlinear Science and Numerical Simulation*, 79:104914, 2019.

Felix Warneken and Michael Tomasello. Helping and Cooperation at 14 Months of Age. *Infancy*, 11(3):271–294, may 2007. ISSN 1532-7078. doi: 10.1111/J.1532-7078.2007.TB00227.X. URL https://onlinelibrary.wiley.com/doi/full/10.1111/j.1532-7078.2007.tb00227.xhttps://onlinelibrary.wiley.com/doi/abs/10.1111/j.1532-7078.2007.tb00227.xhttps://onlinelibrary.wiley.com/doi/10.1111/j.1532-7078.2007.tb00227.x.

Takamitsu Watanabe, Masanori Takezawa, Yo Nakawake, Akira Kunimatsu, Hidenori Yamasue, Mitsuhiro Nakamura, Yasushi Miyashita, and Naoki Masuda. Two distinct neural mechanisms underlying indirect reciprocity. *Proceedings of the National Academy of Sciences of the United States of America*, 111(11):3990–3995, mar 2014. ISSN 10916490. doi: 10.1073/PNAS.1318570111/-/DCSUPPLEMENTAL. URL https://www.pnas.org/content/111/11/3990https://www.pnas.org/content/111/11/3990.abstract.

Bryan Wilder, Nicole Immorlica, Eric Rice, and Milind Tambe. Maximizing Influence in an Unknown Social Network. In *AAAI conference on Artificial Intelligence (AAAI-18)*, 2018a.

Bryan Wilder, Han-Ching Ou, Kayla de la Haye, and Milind Tambe. Optimizing Network Structure for Preventative Health. In *AAMAS*, pages 841–849, 2018b.

Uri Wilensky and William Rand. *An Introduction to Agent-Based Modeling: Modeling Natural, Social, and Engineered Complex Systems with NetLogo*. MIT Press, 2015. ISBN 9780262328128. URL https://ieeexplore.ieee.org/book/7109293.

Gerald S. Wilkinson. Reciprocal food sharing in the vampire bat. *Nature 1984 308:5955*, 308(5955):181–184, 1984. ISSN 1476-4687. doi: 10.1038/308181a0. URL https://www.nature.com/articles/308181a0.

M Wooldridge. Bad equilibria (and what to do about them). *ECAI '12*, pages 6–11, 2012.

Michael Wooldridge. *An Introduction to MultiAgent Systems [Paperback]*. John Wiley & Sons, 2009. ISBN 0470519460. URL https://www.google.com/books?hl=hr{&}lr={&}id=X3ZQ7yeDn2IC{&}oi=fnd{&}pg=PR13{&}dq=Wooldridge,+M.,+An+introduction+to+multiagent+systems.+2009:+John+Wiley+{%}26+Sons{&}ots=WFoevw8u54{&}sig=pW{_}SoLDnMqc6fkrAnJzFJIv8phIhttp://www.amazon.com/Introduction-MultiAgent-Systems-.

J J Wu, B Y Zhang, Z X Zhou, Q Q He, X D Zheng, R Cressman, and Y Tao. Costly punishment does not always increase cooperation. *Proc Natl Acad Sci U S A.*, 106(41):17448–17451, 2009.

Te Wu, Feng Fu, Yanling Zhang, and Long Wang. Adaptive role switching promotes fairness in networked ultimatum game. *Scientific reports*, 3:1550, 2013.

Lily Xu, Shahrzad Gholami, Sara Mc Carthy, Bistra Dilkina, Andrew Plumptre, Milind Tambe, Rohit Singh, Mustapha Nsubuga, Joshua Mabonga, Margaret Driciru, Fred Wanyama, Aggrey Rwetsiba, Tom Okello, and Eric Enyel. Stay Ahead of Poachers: Illegal Wildlife Poaching Prediction and Patrol Planning Under Uncertainty with Field Test Evaluations

(Short Version). *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, 2020-April:1898–1901, apr 2020. ISSN 10844627. doi: 10.1109/ICDE48307.2020.00198.

Paul J. Zak. *Moral Markets*. Princeton University Press, feb 2008. ISBN 9780691135236. URL http://www.jstor.org/stable/j.ctt7swzc.

Ioannis Zisis, Sibilla Di Guida, The Anh Han, Georg Kirchsteiger, and Tom Lenaerts. Generosity motivated by acceptance - evolutionary analysis of an anticipation games. *Scientific reports*, 5(18076), 2015.

A | Appendix A - Additional Results for Chapter 4

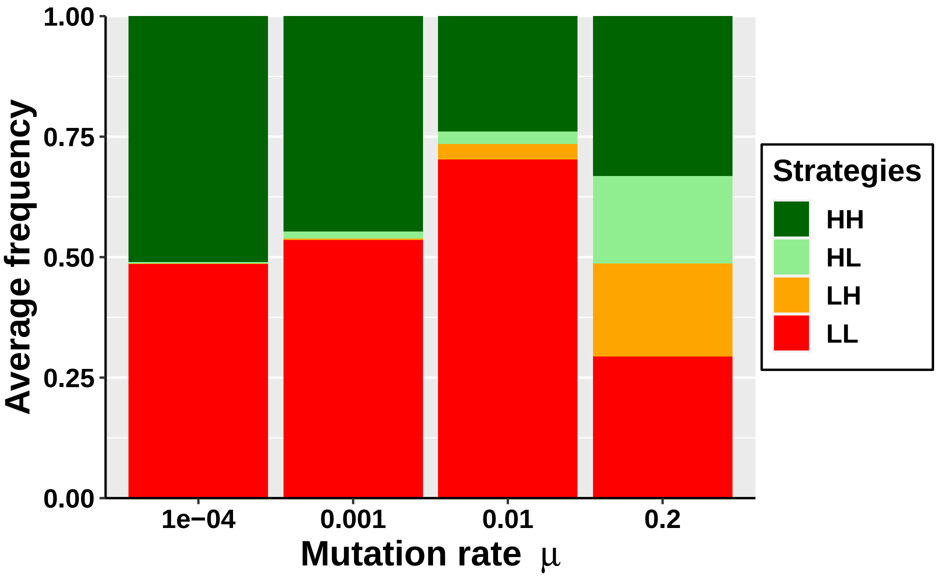

Fig. A.1 Average frequencies of the four strategies HH, HL, LH and LL as a function of mutation rate μ in absence of interference in SL populations.

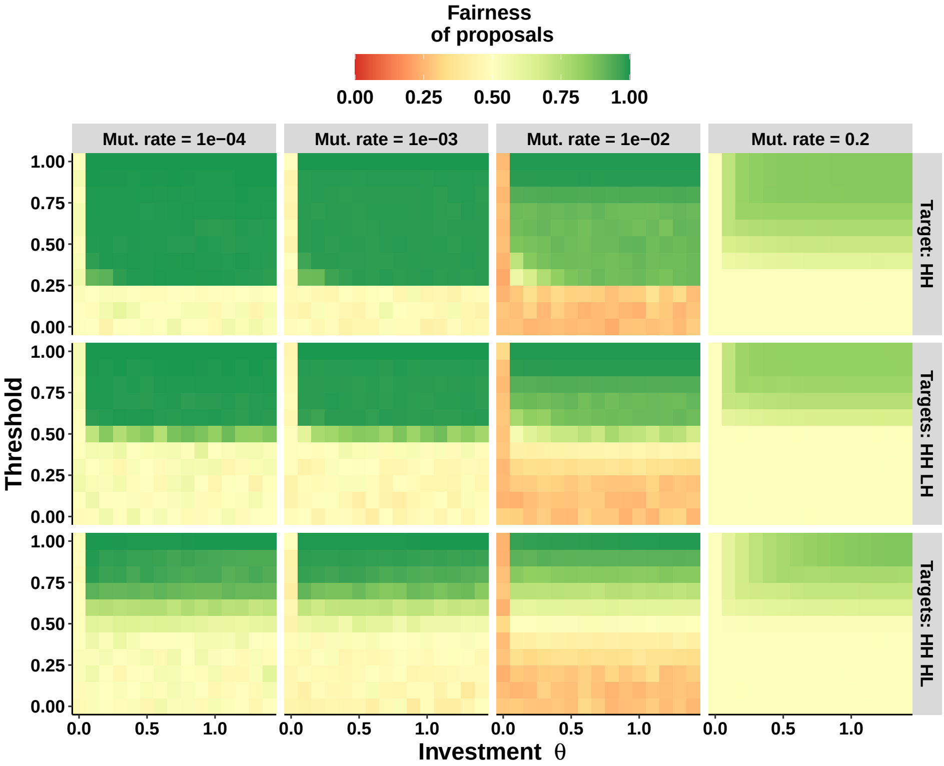

Fig. A.2 Average fairness as a function of the individual endowment θ, the threshold p_f and the mutation rate μ in SL populations (population-based, stochastic update). Each row represents a different targeting scheme.

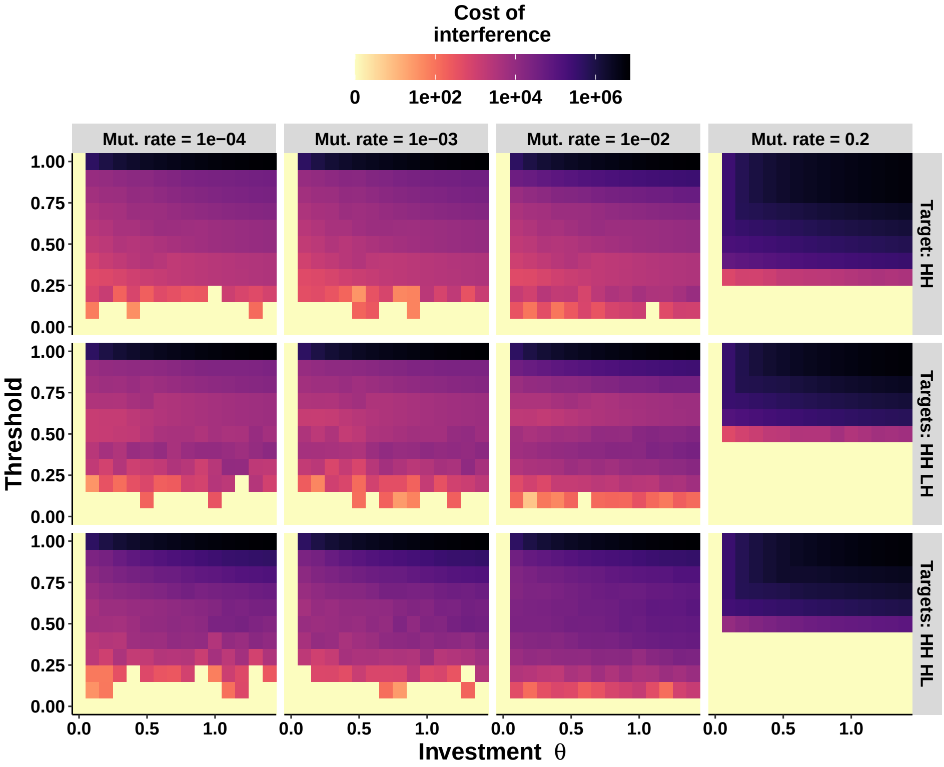

Fig. A.3 Average cost of interference as a function of the individual endowment θ, the threshold p_f and the mutation rate μ in SL populations (population-based, stochastic update). Each row represents a different targeting scheme. The cost of interference is shown on a logarithmic scale.

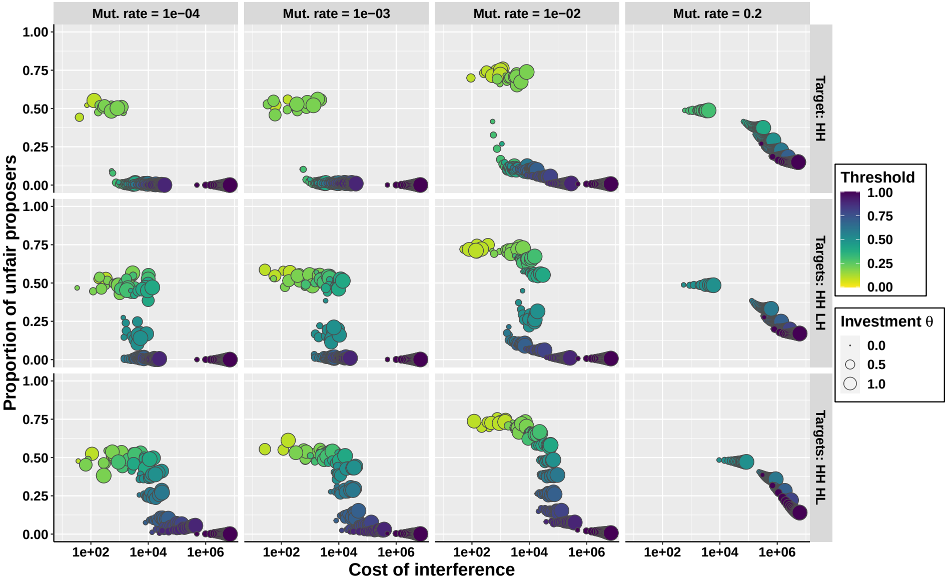

Fig. A.4 Proportion of unfair proposers as a function of average cost of interference for different targeting schemes and mutation rates μ in SL populations (population-based, stochastic update). The size and colour of the circles correspond to investment amount and threshold of investment, respectively. We note that the most desirable outcomes are closest to the origin.

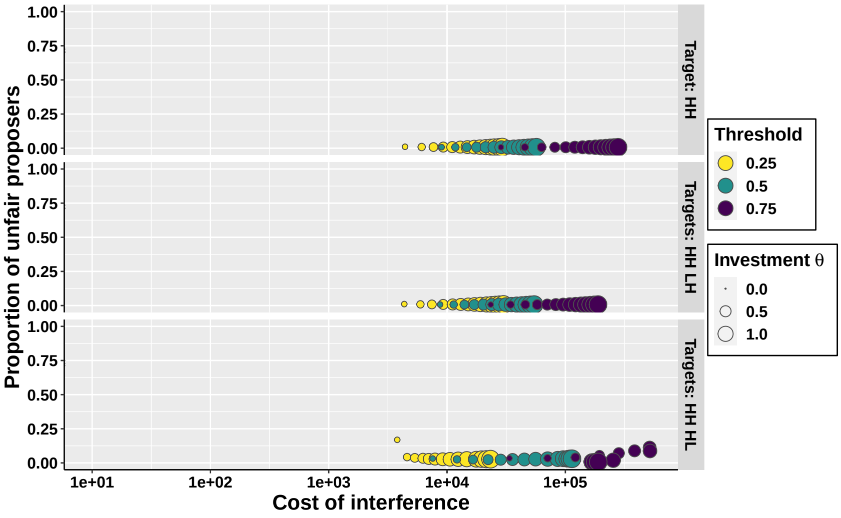

Fig. A.5 Proportion of unfair proposers as a function of average cost of interference for different targeting schemes in SL populations (neighbourhood-based, $\mu = 0.01$, stochastic update). The size and colour of the circles correspond to investment amount and threshold of investment, respectively. We note that the most desirable outcomes are closest to the origin.

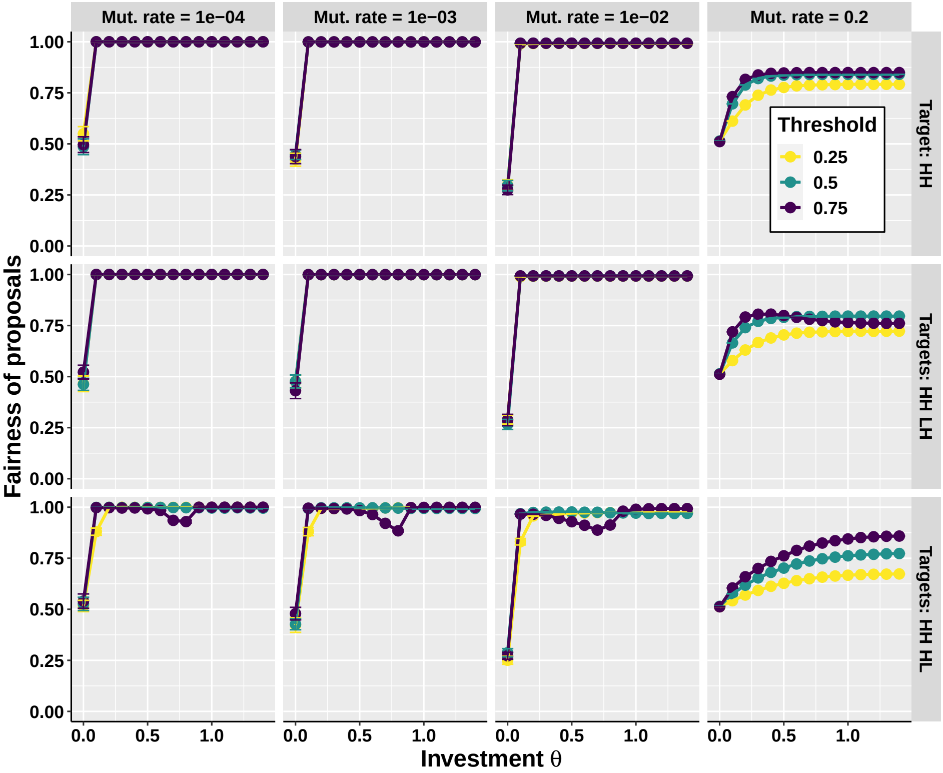

Fig. A.6 Average fairness measured by the sum of frequencies of HH and HL as a function of the individual endowment θ, the threshold p_f and the mutation rate μ in SL populations (neighbourhood-based, stochastic update). Each row represents a different targeting scheme.

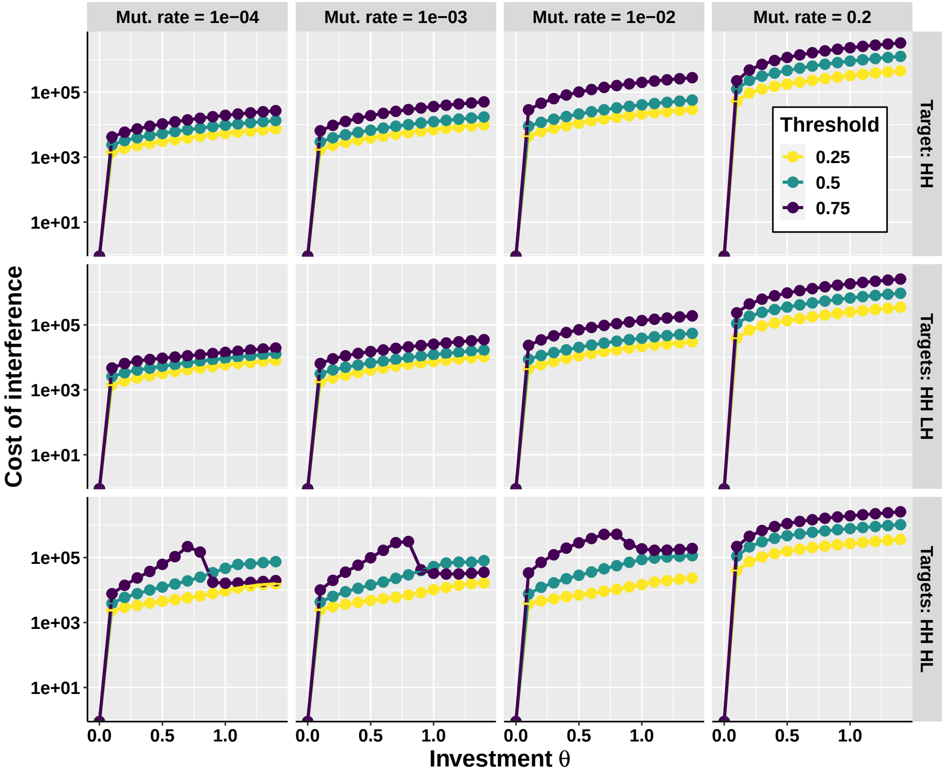

Fig. A.7 Average cost of interference as a function of the individual endowment θ, the threshold p_f and the mutation rate μ in SL populations (neighbourhood-based, stochastic update). Each row represents a different targeting scheme. The cost of interference is on a logarithmic scale for clarity.

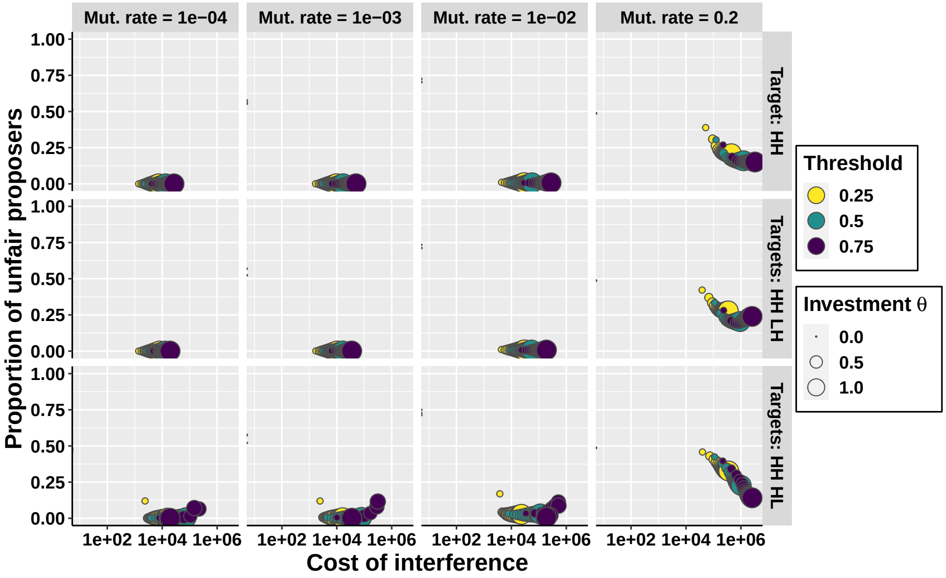

Fig. A.8 Proportion of unfair proposers as a function of average cost of interference for different targeting schemes and mutation rates μ in SL populations (neighbourhood-based, stochastic update). The size and colour of the circles correspond to investment amount and threshold of investment, respectively. We note that the most desirable outcomes are closest to the origin.

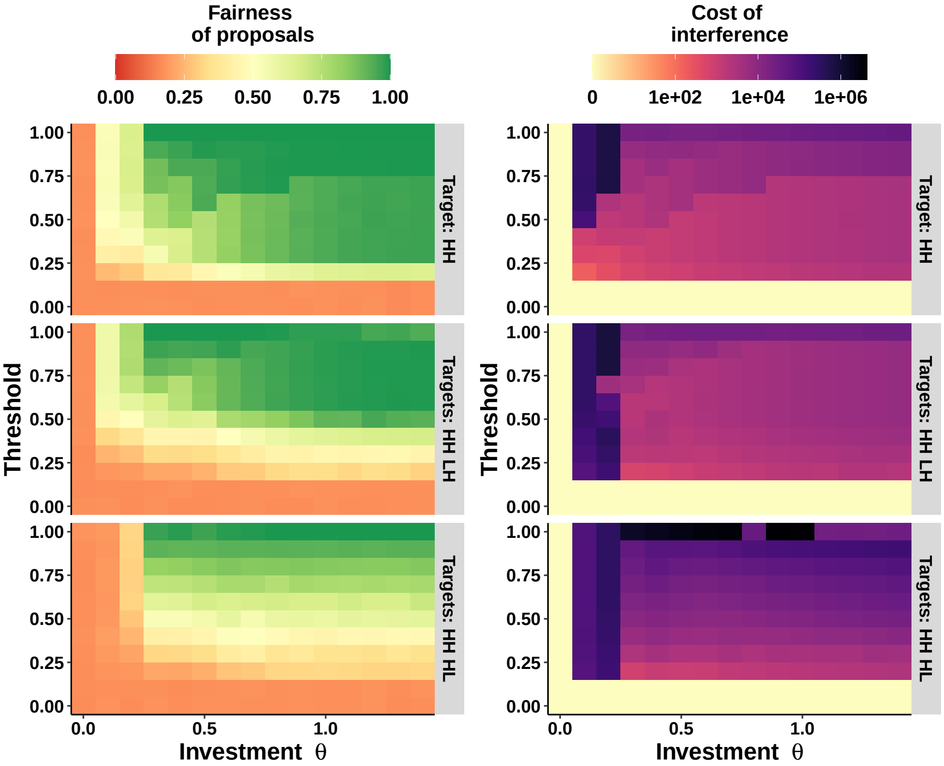

Fig. A.9 Average fairness (left) and average cost of interference (right) as a function of the individual endowment θ and the threshold p_f in SL populations (population-based, deterministic update). Each row represents a different targeting scheme. The cost of interference is shown on a logarithmic scale.

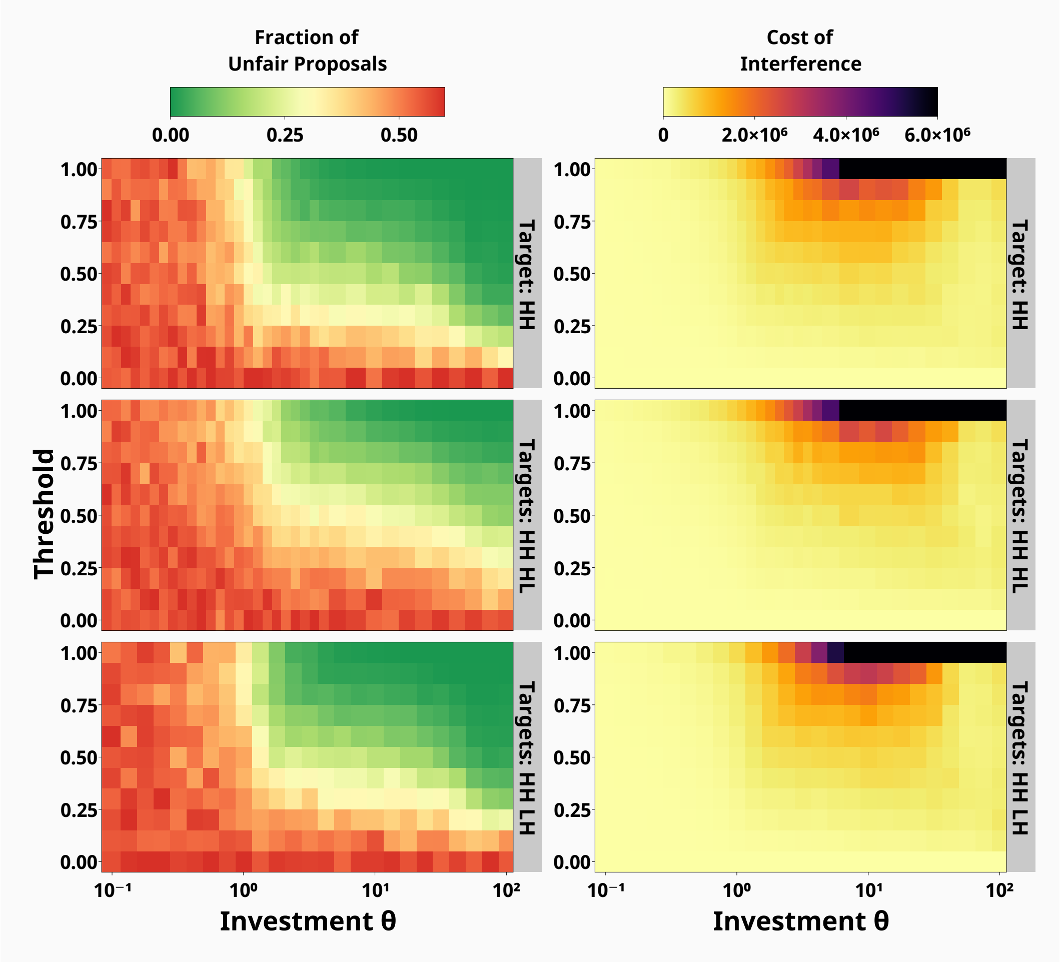

Fig. A.10 Proportion of unfair proposals and total costs of investment for POP-based interference, for all targets (BA networks).

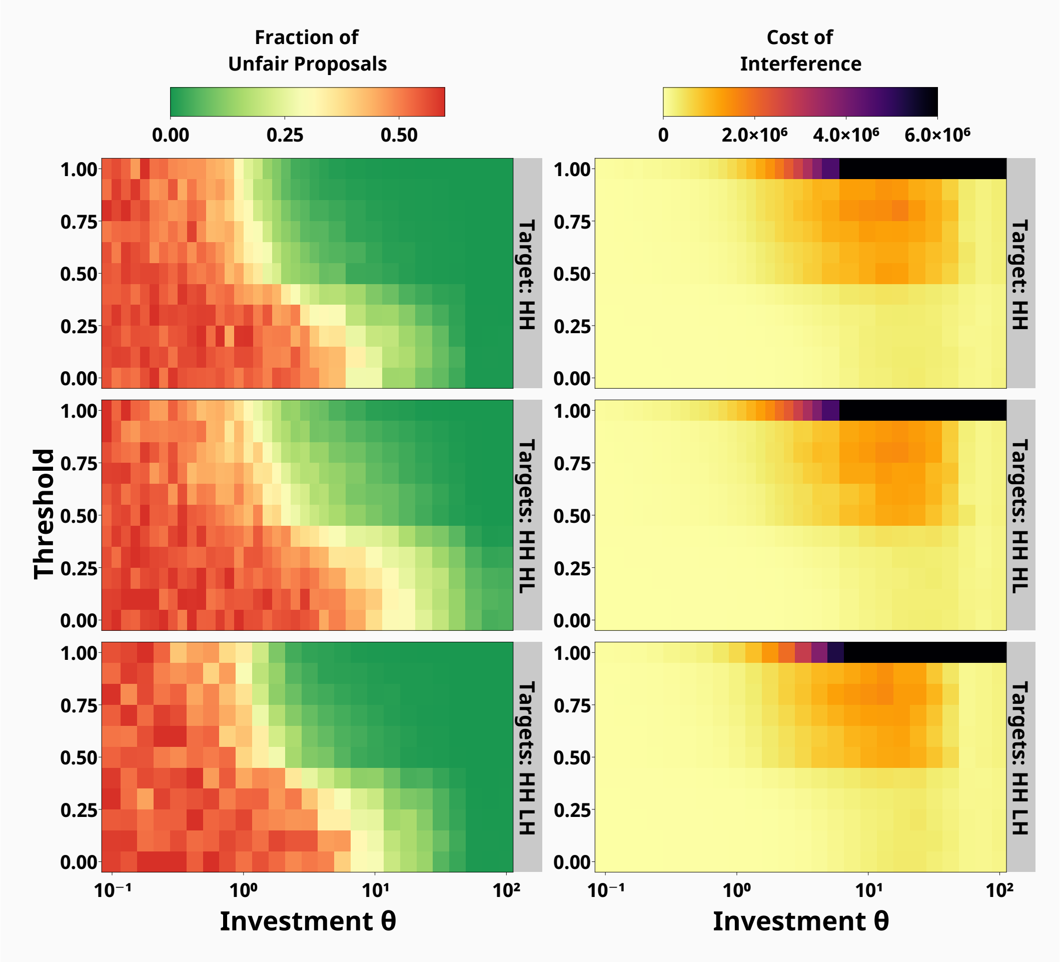

Fig. A.11 Proportion of unfair proposals and total costs of investment for NEB-based interference for all targets (BA networks).

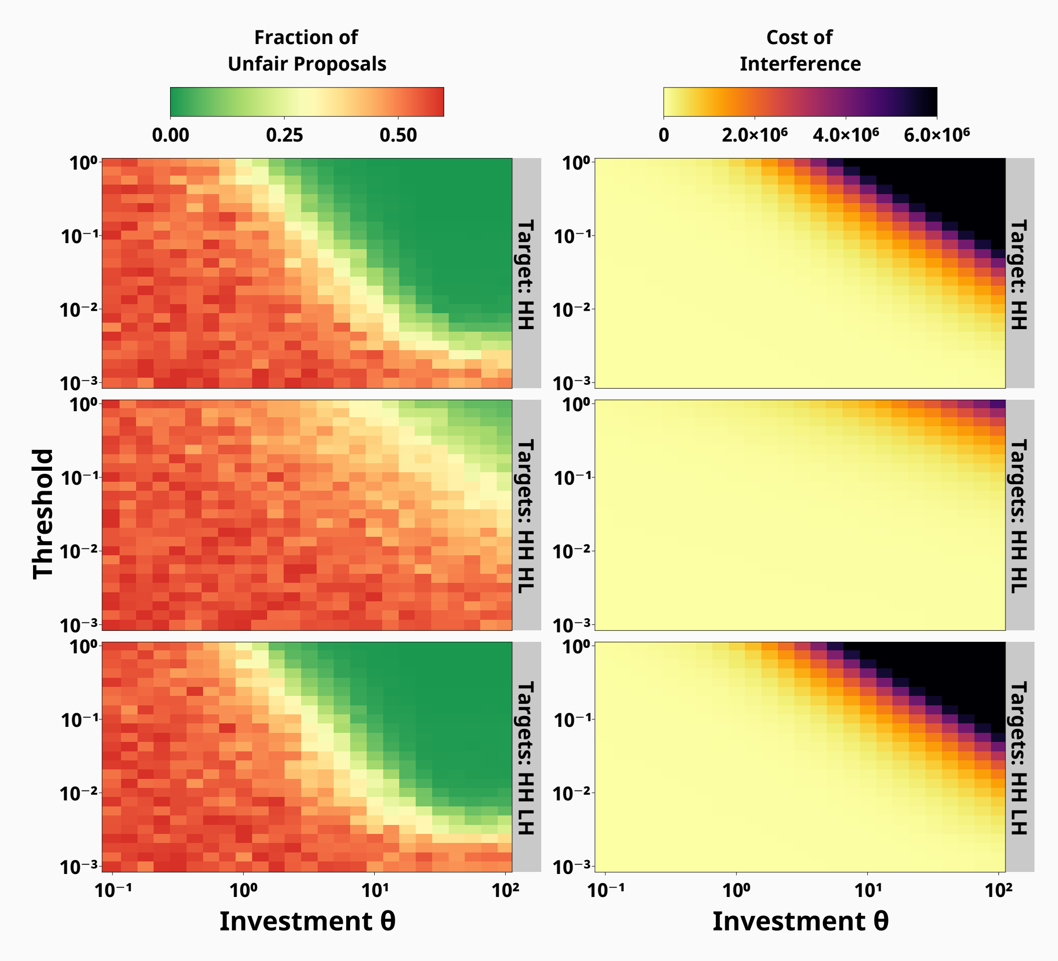

Fig. A.12 Proportion of unfair proposals and total costs of investment for NI-DEG-based interference for all targets (BA networks).

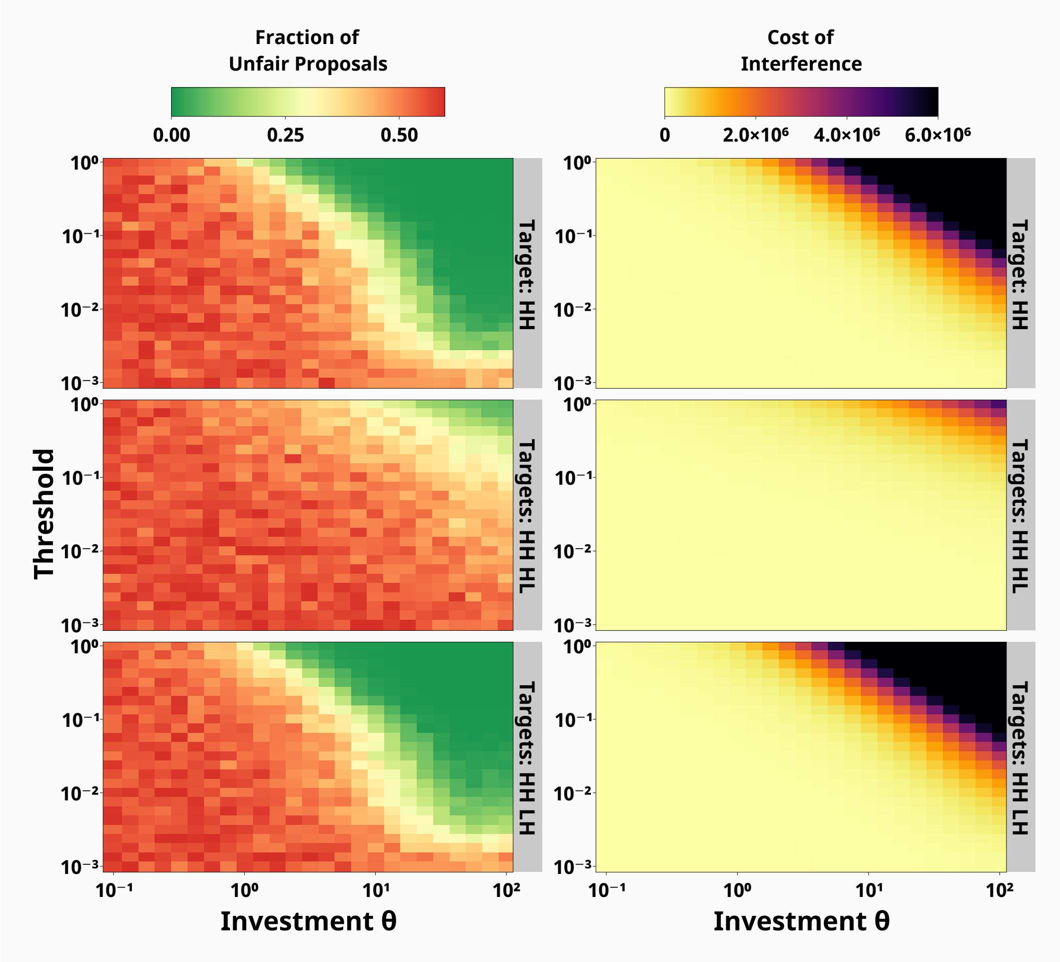

Fig. A.13 Proportion of unfair proposals and total costs of investment for NI-EIG-based interference for all targets (BA networks).

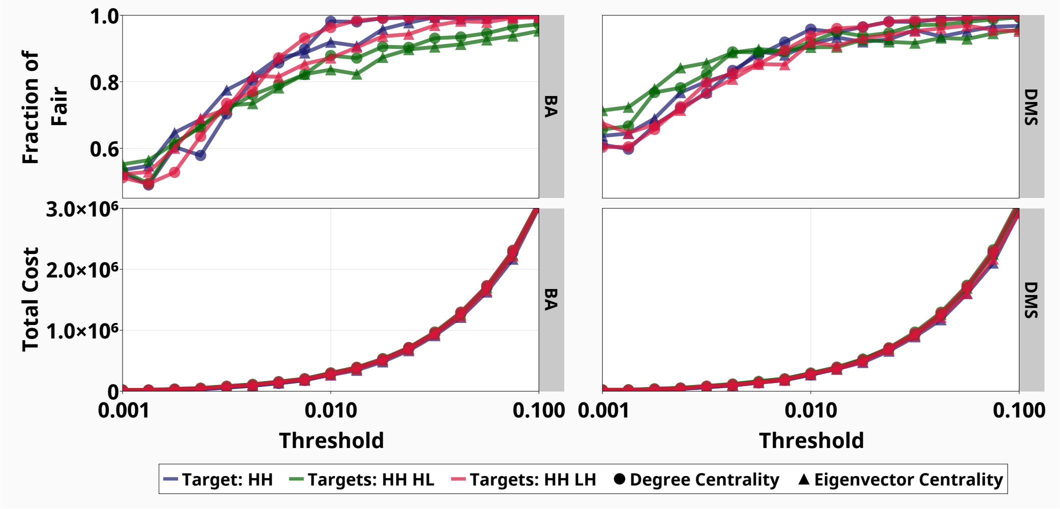

Fig. A.14 Fraction of fair proposals and cost as influential nodes are increasingly targeted in BA and DMS (highly clustered) scale-free networks. Individual investment $\theta = 31.62$ to coincide with the most-efficient value extracted from Table 1.

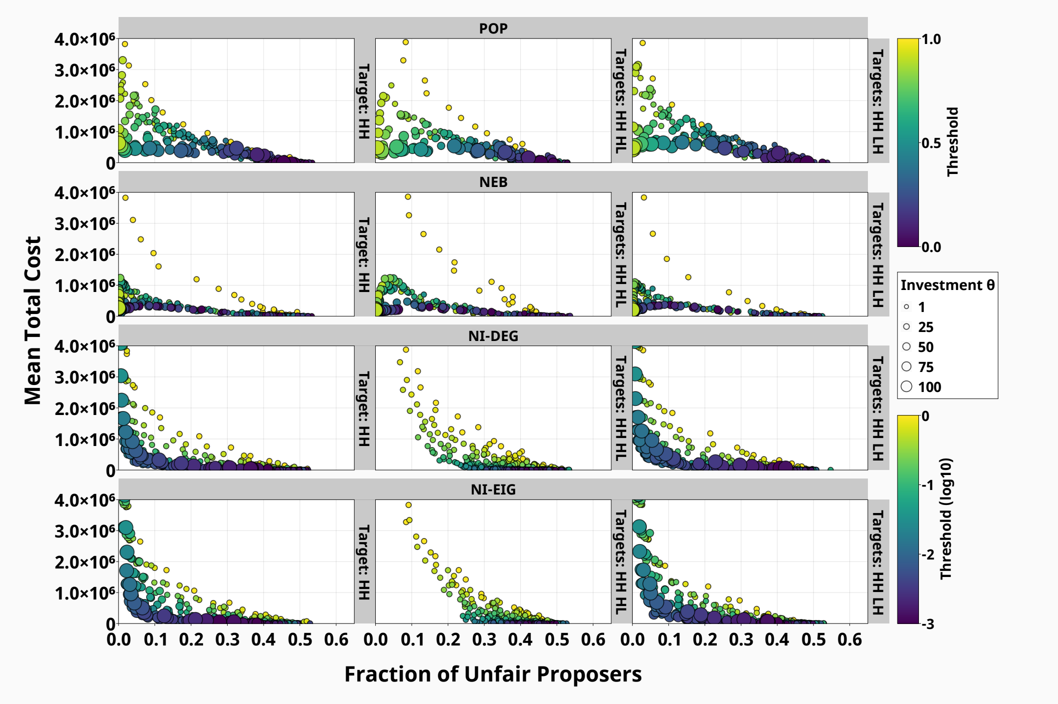

Fig. A.15 Proportion of unfair proposers as a function of the average interference cost for each scheme and target combination (DMS networks). The markers' size is determined by the individual investment θ (grouped to the nearest value), whereas the colour is determined by the threshold. Markers near the origin indicate the optimal solutions. Note that we only show the most cost-effective solutions, by limiting the maximum total cost.

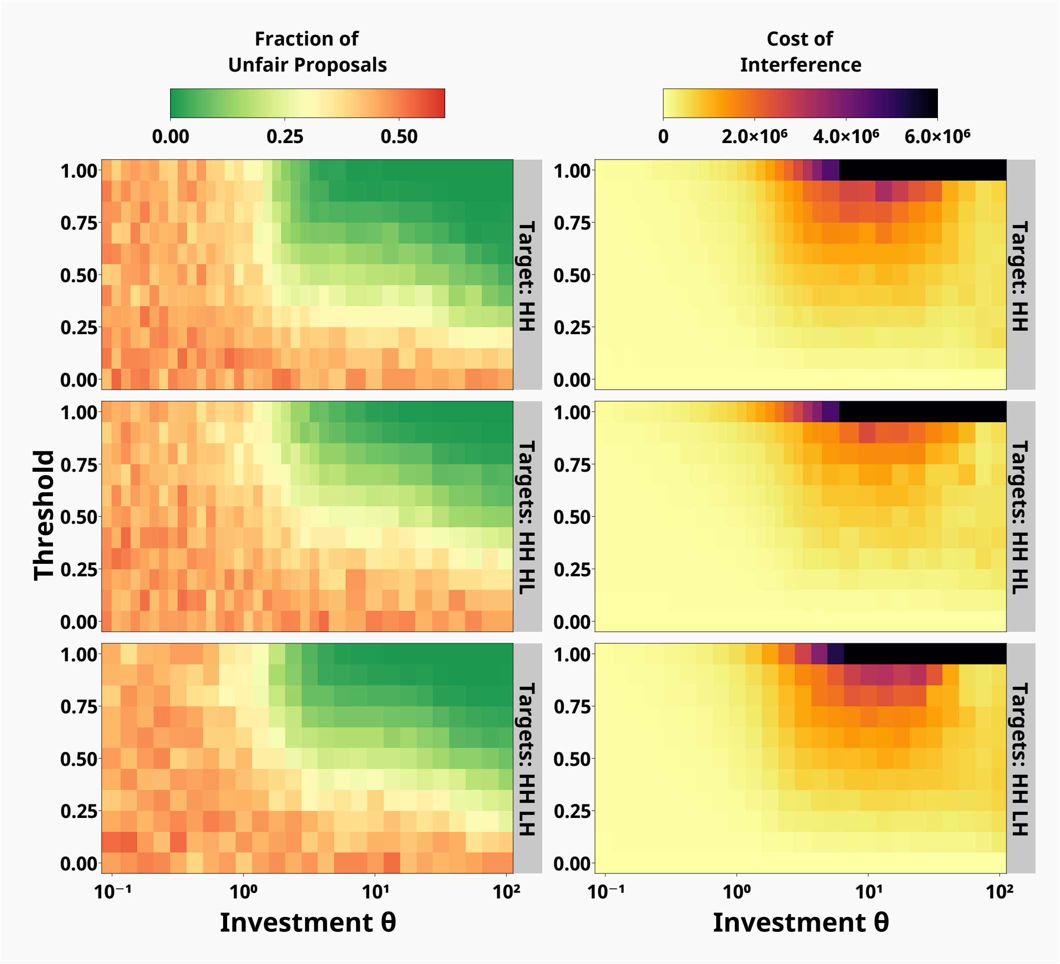

Fig. A.16 Proportion of unfair proposals and total costs of investment for POP-based interference, for all targets (DMS networks).

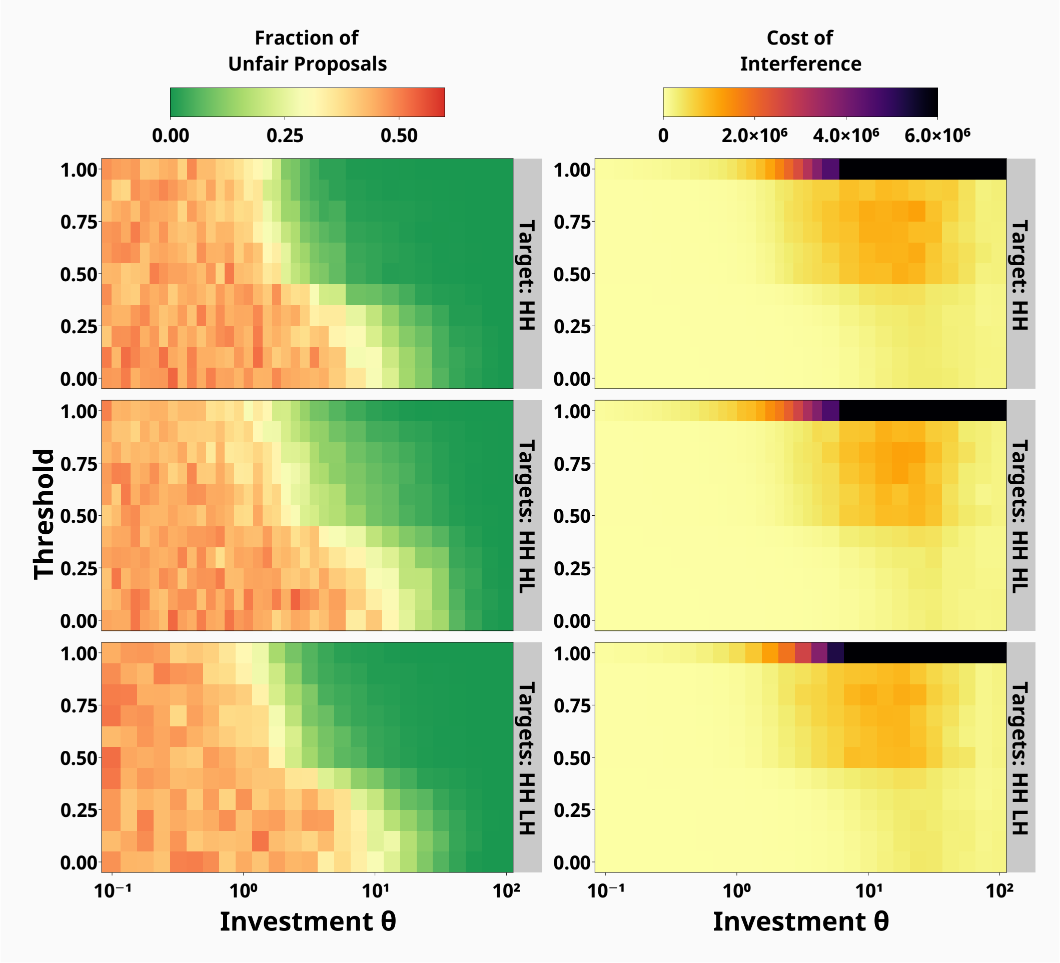

Fig. A.17 Proportion of unfair proposals and total costs of investment for NEB-based interference for all targets (DMS networks).

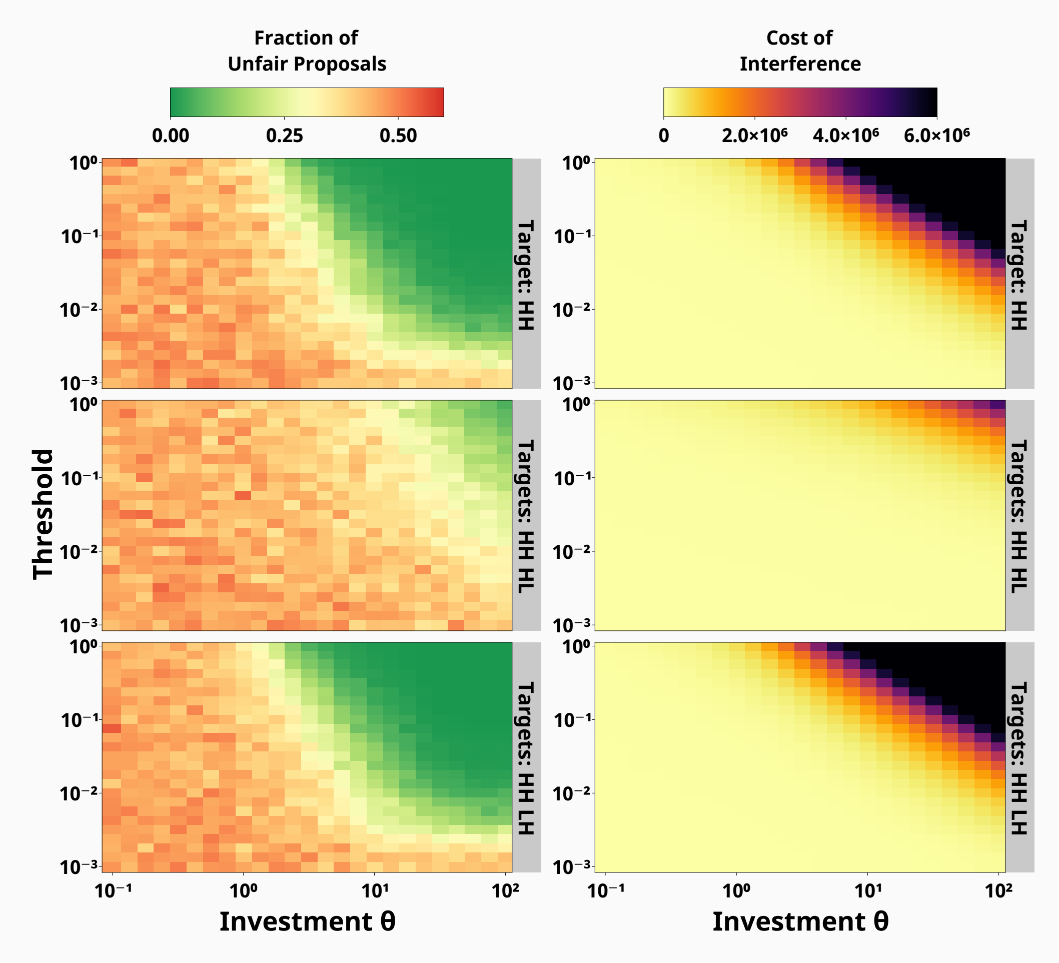

Fig. A.18 Proportion of unfair proposals and total costs of investment for NI-DEG-based interference for all targets (DMS networks).

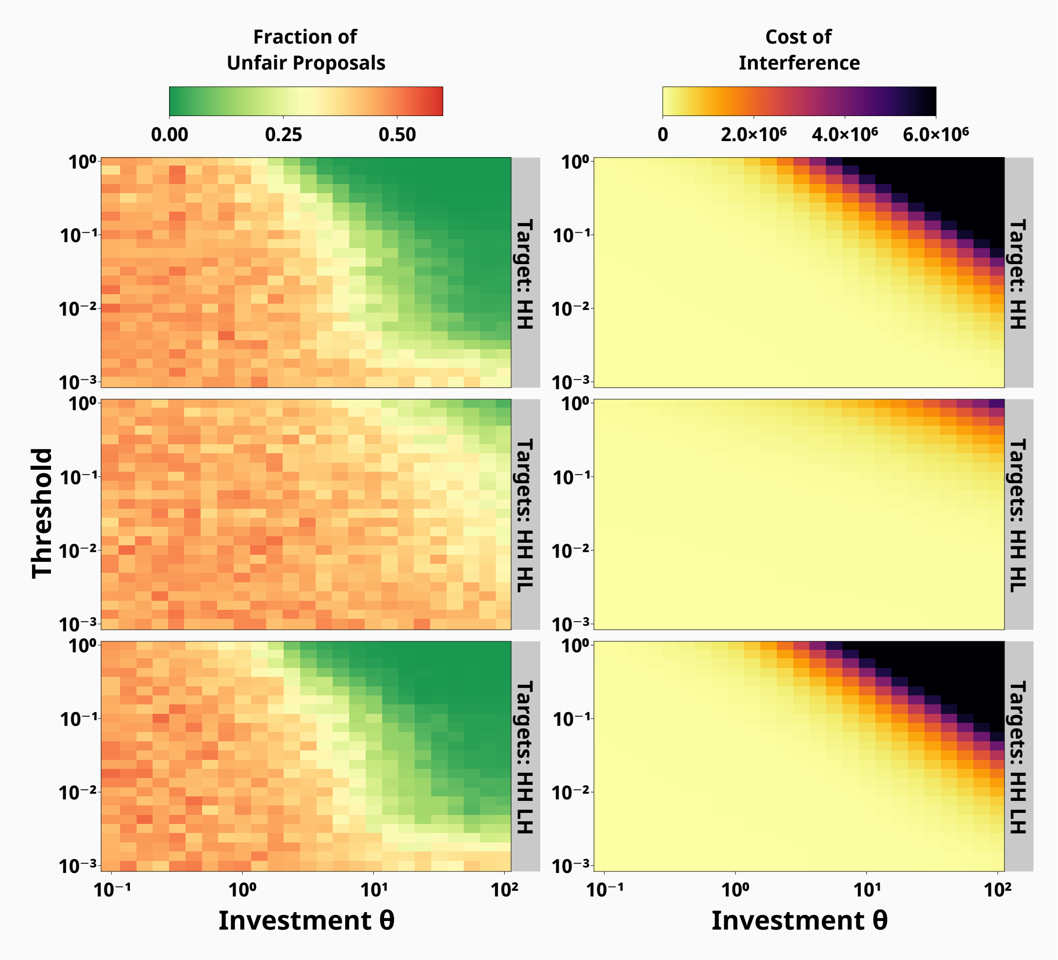

Fig. A.19 Proportion of unfair proposals and total costs of investment for NI-EIG-based interference for all targets (DMS networks).

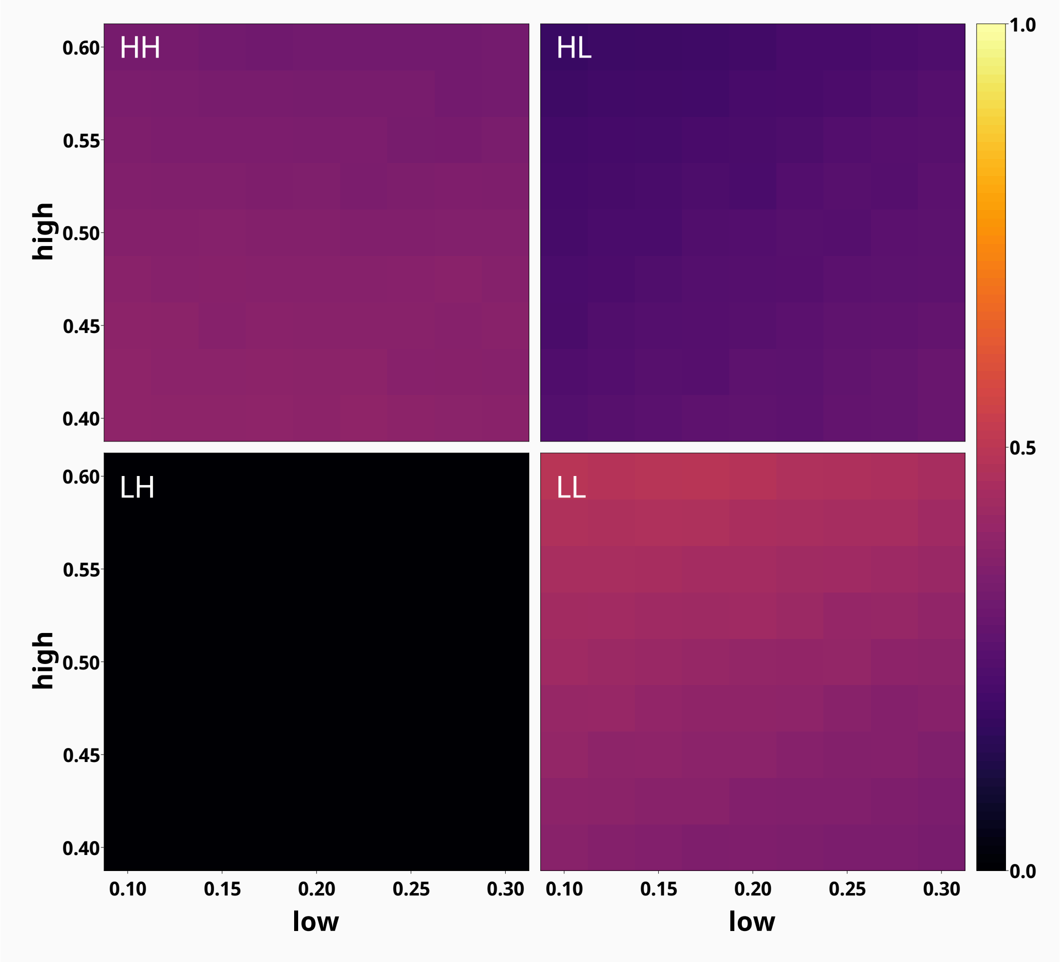

Fig. A.20 Baseline frequencies for each strategy for scale-free networks (DMS). $\mu = 0$.

B | Appendix B - Additional Results for Chapter 5

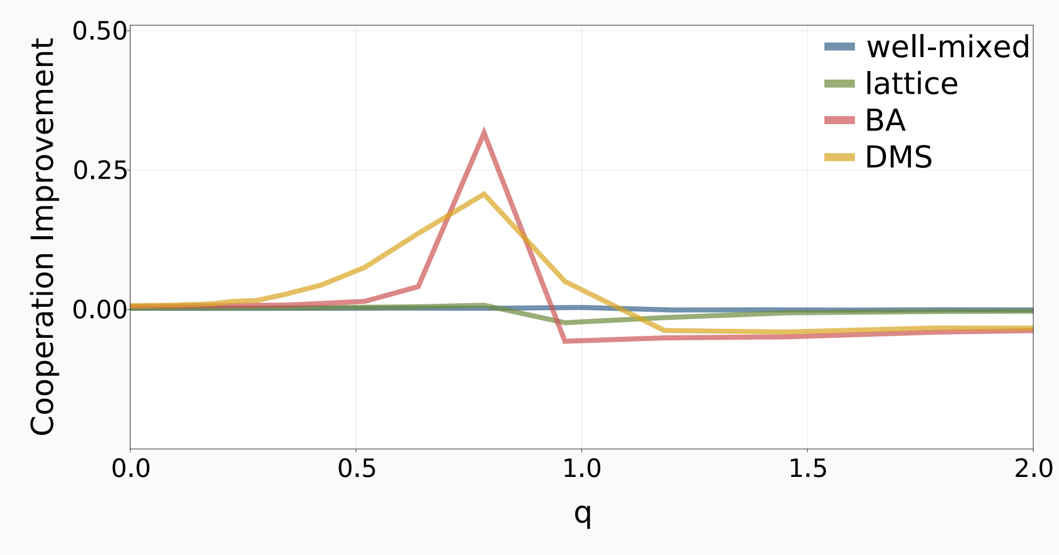

Fig. B.1 **Improvement of cooperative acts relative to baselines in the presence of institutional threat with direct observations.** Depicted are the differences between the fraction of cooperative acts in the presence of threat, and the fraction of cooperation in the baseline case, for different network structures. Parameters: $\rho = 1$.

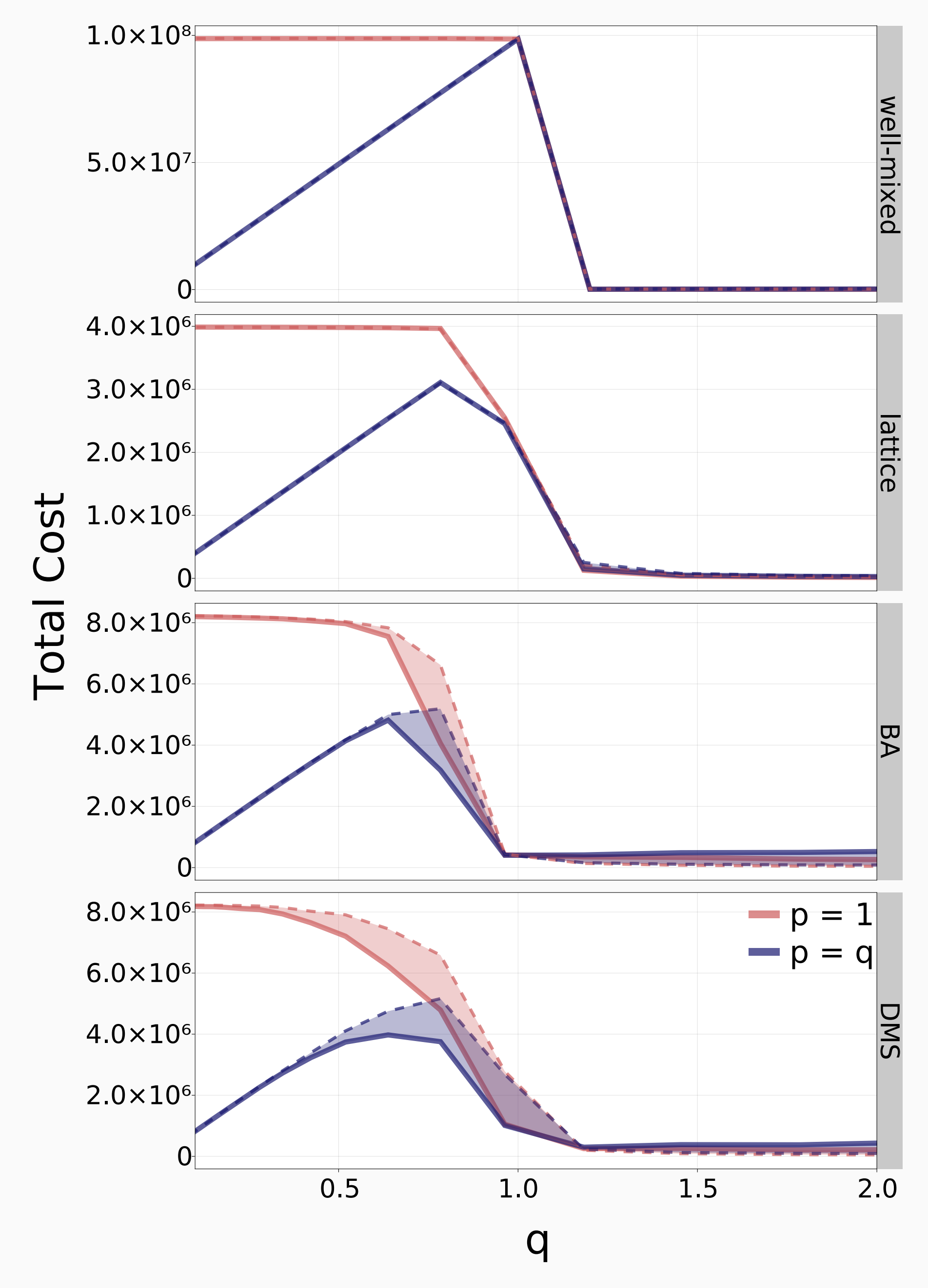

Fig. B.2 **Threat of institutional punishment reduces the total cost of punishment.** Depicted are the accumulated costs of punishment, for different network types with direct observations. The solid lines depict the accumulated costs when p is normalised to 1 and when it is equal to the sanction imposed on defection. Dashed lines depict this cost for the baseline case (in the absence of signalling). The coloured areas highlight the difference between these two lines. Parameters: $\rho = 1$.

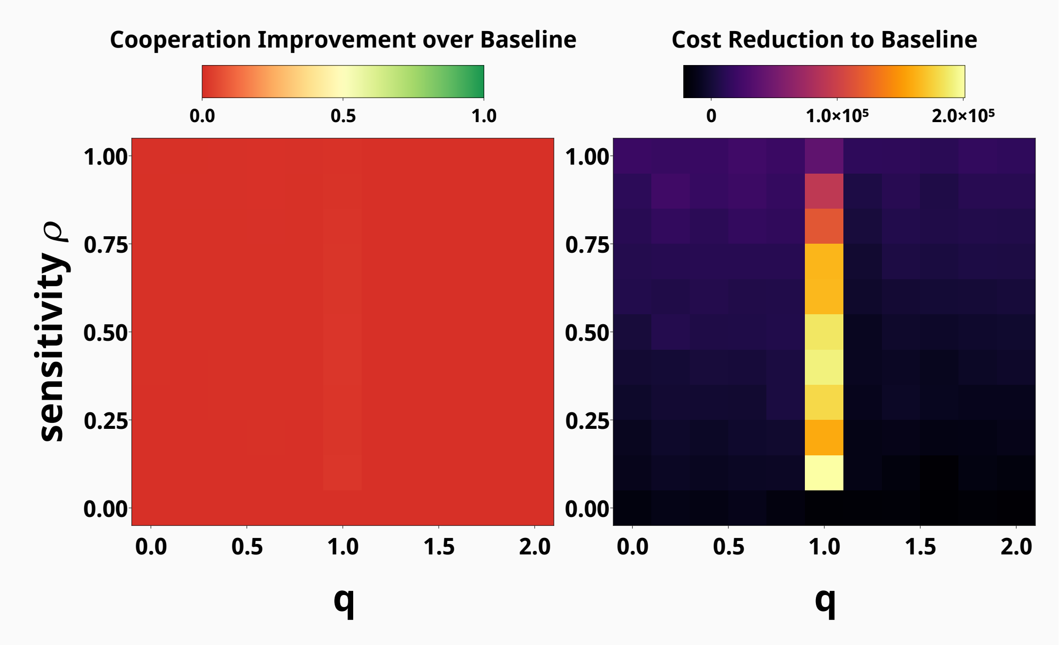

Fig. B.3 **Fear of punishment in homogeneous (WM) populations reduces the cost to institutions.** Panels show the improvement in cooperative acts and the reduction in accumulated cost, compared to the baseline, with varying levels of sensitivity to the signal (ρ), using direct observations. Parameters: $p = 1$.

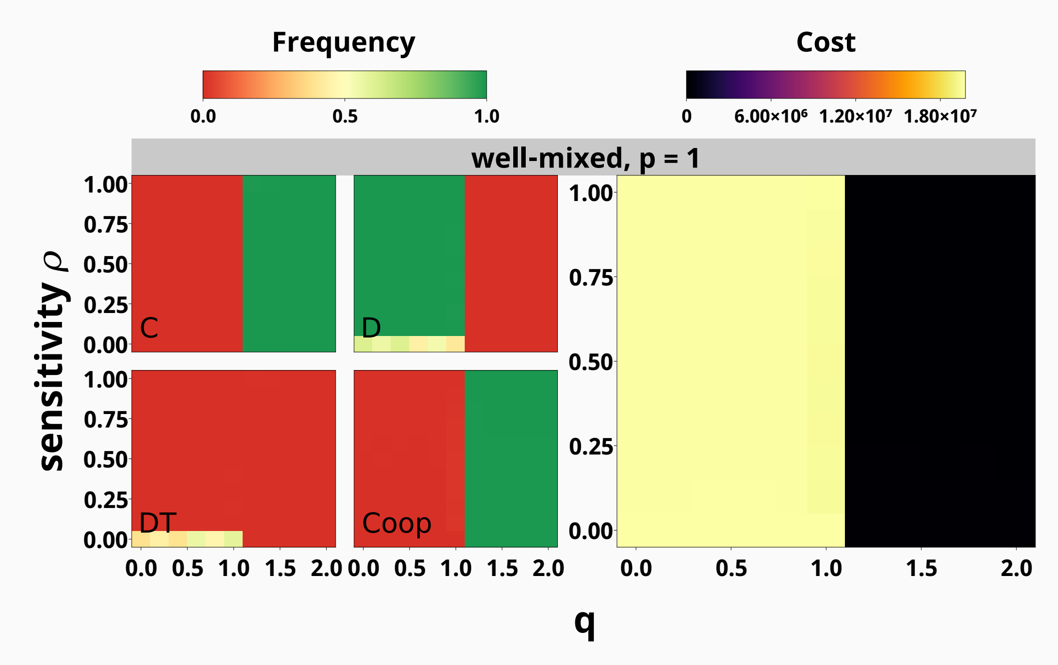

Fig. B.4 **Fear of institutional punishment changes the outcome of evolutionary dynamics in homogeneous populations.** Heatmaps show the fraction of each strategy and overall cooperative acts, as well as the total accumulated cost when signalling the institutional threat of punishment in well-mixed populations with direct observations. Parameters: $p = 1$.

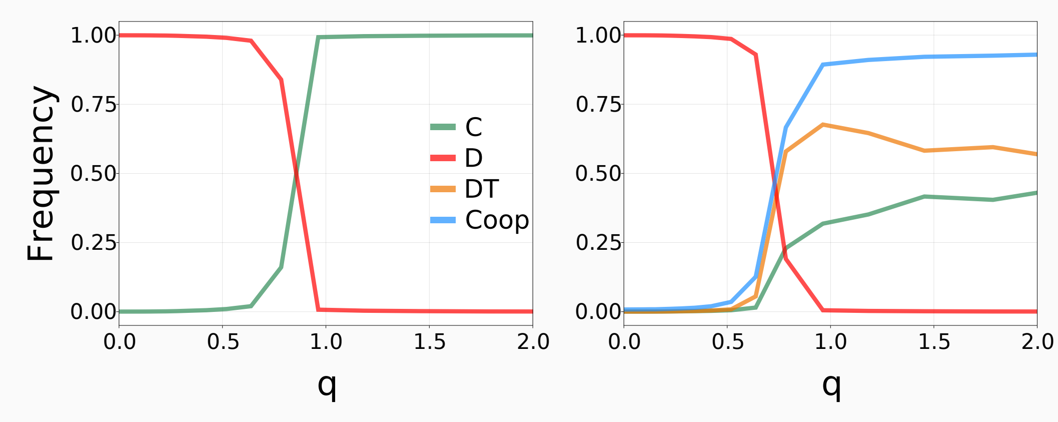

Fig. B.5 **Indirect observations of the threat of punishment allow for the coexistence of fearful defectors and cooperators.** Panels show the frequency of each strategy in BA populations, as well as the fraction of cooperative acts in the presence of threat. The left panel shows the baseline (absence of threat), while the right panel shows threat of punishment with indirect observations. Parameters: $\rho = 1$.

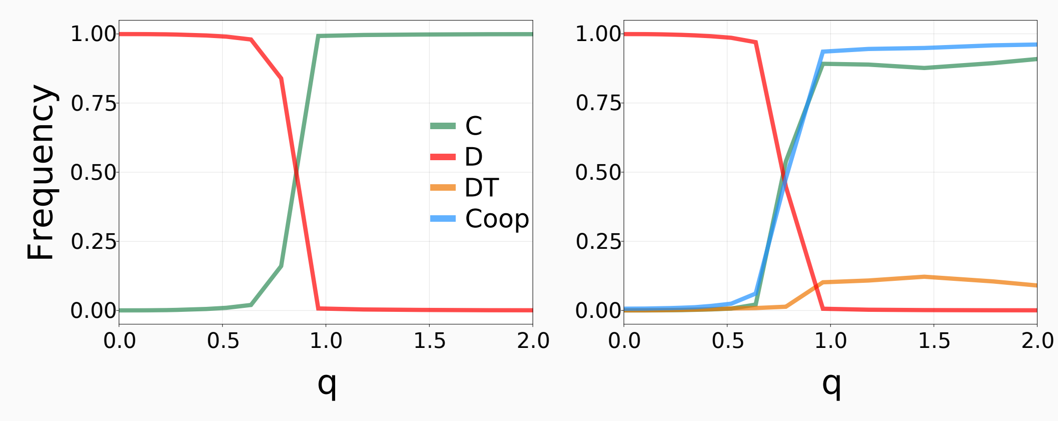

Fig. B.6 **Direct observations of the threat of punishment promote unconditional cooperators.** Panels show the frequency of each strategy in BA populations, as well as the fraction of cooperative acts in the presence of threat. The left panel shows the baseline (absence of threat), while the right panel shows threat of punishment with indirect observations. Parameters: $\rho = 1$.

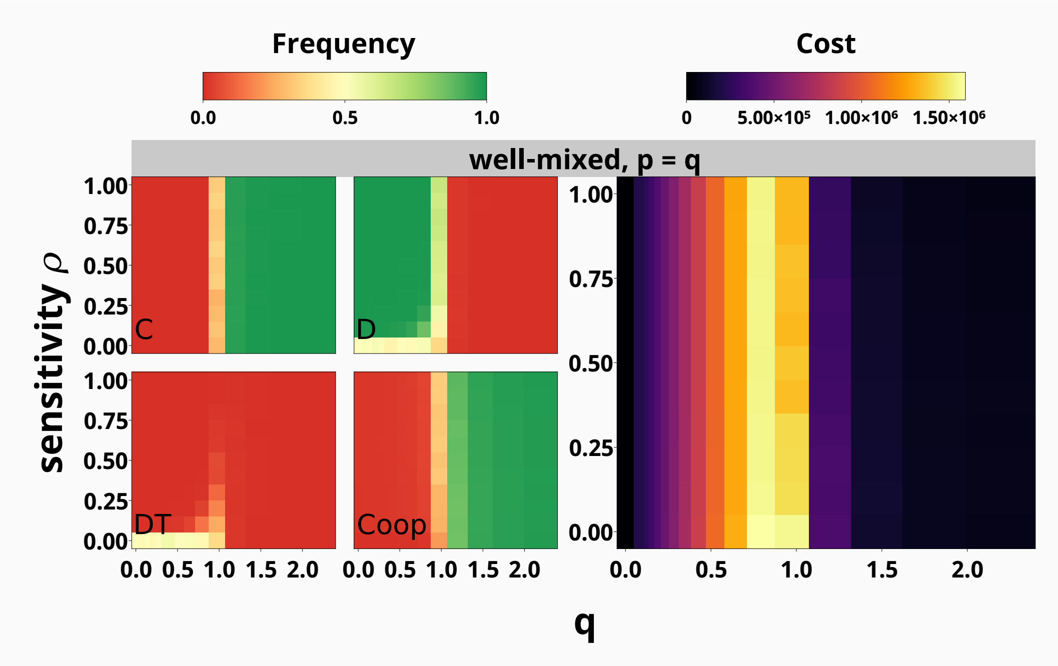

Fig. B.7 **Fear of institutional punishment changes the outcome of evolutionary dynamics in lattice populations.** Heatmaps show the fraction of each strategy and overall cooperative acts, as well as the total accumulated cost when signalling the institutional threat of punishment in lattice populations with direct observations. Parameters: $p = 1$.

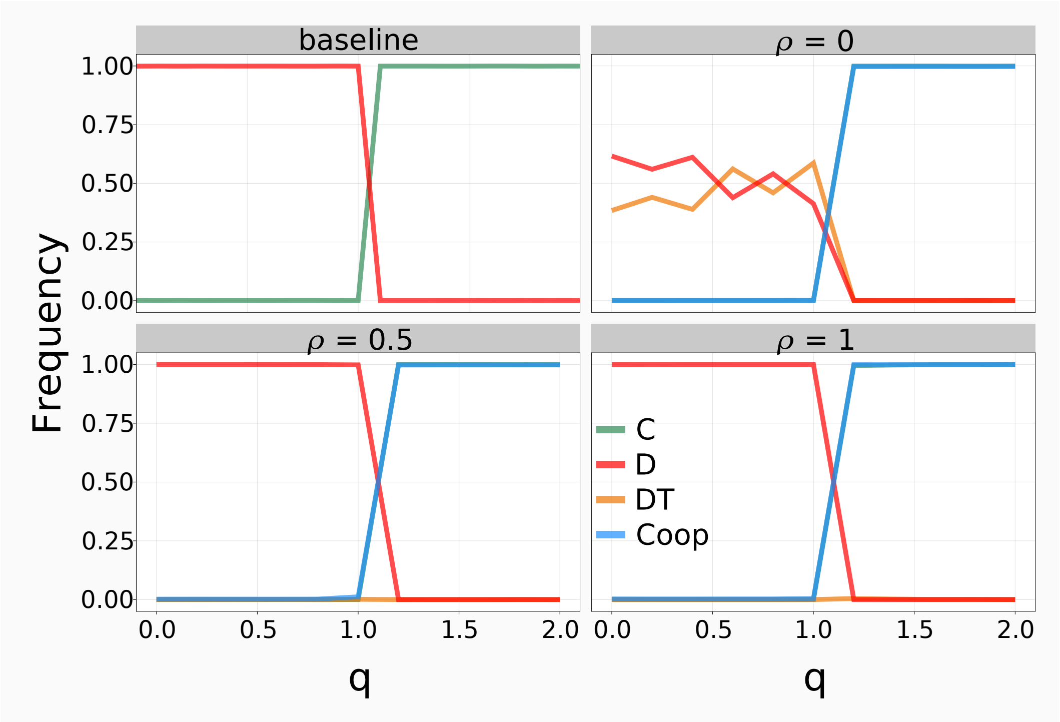

Fig. B.8 Panels show the frequency of each strategy in well-mixed populations with direct observations, as well as the fraction of cooperative acts in the presence of threat for varying levels of signal sensitivity ρ. The top left panel shows the baseline (absence of threat).

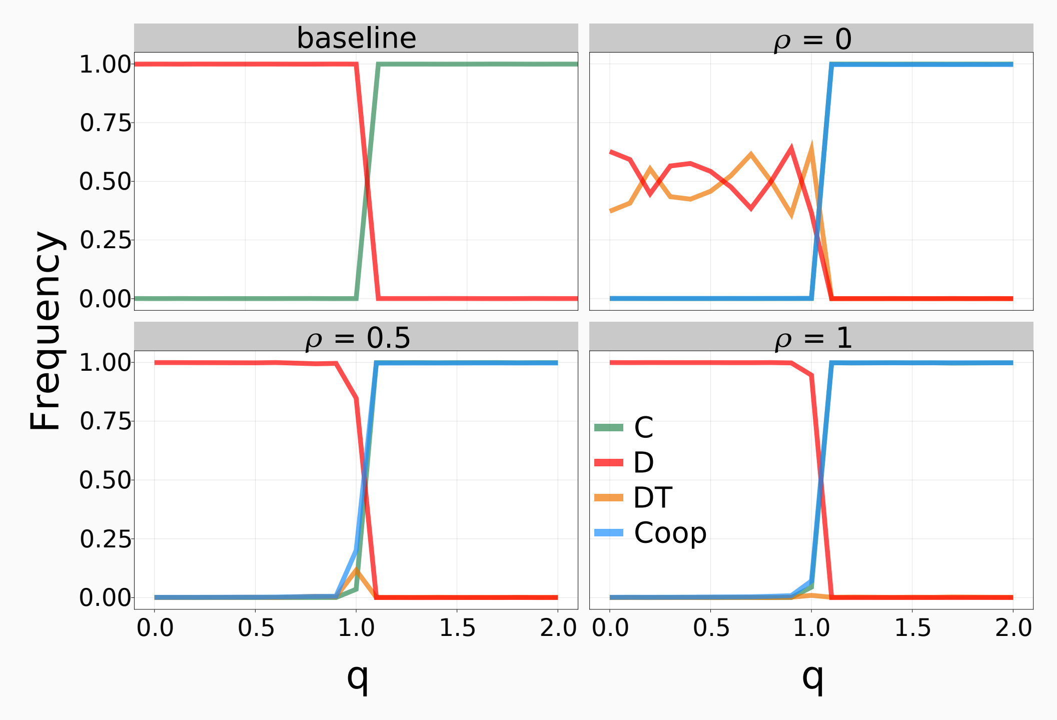

Fig. B.9 Panels show the frequency of each strategy in well-mixed populations with indirect observations, as well as the fraction of cooperative acts in the presence of threat for varying levels of signal sensitivity ρ. The top left panel shows the baseline (absence of threat).

C | Appendix C - Additional Results for Chapter 6

To further illustrate the key differences between each type of network, we plot typical simulation runs for different p_r risk probability values in the area (**II**) of the early AI race (see Figure C.1). It is immediately apparent that the two un-normalised scale-free networks provide significant improvements in safety compliance in the dilemma zone. This is further compounded by the effect of clustering on the threshold at which safe development becomes evolutionarily stable. Specifically, we note that when the risk of a disaster occurring due to inadequate safety compliance is intermediate (see, e.g. $p_r = 0.5$ and 0.65), we see a definitive improvement in highly clustered networks (i.e. DMS) as opposed to the basic BA model.

Figure C.2 confirms the similar trends encountered in the regular square lattice. There are some very minor differences, but there is very little difference between well-mixed, the normal four-neighbour lattice and the eight-neighbour lattice. We confirm the similar late convergence found previously in some cases of the regular lattice.

We see very few improvements over the previously mentioned results on homogeneous populations. Interestingly, there is an area in the late regime where this type of normalised scale-free network produces more unsafe results (undesirably so) than either the well-mixed or lattice variants. We see some slight improvements in area **(II)** of the early regime.

In order to better understand the role and influence of highly connected zealots in the population, as well as to explore any potential for a government or regulatory agency to interfere in the AI race, we artificially accelerate or fund the safety zealots that had been introduced previously. For this analysis, we choose a small number (10% of high-degree nodes) of individuals, to check whether a very small minority can be exploited by an external investor. In addition to the introduction of players following pathological safe behaviour, we either accelerate their development (similarly to how unsafe players gain increased speed, in this case we add $\frac{sB}{W}$ to the influential pathological players' payoffs, where $s = 2$), or heavily invest in these players (to the extent that other players will always imitate them, by increasing their payoffs by a very large amount 10^7). Figure C.7 displays our findings - with very slight improvement throughout. Each approach has its merits in different regions of the early regime, and we see the effectiveness of funding highly connected nodes when the risk for disaster is low. On the other hand, a high risk improves the efficacy of speeding up the development for these dedicated minorities. We note that targeting a very small minority of highly influential players is not sufficient to mitigate the race tensions entirely. Further exploration on this topic would provide more insight into how external interference can be deployed efficiently.

We study a comprehensive view of pathological players (zealots) planted in a well-mixed network (see Figure C.4), but in this case modifying 10% of the total population (not just highly connected nodes). We remove the pathological players from the frequency average to show how these affect the remainder of the population. We see very little effect of pathological players and we suggest that much lower β values would be required to see an effect. With the addition of mutation and more stochasticity, it would be possible for these pathological players to have a significant impact on the outcome.

Figure C.5 shows the evolution over time of unsafe behaviour (AU) in the dilemma zone of an early AI race for different environments (corresponding to varying probability values of a disaster caused by insufficient safety regulation, p_r). High-degree individuals appear to have a higher tendency towards safety compliance (at equilibrium) when compared to their lowly or moderately connected counterparts, except for region (**III**), where highly connected individuals are driving to innovate (optimally so). In spite of this, we see the same trends for regions (**I**) and (**III**). However, in region (**II**), highly connected individuals become important leaders in the shift from unsafe to safe behaviour in the AI race. Specifically, for large p_r values (see $p_r = 0.65$; $p_r = 0.78$), there is an evident disparity between the high degree individuals and the bulk of the population, and indeed, this is the region in which heterogeneity improves safety compliance the most. For low p_r values, heterogeneity fails to improve the outcome, but it does serve as an equaliser for intermediate risk values ($p_r = 0.5$). Regulatory actions would therefore still be required to constrain developers when heterogeneity cannot improve safety enough in region **II**, in the case of low risk of disaster to occur.

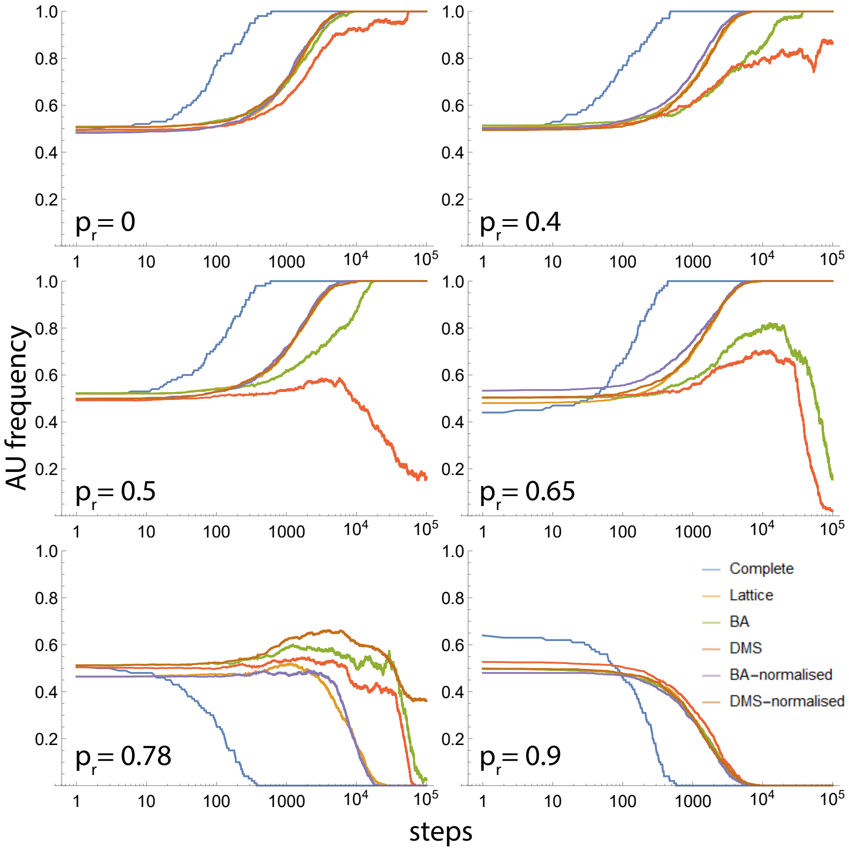

Fig. C.1 Scale-free networks (especially highly clustered networks) reduce unsafe behaviour in the dilemma regions of the early race, shown using typical runs for different risk probability values, for each type of network. Parameters: $c = 1,\ b = 4,\ B = 10^4,\ \beta = 1$.

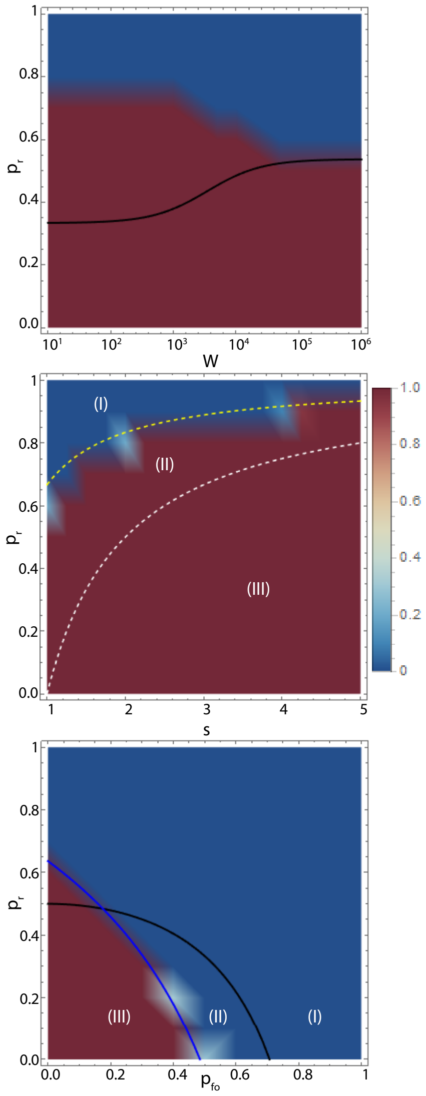

Fig. C.2 Total AU frequencies for the 8-neighbours lattice. The top row reports the spectrum between an early and late AI race (for varying W, with $p_{fo} = 0.1,\ s = 1.5$), the middle row addresses the early regime for varying s and p_r ($p_{fo} = 0.5,\ W = 100$), and the bottom row addressees the late regime for varying p_{fo} and p_r ($s = 1.5$, $W = 10^6$). Other parameters: $c = 1,\ b = 4,\ B = 10^4,\ \beta = 1$.

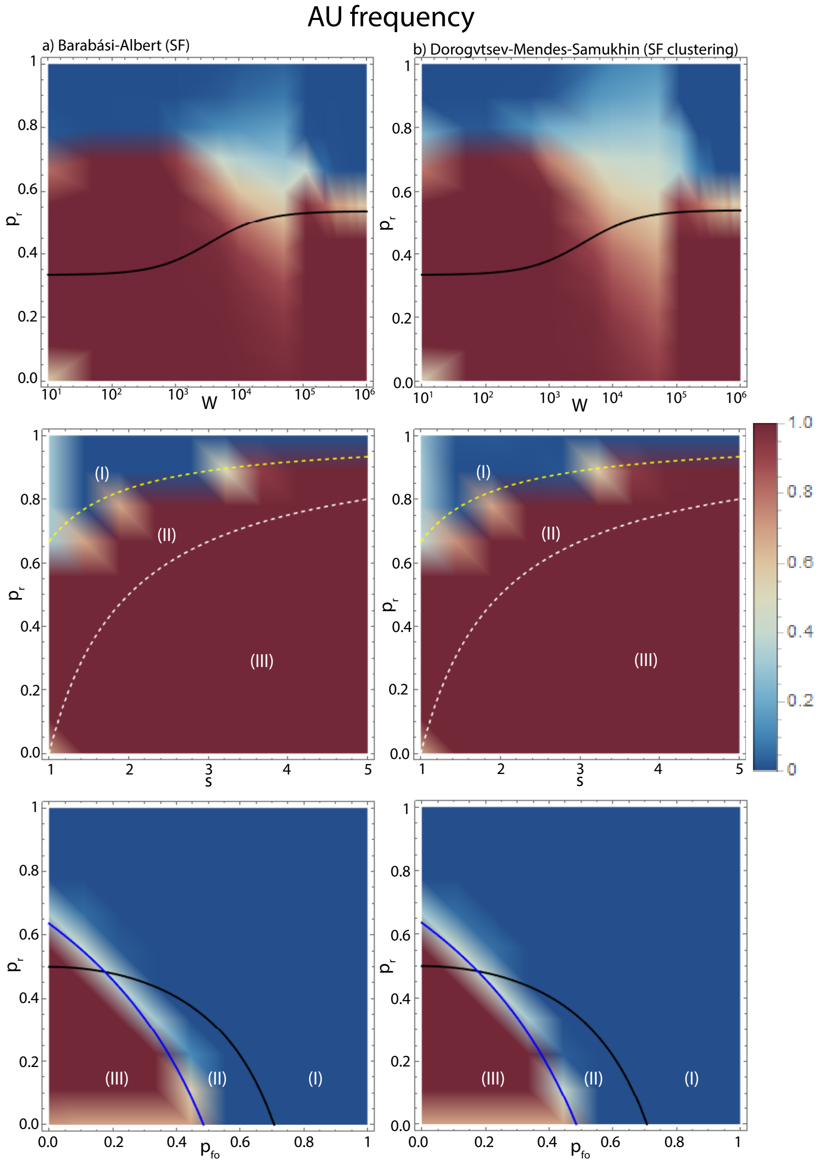

Fig. C.3 Comparison between the two different scale-free networks, BA and DMS. In this case, the payoffs have been normalised. The top row reports the spectrum between an early and a late AI race ($p_{fo} = 0.1,\ s = 1.5$), the middle row addresses the early regime in more detail ($p_{fo} = 0.5,\ W = 100$) and the bottom row considers a late AI race ($W = 10^6,\ s = 1.5$). Parameters: $c = 1,\ b = 4,\ B = 10^4,\ \beta = 1$.

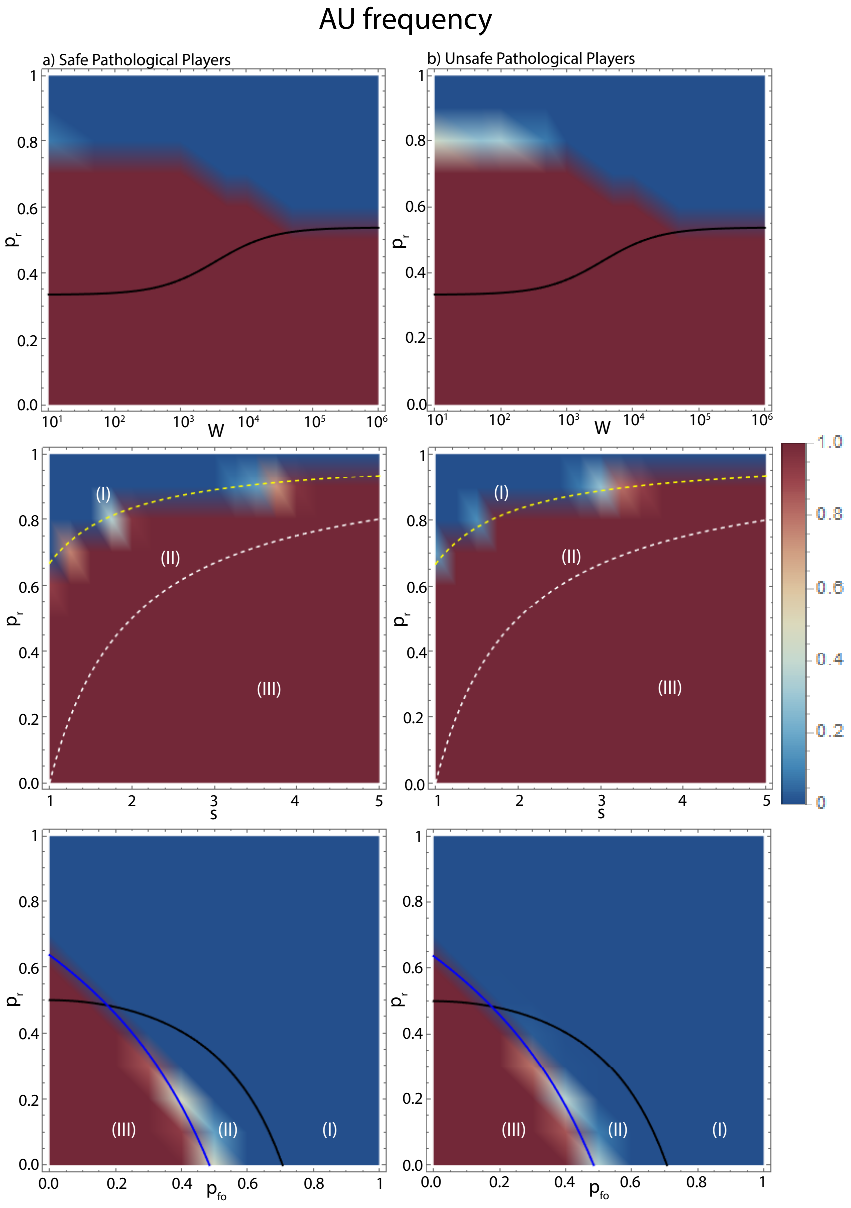

Fig. C.4 Introducing safe and unsafe zealots in the well-mixed scenario. Please note that the pathological players are excluded from these frequencies. The top row reports the spectrum between an early and a late AI race ($p_{fo} = 0.1,\ s = 1.5$), the middle row addresses the early regime in more detail ($p_{fo} = 0.5,\ W = 100$) and the bottom row considers a late AI race ($W = 10^6,\ s = 1.5$). Parameters: $c = 1,\ b = 4,\ B = 10^4,\ \beta = 1$.

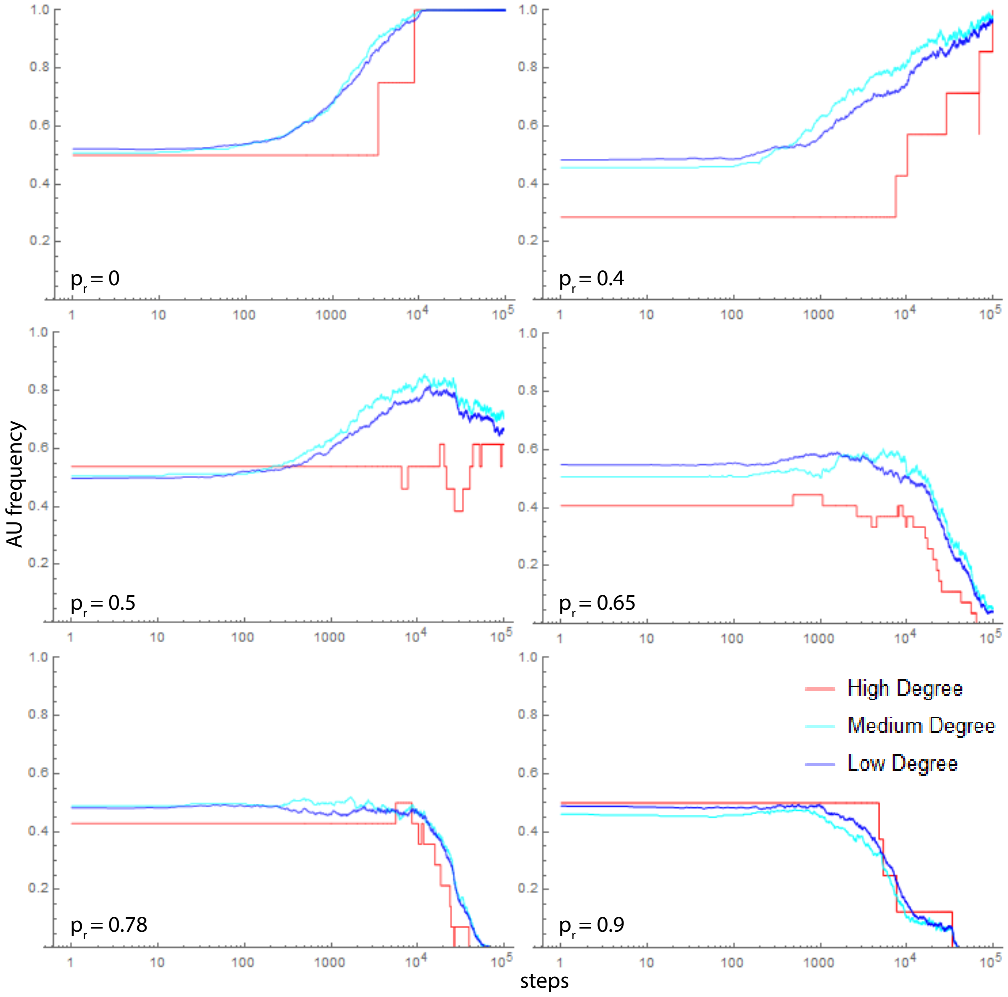

Fig. C.5 Typical runs showing the distribution of unsafe behaviour (AU) in an early AI race, grouped by degree class (connectivity) of the nodes on DMS networks, for different risk probabilities. Parameters: $c = 1,\ b = 4,\ s = 1.5,\ B = 10^4,\ W = 100,\ \beta = 1$.

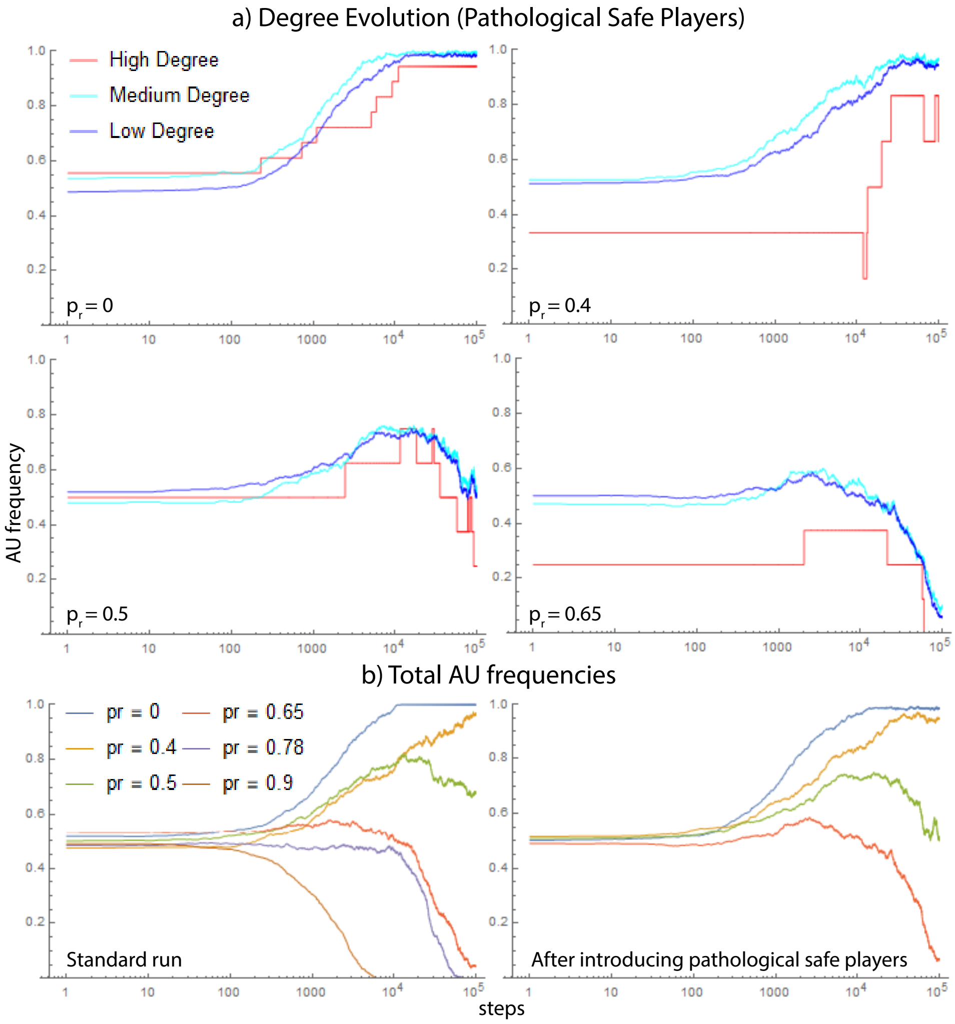

Fig. C.6 Typical runs exploring the evolutionary degree distribution of unsafe behaviour in an early AI race, following the introduction of safety zealots (pathological safe players) in the population of DMS networks. We randomly allocate 10% of high degree individuals as safety zealots. Note that we measure the frequency for the whole population, including the pathological players. Parameters: $c = 1,\ b = 4,\ s = 1.5,\ B = 10^4,\ W = 100,\ \beta = 1$.

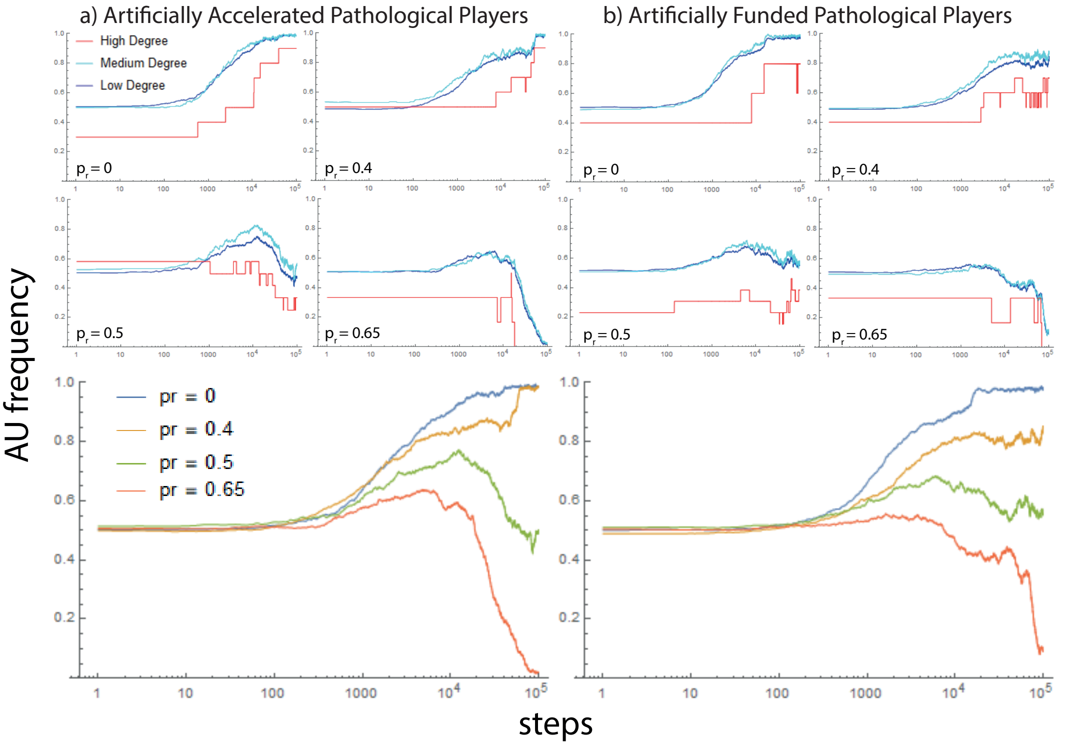

Fig. C.7 Typical runs exploring the evolutionary degree distribution of unsafe behaviour in an early AI race, following the artificial acceleration (or funding) of safety zealots (pathological safe players) in the population interacting in DMS networks. We randomly allocate 10% of high degree individuals as safety zealots. Note that we measure the frequency for the whole population, including the pathological players. Parameters: $c = 1,\ b = 4,\ s = 1.5,\ B = 10^4,\ W = 100,\ \beta = 1$.

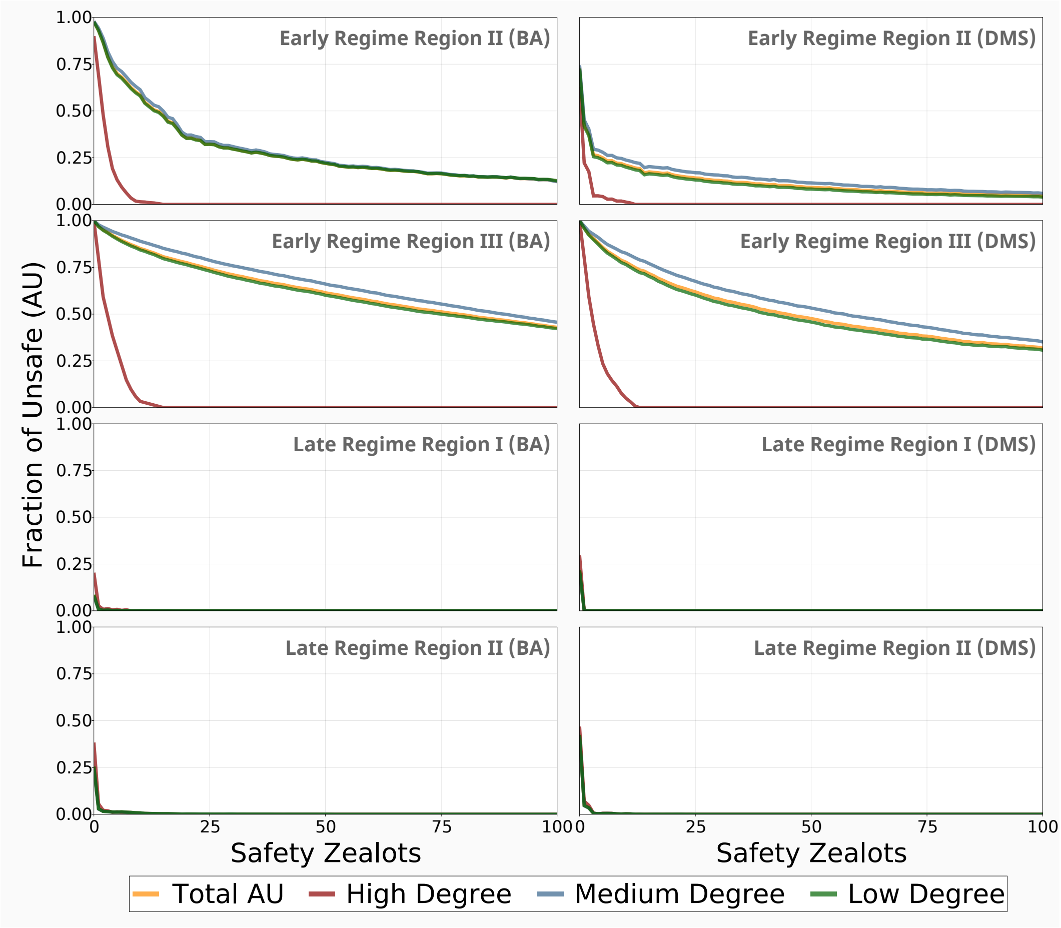

Fig. C.8 Hubs prefer slower, thus safer developments in the early race, and this can be further exploited by progressively introducing safety zealots in highly connected nodes. We show the results for both regimes, as well as the appropriate regions where safety (early region II and late region I), and conversely where innovation (early region III and late region II) are the preferred collective outcomes. The top four panels report the results for the early regime ($p_{fo} = 0.5,\ W = 100$ with $p_r = 0.5$ for region II and $p_r = 0.1$ for region III), and the bottom four do so for the late regime ($p_{fo} = 0.6,\ W = 10^6$ with $p_r = 0.3$ for region I and $p_r = 0.1$ and region II). Other parameters: $c = 1,\ b = 4,\ B = 10^4,\ s = 1.5,\ \beta = 1$.

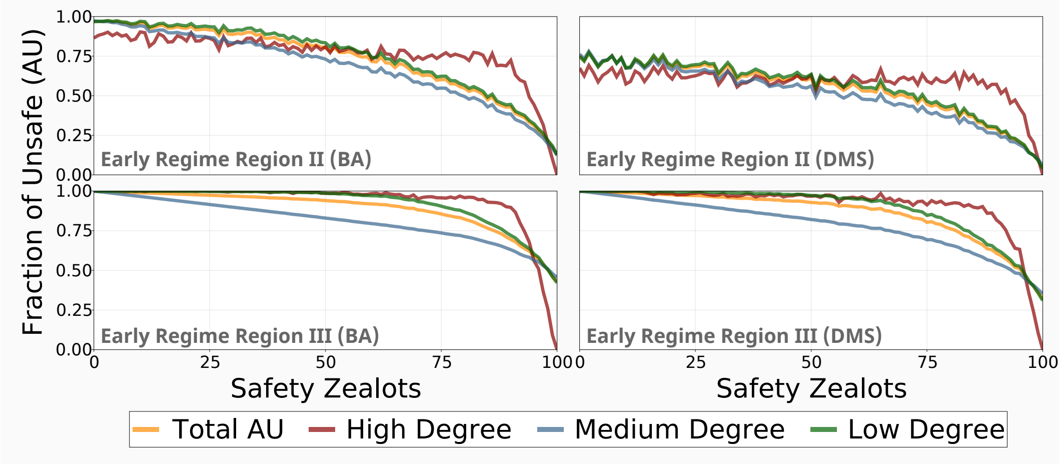

Fig. C.9 Introducing safety zealots in reverse order (still selecting the top 10% of nodes based on degree connectivity) does not produce the same exponential increase in safety that we had seen in Figure 3. We show the results for the early regime, as well as the appropriate regions where safety (region II), and conversely where innovation (region III) are the preferred collective outcomes. Parameters are $p_r = 0.5$ for region II and $p_r = 0.1$ for region III, chosen for clear presentation. Other parameters: $c = 1,\ b = 4,\ B = 10^4,\ s = 1.5,\ \beta = 1,\ p_{fo} = 0.5,\ W = 100$.