Université Paris Cité

Ecole Doctorale 474	Frontières de l'Innovation en Recherche et Éducation
Laboratoires	Institut Jean Nicod, ENS Laboratoire Traitement et Communication de l'Information, Telecom Paris

Models of reputation-based cooperation. Bridging the gap between reciprocity and signaling

Par Julien LIE-PANIS

Thèse de doctorat de Théorie des Jeux evolutionnaire

Spécialité en Sciences cognitives

Dirigée par

Jean-Baptiste ANDRÉ et Jean-Louis DESSALLES

Présentée et soutenue publiquement le 2 Octobre 2023

Devant un jury composé de :

Jean-Pierre NADAL	Université Paris Cité	Président
Pat BARCLAY	University of Guelph	Rapporteur
Daniel NETTLE	Université Paris Sciences & Lettres	Rapporteur
Francesca GIARDINI	University of Groningen	Examinatrice
Eleanor A. POWER	London School of Economics & Political Science	Examinatrice
Jean-Baptiste ANDRÉ	Université Paris Sciences & Lettres	Directeur
Jean-Louis DESSALLES	Telecom Paris	Directeur

Acknowledgements

I am deeply indebted to my advisor, Jean-Baptiste. Jean-Baptiste has taught me everything there is to know about the craft. I am particularly admirative of his broad, all-encompassing vision, and particularly thankful for the unlimited freedom he has left me during my PhD. Jean-Baptiste trusted me enough to let me grow and even seek my own outside collaborations, and supported me enough (i.e. a lot) to make both a possibility. I couldn't have asked for a better training, or for a better mentor.

I would also like to express my sincerest gratitude to my other advisor, Jean-Louis. It's simple: I would not be here were it not for Jean-Louis. Thank you for giving me the confidence to apply my math chops to study social behavior, and for fighting so that I could have the opportunity to learn to do so. And for so many thought-provoking ideas.

I would be remiss if I didn't thank Moshe, who many times felt like an unofficial third advisor. It's still surprising to me how we suddenly came to work together. Thank you for showing me exciting new ways to look at things, and keeping me on my toes. And for the kind mentorship.

I would also like to extend my thanks to Bethany and Nicolas. I feel insanely lucky to have been trained by all these smart people. Coming from very different fields than my own (but quite used to such collaborations), they have kindled my curiosity for other fields, taught me new methods, and ways of integrating these methods with my own.

I am running out of ways to say thank you. Nevertheless, deepest thanks to Coralie and Hugo. Coralie and Hugo are the beating heart of our team. They have instilled a wonderful work atmosphere (and made sure we had desks to anchor us during our ethereal discussions). I am particularly indebted to Coralie for the opportunity to teach at a truly wonderful program, and the opportunity for one additional year of training.

On that note, huge thanks to my comrade in arms, Valentin. Valentin has made our many classes together (6!) effortless, and immensely pleasurable. I have learned from the most passionate and unconventional teacher I could have. Thank you for the good advice: it's best not to leave the acknowledgements for the last minute (I didn't listen). On *that* note, huge thanks to another more recent comrade in arms, Léo. I was tempted to put Léo above, along with my advisors—but he doesn't have his diploma yet. Our collaboration was enlightening and truly thrilling (this might not be the last you hear of me).

There are many others I am indebted to, including some I will forget (I blame the planning fallacy). I would like to thank Olivier, for his thoughtful and deep feedback on my work, and for just the right amount of Bernard Tapie jokes. As well as Christophe, Silvia and Valeria for their equally thoughtful feedback and strong support: thank you for helping me grow as a researcher.

I would like to extend my deepest thanks to Christian, as well as Charlotte, Marta, Nikoleta and Saptarshi for welcoming me, and giving even more reasons to be excited about my line of work. If I'm so eager to return to Plön, it's not for the weather.

I would like to thank Mélusine and Edgar, with whom I have shared an office and a journey. One of the most crazy and intimidating things about this whole experience is how banally exceptional my fellow students were. Thank you both for your unique blend(s) of communicative energy and calm, soothing presence. I would also like to thank other exceptionally admirable, smart, nice and fun people who accompanied me on long stretches of this journey, including Amine, Aurore, Benoît, Charles, Étienne, Jan, Laudine, Maroua, Marc and Mia. And many other equally high quality partners (wink, wink, honest signaling)

such as Adrien, Alexey, Ali, Anne-Sophie, Antoine, Caroline, Léonard, Loïa, Hritika, Isacco, Mathilde, Marius, Mauricio, Mona, Noémon, Olha, Paul, Rita, Sasha and Zoé.

I would like to thank the many who have helped me by giving their time, a kind comment, or simply by doing their job with enthusiasm. I am notably thinking of Daniel and Pat, two particularly accessible and thoughtful scientists, who have generously agreed to review this dissertation. And to the many others who make this job so pleasurable including Alejandro, Amanda, Andrea, Eleanor, Francesca, Helena, Hirotaka, Luuk, Maija, Olympia, Jean-Pierre, Jonathan, Jorge, Peter, Péter, Valentin, and Zach.

Extending the focus, I would like to thank all my friends. I really have chosen the best cooperation partners (or rather, for some reason, they have chosen me). I am indebted to Ulysse, without whom I could not be here, and to Julien, without whom my journey would be much less funny. I would like to thank Aude for our fitness interdependence, and helping me understand the research world and find my place. Equally useful in this regard, Thomas and Marine, as well as Alexandre and Michel for their unwavering support. I am also greatly indebted to Adrien, Agathe, Alexandre, Cécile, Gautier, Mathilde, and so many others.

I would like to thank my parents for their unflinching help and understanding (even when not always understanding my motivations, making their help all the more admirable). And my sisters. And my family in Bretagne.

I would like to finish by thanking Ségo, my longest cooperation partner, unlimited source of motivation, and reviewer I most want to impress. Thank you for pushing me to do this, and thank you making *me* see the world in a different light (I might be stealing this, but it's true!). And thank you Beth, for teaching me to be both more productive and passionate at the same time—I can't wait to teach you game theory in return.[1]

[1]This is, of course, intended to be a reward, and not a punishment.

Contents

Abstract **v**

Résumé **vii**

Résumé détaillé **ix**

1 Introduction **1**
1.1 Economic game theory 2
1.2 Evolutionary game theory 8
1.3 Two views of reputation 14
1.4 Bridging the gap between reciprocity and signaling 16

2 Cooperation as a signal of time preferences **21**

3 The repeated punishment game explains why, and when, we seek revenge **39**

4 Runaway signals: Exaggerated displays of commitment may result from second-order signaling **89**

5 A model of endogeneous institution formation through limited reputational incentives **117**

6 Discussion **169**
6.1 Build on previous work 171
6.2 Use analogies 172
6.3 Embrace your passion for stupid models 173
6.4 Seek smart friends (from other disciplines) 175

References **177**

Liste des éléments retirés **188**

Abstract

Human cooperation is often understood through the lens of reciprocity. In classic models, cooperation is sustained because it is reciprocal: individuals who bear costs to help others can then expect to be helped in return. Another framework is honest signaling. According to this approach, cooperation can be sustained when helpers reveal information about themselves, which in turn affects receivers' behavior. Here, we aim to bridge the gap between these two approaches, in order to better characterize human cooperation. We show how integrating both approaches can help explain the variability of human cooperation, its extent, and its limits. In chapter 1, we introduce the main method used during this thesis: evolutionary game theory. In chapter 2, we show that cooperation with strangers can be understood as a signal of time preferences. In equilibrium, patient individuals cooperate more often, and individuals who reveal higher preference for the future inspire more trust. We show how our model can help explain the variability of cooperation and trust. In chapter 3, we turn to the psychology of revenge. Revenge is often understood in terms of enforcing cooperation, or equivalently, deterring transgressions: vengeful individuals pay costs, which may be offset by the benefit of a vengeful reputation. Yet, revenge does not always seem designed for optimal deterrence. Our model reconciles the deterrent function of revenge with its apparent quirks, such as our propensity to overreact to minuscule transgressions, and to forgive dangerous behavior based on a lucky positive outcome. In chapter 4, we study dysfunctional forms of cooperation and signaling. We posit that outrage can sometimes act as a second-order signal, demonstrating investment in another, first-order signal. We then show how outrage can lead to dishonest displays of commitment, and escalating costs. In chapter 5, we extend the model in chapter 2 to include institutions. Institutions are often invoked as solutions to hard cooperation problems: they stabilize cooperation in contexts where reputation is insufficient. Yet, institutions are at the mercy of the very problem they are designed to solve. People must devote time and resources to create new rules and compensate institutional operatives. We show that institutions for hard cooperation problems can emerge nonetheless, as long as they rest on an easy cooperation problem. Our model shows how designing efficient institutions can allow humans to extend the scale of cooperation. Finally, in chapter 6, we discuss the merits of mathematical modeling in the social sciences.

Key words: game theory, evolution, reciprocity, honest signaling, cooperation, revenge, outrage, institutions

Résumé

La coopération humaine est souvent appréhendée sous l'angle de la réciprocité. Dans les modèles classiques, la coopération est maintenue parce que réciproque : les individus qui assument un coût pour aider les autres peuvent s'attendre à être aidés en retour. Un autre angle est offert par la théorie du signal honnête. Selon cette approche, la coopération peut être maintenue lorsque le fait d'aider informe sur des qualités sous-jacentes, ce qui affecte le comportement des destinataires du signal. Nous visons ici à combler le fossé entre ces deux approches, afin de mieux caractériser la coopération humaine. Nous montrons comment l'intégration des deux approches peut aider à expliquer la variabilité de la coopération humaine, son étendue et ses limites. Dans le chapitre 1, nous introduisons la méthode principale sur laquelle cette thèse est basée : la théorie des jeux évolutionnaire. Dans le chapitre 2, nous montrons que la coopération avec des inconnus peut être comprise comme un signal de préférences temporelles. À l'équilibre, les individus patients coopèrent plus souvent, et les individus qui révèlent une plus grande préférence pour l'avenir inspirent davantage confiance. Notre modèle peut expliquer la variabilité de la coopération et de la confiance. Le chapitre 3 est consacré à la psychologie de la vengeance. La vengeance est souvent comprise comme un moyen d'imposer la coopération ou, de manière équivalente, de dissuader les transgressions : les individus vengeurs assument des coûts, qui peuvent être compensés par l'avantage d'une réputation vengeresse. Pourtant, la vengeance ne semble pas toujours conçue pour une dissuasion optimale. Notre modèle réconcilie la fonction dissuasive de la vengeance avec ses bizarreries apparentes, comme notre propension à réagir de manière excessive à des transgressions minuscules, ainsi que la tendance à pardonner un comportement dangereux lorsqu'il aboutit de manière fortuite à un résultat positif. Dans le chapitre 4, nous nous penchons sur les formes dysfonctionnelles de coopération et de signal. Nous postulons que l'indignation peut parfois servir de signal de second ordre, en démontrant l'investissement de l'individu dans un autre signal du premier ordre. Nous montrons ensuite comment l'indignation peut conduire à des signaux malhonnêtes et à une escalade des coûts. Dans le chapitre 5, nous étendons le modèle du chapitre 1 aux institutions. Les institutions sont souvent invoquées comme des solutions à des problèmes de coopération difficiles : elles stabilisent la coopération dans des contextes où la réputation est insuffisante. Cependant, les institutions sont à la merci du problème même qu'elles sont censées résoudre. Les individus doivent consacrer du temps et des ressources à l'élaboration de nouvelles règles et à la rémunération des acteurs institutionnels. Nous montrons que des institutions pour des problèmes de coopération difficiles peuvent néanmoins émerger, à condition qu'elles reposent sur un problème de coopération facile. Notre modèle montre comment la conception d'institutions efficaces permet aux humains d'étendre l'échelle de la coopération. Enfin, dans le chapitre 6, nous discutons des mérites de la modélisation mathématique en sciences sociales.

Mots clefs : théorie des jeux, évolution, réciprocité, signal honnête, coopération, vengeance, indignation, institutions

Résumé détaillé

Ce document est structuré comme suit.

Chapitre 1 : Introduction

Dans ce chapitre, j'introduis la théorie des jeux, un ensemble d'outils mathématiques utilisés pour étudier des situations d'intéraction—des situations où les résultats de chacun dépendent des choix des autres, et non pas seulement de ses propres choix. J'introduis tout d'abord la théorie des jeux dite économique, où les individus sont supposés rationnels et la rationalité de chacun est connaissance commune. Dans ce cadre, on peut s'attendre à atteindre un équilibre de Nash : une situation de laquelle aucun individu n'a d'intérêt à dévier unilatéralement. J'illustre ceci avec le dilemme du prisonnier : lorsque la coopération est individuellement coûteuse, on peut s'attendre à ce que les individus ne coopèrent pas à l'équilibre. Je conclus cette première sous-section en révélant les apparentes limites de cette approche. Bien souvent, les individus ont des comportements qui peuvent paraître irrationnels—si on se réfère à la définition restreinte de la rationalité de la théorie des jeux économique. En outre, ils coopérent souvent dans des situations ressemblant au dilemme du prisonnier.

J'introduis ensuite un nouveau cadre : la théorie des jeux évolutionnaire. Dans celui-ci, on ne considère plus des individus rationnels, qui raisonnent de manière à adopter des comportements optimaux. Au lieu de cela, on suppose que le jeu représente une intéraction importante pour un organisme, et que la propension à prendre une décision plutôt qu'une autre est héritable. Ceci amène à une toute autre interprétation de l'équilibre de Nash. Lorsque ces conditions sont réunies, un processus évolutif ne peut aboutir qu'à un état stationnaire. L'évolution par sélection naturelle doit alors mener les individus à prendre des décisions globalement cohérentes avec la notion d'équilibre de Nash—ou, comme je l'explique, avec des raffinements de ce concept. J'illustre ceci avec le dilemme du prisonnier répété. Lorsque les intéractions se répètent, la coopération est possible à l'équilibre. L'évolution peut mener à des comportements coopératifs dans des situations ressemblant au dilemme du prisonnier répété. Ceci n'exclut pas la coopération dans le dilemme du prisonnier classique (non-répété) : l'évolution mène les individus à prendre de bonnes décisions sur une classe d'interactions, et non pas à prendre des décisions optimales dans chaque intéraction individuelle.

Je finis ce chapitre en contrastant deux explications à la coopération. Selon la première, la coopération est maintenue parce que réciproque : les individus qui assument un coût pour aider les autres peuvent s'attendre à être aidés en retour. Un autre angle est offert par la théorie du signal honnête. Selon cette approche, la coopération peut être maintenue lorsque le fait d'aider informe sur des qualités sous-jacentes, ce qui affecte le comportement des destinataires du signal. J'explique en quoi il peut être intéressant de combiner ces deux approches, avant d'introduire les articles et projets dévéloppés pendant cette thèse (les chapitres 2 à 5).

Chapitre 2 : La coopération comme signal honnête de long-termisme

Dans ce chapitre, Jean-Baptiste André et moi-même nous intéressons à la coopération entre inconnus ; un phénomène qui semble paradoxal car il implique des coûts sans espoir de réciprocité. L'explication traditionnelle repose sur la notion de réciprocité indirecte : aider ceux qui ont aidé les autres. Le chapitre introduit un nouveau modèle qui considère la coopération comme un signal des préférences temporelles des individus.

Dans ce modèle, il existe un équilibre évolutionnairement stable où les individus long-termistes coopèrent, tandis que les court-termistes font défection. Ce résultat est utilisé pour expliquer la variabilité du comportement coopératif. Par exemple, le modèle aide à comprendre pourquoi les personnes dans des environnements plus aisés sont plus enclines à aider les inconnus et à donner à des œuvres caritatives. Les besoins les plus pressants de ces individus sont satisfaits. En conséquence, ils sont plus patients : ils ont la liberté d'explorer d'autres opportunités, comme investir dans leur réputation ou leur réseau social.

Le modèle explique également pourquoi nous faisons confiance aux personnes sur la base de proxies pour le contrôle de soi, et pourquoi certaines formes de coopération sont plus propices à la confiance. Par exemple, des formes plus subtiles de coopération prennent plus de temps à être observées en moyenne, révélant ainsi une plus grande préférence pour l'avenir ainsi qu'une plus grande motivation à coopérer.

Chapitre 3 : Le jeu de punition répété explique pourquoi et quand nous nous vengeons

L'une des principales explications de la vengeance est la dissuasion de transgressions futures. En cherchant à punir ceux qui nous nuisent, nous pouvons acquérir une réputation de représailles, dissuadant ainsi d'autres individus de nous nuire à l'avenir.

Dans ce chapitre, trois collègues et moi-même étudions ces dynamiques à l'aide du jeu de punition répété, qui met aux prises un acteur et de multiples partenaires. Nous montrons que pour que la coopération ait lieu à l'équilibre, la vengeance doit avoir une fonction dissuasive : l'acteur punit après une transgression, rendant ainsi moins probable que de futurs partenaires transgressent.

Le modèle est ensuite étendu pour expliquer des caractéristiques plus fines de la vengeance humaine. Ceci inclut le coût des excuses : la vengeance n'est pas toujours avantageuse car la représaille peut entraîner une contre-représaille, nuisant aux relations mutuellement bénéfiques. Les excuses jouent alors un rôle crucial. Le modèle prédit que les excuses doivent être suffisamment coûteuses pour dissuader de futures transgressions. Elles doivent être coûteuses lorsque les transgressions profitent à l'offenseur, mais peuvent être gratuites si l'offenseur n'en bénéficie pas.

Le modèle est également étendu pour expliquer le fait de négliger des informations en apparence importantes : parfois, lorsque nous décidons de punir ou ne pas punir, nous négligeons des informations importantes, ou bien nous nous appuyons sur des informations qui devraient être sans pertinence. Un exemple est donné par le phénomène de "chance morale", selon lequel le même comportement nuisible est jugé moins sévèrement en fonc-

tion du résultat plutôt que de l'intention. Nous montrons, à l'aide de notre modèle, que la vengeance doit être basée sur des informations qui sont de notoriété publique, et non sur des connaissances privées. Cela pourrait expliquer pourquoi nous observons des phénomènes comme la chance morale, puisque les résultats d'une action indiviudelle sont plus souvent de notoriété publique que l'intention de l'individu.

Chapitre 4 : Une fuite en avant—Signaler au second ordre peut mener à des signaux exagérés

Certains de nos comportements sont assimilables à des signaux dysfonctionnels. C'est le cas par exemple lorsque tous les membres d'un groupe participent au même rite initiatique. On obtient alors un signal en apparence uniforme : le rite n'informe en rien sur les qualités intrinsèques des individus par rapport à d'autres membres du groupe, puisque chacun accomplit le même rite. D'après la théorie du signal honnête, ce genre de signal uniforme devrait être abandonné.

Dans ce chapitre, Jean-Louis Dessalles et moi-même montrons qu'on peut expliquer les signaux uniformes à travers ce que nous appelons un signal du second ordre. Dans notre modèle, les signaleurs peuvent s'engager dans deux types de manifestations : ils peuvent investir dans une manifestation coûteuse (signal), et ils peuvent exprimer leur indignation envers ceux qui n'y participent pas (signal du second ordre).

L'indignation, dans notre modèle, sert à faire connaître l'investissement de l'individu dans le signal original. Lorsque l'indignation est trop coûteuse pour ceux qui n'investissement pas dans le signal du premier ordre, on peut déduire de son indignation que l'individu indigné doit avoir investi dans le signal du premier ordre, sans avoir besoin de l'observer directement. L'indignation permet alors aux individus d'atteindre un public plus large.

En outre, lorsque les signaleurs sont indignés, l'incitation de tous à envoyer le signal augmente. Nous montrons alors qu'un signal uniforme peut émerger, sous certaines conditions. Nous montrons également que l'indignation peut engendrer une fuite en avant des coûts du signal du premier ordre.

Chapitre 5 : Un modèle endogène de formation d'institutions coopératives

Pour stabiliser la coopération, les sociétés humaines s'appuient sur des institutions. Cependant, celles-ci ne sont pas une solution miracle. Une structure institutionnelle, même parfaite sur le papier, ne peut pas créer de la coopération *ex nihilo*. Pour que les institutions fonctionnent, les individus doivent être dévoués au bien commun. Ils doivent consacrer du temps et des ressources à la conception des règles institutionnelles et à la récompense des agents institutionnels, qui doivent à leur tour résister à la corruption.

Les institutions sont donc des interactions coopératives du second ordre, qui émergent des communautés qu'elles sont censées réguler. Dans ce chapitre, trois collègues et moi-même élaborons un modèle endogène de formation institutionnelle. Nous partons du principe que les dilemmes coopératifs varient en difficulté. Certains sont faciles à résoudre : les incitations réputationnelles sont suffisantes pour pousser les gens à coopérer. D'autres

dilemmes sont difficiles : les incitations réputationnelles ne suffisent pas toutes seules. Dans notre modèle, les individus s'engagent dans une coopération de premier ordre difficile et dans une coopération de second ordre facile. Ce faisant, ils contribuent à une institution qui génère de nouvelles incitations pour la coopération de premier ordre.

Notre modèle suggère que les institutions peuvent être comprises comme des technologies inventées par les humains pour construire l'organisation sociale la plus bénéfique possible, soutenue uniquement par la réputation. Tout comme un système de poulies permet de soulever de lourdes charges, les institutions maximisent le potentiel des incitations réputationnelles, aidant les humains à étendre l'échelle de la coopération.

Notre modèle génère des prédictions spécifiques sur les caractéristiques, les mécanismes sociaux et les variations culturelles des institutions, qui vont dans le sens de régularités observées par les chercheurs et chercheuses en sciences sociales.

Chapitre 6 : Discussion

Pour conclure, je discute des mérites de la modélisation en sciences sociales dans ce chapitre. Je suggère que des modèles simplistes, comme ceux développés au cours de cette thèse, apportent une valeur ajoutée aux sciences sociales—et que leur utilité vient précisément de leur simplicité. La simplicité permet d'illuminer la logique sous-tendant un phénomène de manière claire, compréhensible, et facile à communiquer. Elle favorise la suggestion d'analogies et se révèle être un outil essentiel pour les collaborations interdisciplinaires en sciences sociales.

Chapter 1

Introduction

Now, why do the various animals do what seem to us such strange things , in the presence of such outlandish stimuli? [...] Why do men always lie down, when they can, on soft beds rather than on hard floors? Why do they sit round the stove on a cold day? [...] Not one man in a billion, when taking his dinner, ever thinks of utility. He eats because the food tastes good and makes him want more. If you ask him why he should want to eat more of what tastes like that, instead of revering you as a philosopher he will probably laugh at you for a fool. [...] It takes, in short, what Berkeley calls a mind debauched by learning to carry the process of making the natural seem strange, so far as to ask for the why of any instinctive human act. To the metaphysician alone can such questions occur as: Why do we smile, when pleased, and not scowl? Why are we unable to talk to a crowd as we talk to a single friend? Why does a particular maiden turn our wits so upside-down? The common man can only say, "Of course we smile, of course our heart palpitates at the sight of the crowd, of course we love the maiden, that beautiful soul clad in that perfect form, so palpably and flagrantly made from all eternity to be loved!"

– William James, *The Principles of Psychology*, pp. 386-7, 1890/2007

1.1 Economic game theory 2
1.2 Evolutionary game theory 8
1.3 Two views of reputation 14
1.4 Bridging the gap between reciprocity and signaling 16

Why do we rush to help our friends when they are in need? Why do our hearts and wallets open to fund research for childhood illnesses? Why does our chest tighten with pride when we stand up against an act of injustice, only to subsequently dampen our triumph by affecting an air of modesty? Why does a wave of regret wash over us when we realize we've missed an opportunity to be kind, and why does guilt gnaw at us when we reflect on the world's inequalities?

Just like the questions raised by William James, these questions have obvious answers. "Of course we rush to their aid, they're our friends!" "Of course we are moved by sick children, who wouldn't be?" And so on. When I tell people my work is to go beyond the obvious, they can be surprised, but are more often intrigued, and keen to contribute their own theories to the conversation. When I tell people about my day-to-day however, the conversation usually stops.

Simply mentioning mathematics often does the trick. It just seems like such a strange way of going at things. I wouldn't say people take me for a fool—if anything, I feel a kind of reverence for having mastered the arcane, and socially valued, art. But what could equations possibly have to tell us about real-life human behavior? Only a mind 'debauched by learning' would think of studying human behavior using mathematical models.

This dissertation is organized as an answer to this question. I will begin by introducing game theory, the specific mathematical framework I have relied on during my PhD, and quickly move on to defend the use of game theory for studying human social behavior. Of course, I don't actually mean to convince people in fictitious conversations of the merits of game theory—if anything, it would be presumptuous of me to think that people outside of my specific line of work would care.

Rather, the idea is to provide you, the reader, and comrade in debauchery (I presume), with an accurate picture of my PhD work, and how I have come to understand my main research method. I will start from the ground up, with the prisoner's dilemma, pretending this is the first you hear of this model. I will rely on simple, abstract models and short mathematical demonstrations throughout the introduction—this is just how I think, as one debauched by mathematical learning in particular. I will also heavily rely on my PhD advisors, Jean-Baptiste André and Jean-Louis Dessalles, as well as Moshe Hoffman, who have shaped my understanding of game theory and its application to our social psychology.

1.1 Economic game theory

1.1.1 An example: the prisoner's dilemma

Game theory is a set of analytical tools designed to help us understand how people, companies, or even countries behave in interactive settings—when it matters not only what they do but also what others do (M. Hoffman & Yoeli, 2022; Osborne & Rubinstein, 1994). The models it produces, games, are abstract representations of classes of real-world situations.

For example, the prisoner's dilemma is famously used to study cooperation: classes of

situations in which individual decision-makers can work together for mutual gain, but each has an incentive to behave selfishly. In the simplest possible case, captured by the payoff matrix of Figure 1.1, two players simultaneously decide between two actions: cooperate or defect. Cooperation costs $c > 0$ to either player, and brings benefit $b > c$ to the other player.

		Player 2	
		Cooperate	Defect
Player 1	Cooperate	$b - c$	$-c$
	Defect	b	0

Figure 1.1: **Payoff matrix for a two player prisoner's dilemma**. The payoff for player 1 is given in each of the four possible outcomes: for instance, when player 1 cooperates and player 2 defects, player 1 earns payoff $-c$. Player's 2's payoff are deduced by symmetry. Note that this is the simplest possible version of the dilemma, in which payoffs can be summarized using two mathematical variables b and c, representing the benefit and cost of a cooperative act. More generally, reading the table from top to bottom and left to right, payoffs can be noted R, S, T and P (Axelrod & Hamilton, 1981). A prisoner's dilemma occurs when $T > R > P > S$ and $R > (S + T)/2$. Here, these inequalities follow from $b > c > 0$; we verify that $b > b - c > 0 > -c$ and $b - c > (-c + b)/2$.

Game theory uses mathematics to express its ideas formally. The prisoner's dilemma is an example of a simultaneous game[1]; that is, a game in which decision-makers choose one plan of action once and for all, and these choices are made simultaneously (or equivalently, without knowledge of others' choices). A simultaneous game consists in a set of players N; for each player $i \in N$, a non-empty set of available actions A_i, and a set of obtained payoffs over every possible outcome of the players' individual decision-making. Here, the set of players is $N = \{1, 2\}$. For both players $i = 1, 2$, the set of available actions is $A_i = \{\text{Cooperate}, \text{Defect}\}$, and the set of obtained payoffs is given by the payoff matrix of Figure 1.1—an outcome being an element of the set $A_1 \times A_2$ like {Defect, Cooperate}, in which case player 1 earns b and player 2 earns $-c$.

Mathematical formalism offers many advantages. Among other things, mathematics allow us to precisely define concepts like simultaneous games, to verify the consistency of our ideas using these concepts, and to rigorously explore the logical implications of our assumptions (Smaldino, 2017). However, I must acknowledge that in this context, mathematical formalism can be a hurdle to readability. Most textbooks are composed of a long list of slightly varying game concepts and assumptions, arranged in order of increasing difficulty to cater to students who are learning the ropes. To keep this document reader-friendly, I will introduce only the most relevant concepts and mathematical notations, using illustrative examples like the prisoner's dilemma.

For similar reasons, I have assigned a gender to each player based on the result of a coin toss. In every two-player game presented here, when useful, I will use feminine pronouns (she/her) to refer to player 1, and masculine pronouns (he/him/his) to refer to player 2.

[1]These games are also called strategic or normal-form games.

1.1.2 The core assumptions of economic game theory

Traditionally, game theory rests on the assumption of individual *rationality*. The individual players in a game are assumed to possess complete and consistent preferences over every possible outcome of the game, and to behave according to these preferences. One easy way to introduce such preferences is to give a real value to each possible outcome—that it, to introduce a utility function (the other way is to introduce a preference relationship, which must be complete and transitive to ensure that preferences are complete and consistent).

In the prisoner's dilemma presented above, payoffs then represent players' utility. Using this traditional interpretation of game theory, we can predict individuals' behavior. In this particular case, since cooperation is costly by assumption, we can predict that each individual player will invariably choose defection over cooperation. For example, whatever the action of player 2, playing defect rather than cooperate leads to a utility gain of c for player 1, as evidenced by subtracting the second line of Figure 1.1 to the first line: as a result, we can predict that she will invariably choose to defect.

In addition to rationality, game theory traditionally assumes that players *reason strategically*; that is, that they recursively take into account their knowledge or expectations of other players' behavior. In the prisoner's dilemma, player 2 is implicitly assumed to know the utility function of player 1. He can therefore make the same predictions that we can: player 2 can predict that player 1 will defect in every situation. And player 1 knows that player 2 knows this: she knows that player 2 can expect her to defect. And so on: player 2 knows that player 1 knows that player 2 knows this...

Another way of stating this assumption: rationality (here, each player's utility function) is assumed to be *common knowledge* among the players. Unless specified otherwise, game theory assumes that important parameters (e.g., the payoffs) and other pieces of information (e.g., the result of previous interactions, if any) are common knowledge: players are assumed to know about them (observe them without error), and know that others know, and so on. This assumption is crucial because it shapes individual expectations, and therefore shapes possible behavior at equilibrium (for a general discussion, see Aumann & Brandenburger, 1995). We will return to common knowledge in chapter 3, in which we'll show (in a specific case) that deviations that are common knowledge affect behavior at equilibrium, and deviations that are not common knowledge do not—first, we need to define the word equilibrium, which we'll do in the section just below.

These assumptions, rationality and strategic reasoning, are at the core of game theory since it became a distinctive field in the first half of the twentieth century. Since economists, by in large, continue to rely on this interpretation of game theory (Page, 2022), I refer to it as 'economic game theory' throughout this document.[2]

[2]The reader might be familiar with the terms 'classical game theory', which I'm told is frowned upon by economists, and 'noncooperative game theory', which I find confusing in the context of a dissertation on cooperation.

1.1.3 The Nash equilibrium

The most important concept in game theory is that of the *Nash equilibrium* (Nash, 1951). In a Nash equilibrium, each player does as well as he or she can, taking as given what other players are doing. For instance, in the prisoner's dilemma we have been discussing, {Cooperate, Defect} is not a Nash equilibrium: player 1 would do better by deviating to defection, taken as given player 2's choice to defect. The same is immediately true for {Defect, Cooperate} and {Cooperate, Cooperate}. In contrast, {Defect, Defect} is a Nash equilibrium because neither player can benefit by unilaterally deviating to cooperation.

The Nash equilibrium represents a steady state of the game. In the context of the prisoner's dilemma, it tells us whether a pair of chosen actions can constitute an endpoint of the game, assuming, as we have, that players are rational and reason strategically. As we've just seen, the only possible endpoint is {Defect, Defect}. In classes of situation that resemble the prisoner's dilemma, we predict that individuals will not cooperate: no rational person would chose to lower his or her utility without earning anything in return.

For another example, consider the following sequential game (as in not simultaneous), which we'll call the punishment game and represent using the game tree below, in Figure 1.2. The punishment game involves two players, who play in succession. First, player 1 decides whether to cooperate or defect. Cooperation costs her $c > 0$, and brings benefit $b > 0$ to player 2. Second, player 2 decides whether or not to punish. Punishment costs him $\gamma_2 > 0$, and entails a cost $\gamma_1 > c$ to player 1.

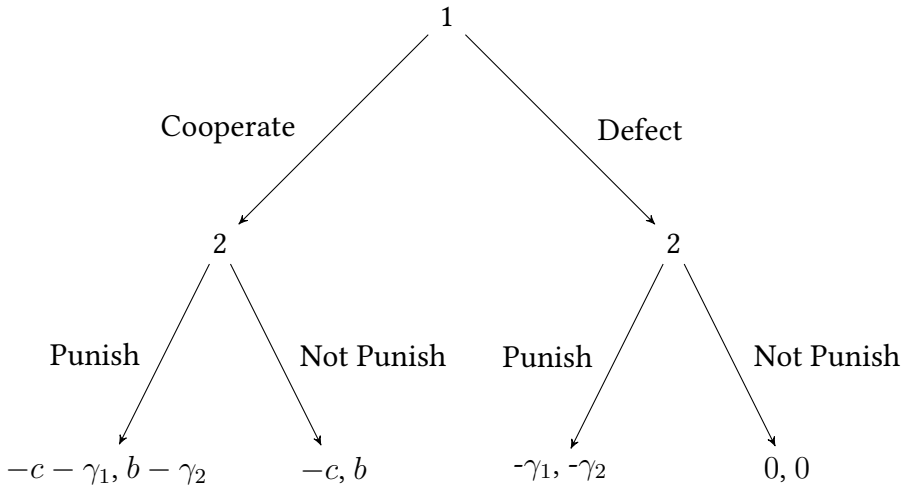


Figure 1.2: **Game tree for a two player punishment game**. Payoffs for both players are noted on the branches of the tree, after both players have played—player 1's payoff on the left of the comma, and player 2's payoff on the right. For instance, if player 1 cooperates and player 2 subsequently punishes, player 1 earns $-c-\gamma_1$, and player 2 earns $b-\gamma_2$.

Player 2 can now decide to punish or not in two separate situations: after player 1 has cooperated, or after she has defected. To completely describe player 2's behavior, we need to explicitly define his strategy—a complete action plan specifying what to do in either case, such as 'Never punish' or 'Punish only if 1 defects'. The Nash equilibrium concept

naturally extends to player strategies (in the previous example, a player's strategy was just her chosen action, so these two concepts didn't need to be distinguished).

The punishment game has two Nash equilibria (three, in fact, if you count {Defect, Punish only if 1 cooperates}). First, {Defect, Never punish} is a Nash equilibrium: neither player can do better since both cooperation and punishment are costly. Second, {Cooperate, Punish only if 1 defects} is also a Nash equilibrium. Player 1 cannot do better by deviating to defection since this would entail saving on c to lose $\gamma_1 > c$. Player 2 cannot do better either by deviating to another strategy, since player 1 will in fact always cooperate in this situation: player 2 always plays after player 1 has cooperated and, following his action plan, opts not to punish and therefore not to pay the cost γ_2.

1.1.4 Subgame perfection and backwards induction

In contrast to the prisoner's dilemma, the punishment game admits a Nash equilibrium in which one player cooperates. Player 1 cooperates because player 2 has committed to punish her otherwise, as per his strategy. This equilibrium is unsatisfying, however, because player 2's commitment to punish is hollow. Since player 1 never defects, player 2 never actually has to follow through on his commitment and pay the cost of punishment.

For this reason, game theorists have come up with a stricter equilibrium concept: that of a *subgame perfect equilibrium* (originally proposed by Selten, 1965). In a subgame perfect equilibrium, each player does as well as he or she can, taking as given the strategies of other players, in every possible situation—*even* after sequences of events that aren't supposed to happen. Since punishment is costly for player 2, the only subgame perfect equilibrium is {Defect, Never punish}.

Formally, this can be proven by a procedure called backwards induction. Let's start from the end: what happens after a player 1 defection, when it's player 2's turn to act? Player 2 has then promised punishment. But this would entail him paying a cost, without affecting what already happened. Since player 2 is expected to behave rationally, he would not punish in this hypothetical situation. Moving backwards one step (hence the name backwards induction), player 1 should defect rather than cooperate: doing so allows her to save on the cost of cooperation, without any later repercussions.

If we assume that player 1 uses backwards induction, we are left with only one possible endpoint: {Defect, Never punish}. As we have seen, {Cooperate, Punish only if 1 defects} is Nash, but not subgame perfect. Strictly speaking, strategic reasoning may lead player 1 to cooperate simply because player 2 has made an idle threat: even though player 1 is allowed to put herself in the shoes of player 2 using strategic reasoning, she is not allowed to consider the hypothetical scenario in which the threat would actually have to be carried out to compute its cost. With the way that economic game theory models players, it is more natural to assume that they use backwards induction, and to use the concept of subgame perfect equilibrium when dealing with sequential interactions. As a result, just as with the prisoner's dilemma, we predict that individuals should refrain from cooperation in situations resembling the punishment game.

1.1.5 The failures of conscious optimization

As we've seen, game theory uses games to model classes of interactive situations, and the concept of Nash equilibrium or the relevant refinement to find these games' endpoints. Because it rests on the assumptions of rationality and strategic reasoning, economic game theory rests on a process of conscious optimization by its players (M. Hoffman & Yoeli, 2022). A valid endpoint is one which resists conscious optimization: if players are in this situation, they will not find any reason to behave differently by deliberately reasoning about their rational objectives and those of their partners.

Based on this view of individual rationality, we have made two predictions. We have predicted that individuals should never cooperate in situations resembling the prisoner's dilemma, and that individuals should use backwards induction to find the best course of action in sequential games—leading them to never cooperate either in situations resembling the punishment game.

One need only turn to the lab to contradict these predictions. In carefully controlled lab experiments designed to reproduce the settings of the prisoner's dilemma, people do cooperate to some extent (for a review, see Raihani & Bshary, 2015). And similarly, they often fail at backwards induction when lab experiments implement sequential games (for a review, see Klein Teeselink et al., 2023). These findings extend to high stakes situations outside the lab, like the popular US show 'The Price is Right'[3] (for more anecdotal evidence on a British show resembling the prisoner's dilemma, see Raihani, 2021, pp. 128-9).

More largely, decades of research has revealed the many ways in which people's decisions depart from economic rationality, and the many ways in which their decision processes rely on heuristics and gut feelings rather than conscious deliberation (for a review, see Page, 2022). With such erroneous assumptions, can economic game theory tell us anything about people's behavior? If people are irrational and conscious optimization fails, what is the point of looking for endpoints?

[3] A few more details for the interested reader. At one point in the show, three contestants compete in what is essentially a sequential game, for an opportunity at winning tens of thousands of dollars. They take turns spinning a wheel that contains all multiples of 5 in the range $5-100$. The winner is the one who gets closest to 100 in total using one or two spins, without going over. After each contestant has spun the wheel once, things start to get more complicated. The first contestant must decide whether to stop, and submit only the score she got in her first spin, or continue, and aim for a higher aggregate score whilst risking going over 100 and losing everything. She must do this knowing the scores of the other contestants, and estimating their chances of stopping or continuing when it will be their turn. Evidence from the last four decades shows that people tend to be overly cautious when in the role of contestant $\#1$ (they stop when subgame perfection says they should spin), and that this is consistent with a failure of backwards induction, wherein people erroneously represent the game as having only one remaining contestant instead of two (Klein Teeselink et al., 2023).

1.2 Evolutionary game theory

1.2.1 The core assumptions of evolutionary game theory

The solution is to consider that something else is doing the optimizing, namely: evolution. Let's return to the prisoner's dilemma, with another interpretation in mind. This time, we consider a population of individual organisms, from an unspecified species—let's say they are guppies, who we'll assume have little ability for conscious optimization. During their life, guppies are paired once with another guppy, and the pair play the prisoner's dilemma. Crucially, payoffs now represent a biological currency, namely their fitness: each time guppy #1 cooperates, this negatively impacts its life expectancy and chances of reproduction, and positively impacts the life expectancy and chances of reproduction of guppy #2.

In addition, we assume that individual strategies are inherited, with some small chance of mutation: the offspring of guppies who cooperate will also cooperate, unless a mutation causes them to instead become defectors. We also assume that both players play the same strategy, since there is no reason for evolution to distinguish between both of these symmetric roles. We then refer to the two remaining candidate endpoints as Cooperate and Defect rather than {Cooperate, Cooperate} and {Defect, Defect}.

Under these assumptions, Defect is again the only possible endpoint. Guppies who cooperate have lower expected fitness than guppies who defect: on average, they have fewer offspring, and, over time, evolution should lead the entire population of guppies to defect (minus a few rare mutants).

Evolution can thus rescue game theory and the search for possible endpoints. In this simple example, evolution should lead to a Nash equilibrium (as we'll soon see, other equilibrium concepts are warranted when things get more complicated). We can dispense with conscious optimization and individual rationality altogether. Instead, we only need two things. First, the game must represent an important aspect of an organism's life, as well as that of its descendants, so that it makes sense to reason in terms of biological fitness. Second, behavior must have some non-degenerate underlying genetic basis (Grafen, 1991). This is the default assumption. By default, a guppy's cooperativeness is susceptible to be controlled by many loci, without strong non-additive effects (e.g., over-dominance), and without strong constraints preventing some possibilities from being explored (i.e., preventing mutation towards alleles that hinder cooperative behavior). By default, we can abstract away from any and all genetic details (Richter & Lehtonen, 2023), and reason, as we did above, as if the very simplest haploid genetic system controlled individual strategy—leading the population to reach an endpoint through the laws of natural selection.[4]

[4]More precisely, we have made an assumption termed the phenotypic gambit by Alan Grafen. For a discussion of its validity, see Grafen, 1991.

1.2.2 Another example: the repeated prisoner's dilemma

How then can we explain cooperation by guppies in their natural environment? (We'll return to humans in the lab in the next section.) The key is to recognize that a crucial aspect is missing from our model. The prisoner's dilemma has been applied to predator inspection behavior, whereby guppies advance in group to obtain information about a sunfish whilst substantially reducing the risk of being eaten (because sunfish are confused by the presence of more than one prey) (Dugatkin, 1988; McCullough, 2008, chapter 4-5).

Advancing with the group is a cooperative behavior: each guppy has an incentive to stay behind, and let the others bear the brunt of the risk. Yet guppies don't advance in one go. Instead, guppies advance one step at a time, with one guppy advancing a little, then another, and so on until the group is close enough for inspection.

This leads us consider a variant of the model, whereby individuals engage in many successive interactions: the repeated prisoner's dilemma. Let's imagine, as we have, two players who engage in a prisoner's dilemma. They choose whether to cooperate, and pay $c > 0$ for the other player to receive $b > c$. Crucially, we don't just end things there. The game can repeat: once an interaction is over, the players again play the prisoner's dilemma with a certain probability r ($0 < r < 1$). And so on. Each time players interact, there's a r percent chance that they will interact again, and a $1 - r$ percent chance that this will be their last interaction.

When $r \geq c/b$, cooperation becomes possible: the repeated prisoner's dilemma has Nash equilibria in which players cooperate to some extent. For instance, let's consider the case in which both players adopt the 'Tit-for-Tat' strategy (Axelrod & Hamilton, 1981): they cooperate in the first interaction, and in all other interactions after that they simply copy the other player's previous action—playing cooperate if their co-player cooperated just before, and defect otherwise.

When $r \geq c/b$, both players' playing Tit-for-Tat is a Nash equilibrium (and, in fact, vice-versa: if Tit-for Tat is a Nash equilibrium, then it must be that $r \geq c/b$). The demonstration is relatively quick. When both players use this strategy, they end up cooperating every time: they start out cooperating, then, every time the game repeats, they chose to cooperate again by copying their co-player's previous action. In contrast, if player 1 decides to defect once (because of the recursive nature of the game, it doesn't matter when), players end up alternating: player 2 will copy that move and defect the next time (while player 1 reverts back to cooperating), provided there is a next time, then player 1 will copy that move (while player 2 reverts back to cooperating), and so on. By defecting once instead of sticking to the game plan, player 1 ends up saving on the cost of cooperation every other time, starting in this interaction, but loses on the benefit of cooperation every other time, starting in the next interaction. The payoff differential is then equal to: $c/(1-r^2) - rb/(1-r^2) \propto c - rb$. Tit-for-Tat is a Nash equilibrium if and only if this difference is negative, which is equivalent to $r \geq c/b$.

The repeated prisoner's dilemma shows that cooperation can emerge when the chances of repeat interaction r are high enough. Its creator, Robert Axelrod, famously used it to explain cooperation in the trenches dug during World War I (Axelrod, 1982/2006, chapter

4). Despite the unforgiving context, an extended stalemate led to favorable conditions, at least in terms of probability of repetition r: the same units faced each other over months at a time, and were able to devise ways to cooperate.

The key is reciprocity (Trivers, 1971): when interactions are repeated, individuals can respond to a partner defecting. By engaging in a reciprocal form of cooperation like the one prescribed by the Tit-for-Tat strategy, whereby defection is followed by one reciprocal defection, individuals create an incentive not to defect in the first place. For guppies, evolution has found an analogous solution: guppies advance towards predators only as long as other also advance when its their turn (Dugatkin, 1988).

1.2.3 Adaptation-executers

> Individual organisms are best thought of as adaptation-executers rather than as fitness-maximizers. (Tooby & Cosmides, 1992)

Evolutionary game theory can thus explain cooperative predator-inspection by guppies, and the emergence of cooperation in the many long-term interactions of our species (Lehmann et al., 2022). How however can we explain the existence of cooperation when experiments set $r = 0$?

The short answer is that we don't need to. Evolution has not led to optimal strategies for the prisoner's dilemma so that humans could behave optimally in the lab, in perfect instantiations of the mathematical model. Rather, as the quote above makes bare,[5] evolution has embedded optimal strategies into elements of our psychology, pushing people to make good decisions for their fitness across the cooperative interactions that are typical for our species—situations that we can model using the repeated prisoner's dilemma.

'Adaptation-executers' will sometimes make mistakes. Our adaptive psychological mechanisms, through which evolution has embed optimal strategies, can be triggered out of context—including in artificial lab settings that mimic key features of cooperation. Without such misfires, lab experiments wouldn't be of much use: it's by tricking our cooperative psychology into spilling over from real life to the lab, that we are able to document its many interesting features (Axelrod, 1982/2006, p. 28; E. Hoffman et al., 1998, M. Hoffman and Yoeli, 2022, p.26-27). For subjects, cooperating when $r = 0$ is then arguably a small, and evolutionarily inevitable, mistake (for counter-arguments, see Raihani & Bshary, 2015).

1.2.4 Qualitative predictions

> The best way to test a model is to break its assumptions and see what happens. (M. Hoffman & Yoeli, 2022, p. 48)

[5] A quote that is conveniently displayed right in front of the PhD student office of the *Evolution and Social Cognition* team, in which I've spent most of my time.

The longer answer to one-shot cooperation in the lab begins by recognizing the limits of our approach, and setting the right level of analysis and predictive targets. Evolutionary game theory involves two kinds of abstraction: not only are we considering entire classes of situations all at once in a highly abstract model called a game, but we are also abstracting away from any and all details of the evolutionary process through which organisms may come to execute adaptive strategies (Richter & Lehtonen, 2023). By now, it should be clear that we are using game theory to answer ultimate (Why do humans cooperate?) rather than proximate (How do humans decide to cooperate?) questions (Tinbergen, 1963/2010)—in the preceding discussion, I did not stop once to name a proximate mechanism (empathy?) through which evolution may have embed an adaptive strategy onto our psychology.

As a result of the crudeness of our abstract models, we can only derive qualitative predictions from them. The repeated prisoner's dilemma predicts that cooperation can emerge if interactions are repeated sufficiently often; at the very least, we need $r > 0$. To test that model, we can follow the advice above and break that assumption: when $r = 0$, we expect less cooperation, not 0 cooperation, since our model is too crude to yield any quantitative prediction.

This prediction is verified in the lab: people cooperate more in the repeated prisoner's dilemma ($r > 0$) than in the simple, one-shot prisoner's dilemma ($r = 0$) (Gächter & Falk, 2002). These lab experiments manage to trigger our evolved cooperative psychology: a different result would have led us to question the validity of these experiments at least (if not the model). More broadly, seeing the important role played by r in the mathematical model, and the crucial manners in which r can vary during our social lives—for instance, from one relationship to the next—we can make a stronger prediction (and verify it, in defense of the model): all other things being equal, people's cooperative motivation should vary with the potential for repeat interactions (Balliet et al., 2017; Barclay, 2013; Delton et al., 2023).

1.2.5 Subgame perfection and noise

Before we move on to the limits of the repeated prisoner's dilemma and the Tit-for-Tat strategy, let's briefly return to our other model, the punishment game. We do so because this game, and the concept of subgame perfection, will be useful in Chapter 3.

Recall that punishment game had a cooperative Nash equilibrium, in which player 1 cooperates because player 2 threatens to punish her if she defects. Player 2's threat turned out to be idle, though—he never has to carry it out because when things go according to plan player 1 cooperates—making this equilibrium unsatisfying. To eliminate this unsatisfying possibility, we introduced the notion of subgame perfect equilibrium, which was justified by assuming that players could engage in a complex process of reasoning called backwards induction.

There is however another justification for subgame perfection: noise. Sometimes, things do not go according to plan. If there is even the tiniest of positive probabilities ε that player 2 finds himself in a situation where player 1 has defected (to the right of our game tree represented in Figure 1.2), then he will have to carry out his threat ε percent of the time, and

lose $\varepsilon\gamma_2 > 0$ on average. To save on this cost, player 2 should deviate to never punishing.

Noise can occur if player 1 makes mistakes in implementation (even though her strategy is to cooperate, she accidentally defects with probability ε), if player 2 makes mistakes in interpretation (even though player 1 cooperates, it looks like she defected to player 2, for instance because he is relying on imperfect information from a third-party) or a mixture of both (for instance if there is an attenuating circumstance under which player 1 might defect, which player 2 might fail to see). When we allow for such rare mistakes, only subgame perfect equilibria are evolutionarily stable (Selten, 1983).

1.2.6 Evolutionarily stable strategies

Introducing evolution has allowed us to do three things. First, as we just saw, evolution can rescue the concept of subgame perfection, as long as it makes sense to also introduce rare mistakes. Second, as we saw when introducing evolution for the first time, evolutionary game theory can rescue the very notion of endpoint. Although evolutionary game theory and related notions are often used to look at dynamics (for a review, see Traulsen & Glynatsi, 2023), I have exclusively relied on a static perspective during my PhD.[6] In the simple prisoner's dilemma, Defect is both the only Nash equilibrium and the only possible evolutionary endpoint.

Third, introducing evolution and the repeated prisoner's dilemma have allowed us to explain cooperation. To allow for a more gradual presentation, I have however made an omission: although Tit-for-Tat is a Nash equilibrium, it is not a stable evolutionary endpoint. This is because, there are strategies that do just as well, or rather, with this evolutionary perspective, mutants that can invade via neutral drift. For instance, a mutant who always cooperates, unconditionally, does just as well as the resident who plays Tit-for-Tat, both against itself and against the resident: whatever the pairing, individuals end up cooperating every time. This is a problem for the stability of this particular cooperative strategy: over time, this neutral mutant can increase in frequency. This is also a problem for the stability of cooperation: once there are too many unconditional cooperators, it becomes beneficial to exploit them, and a mutant who plays Defect can invade.[7]

Because of such considerations, evolutionary game theorists have come up with the concept of an *evolutionarily stable strategy* (or ESS, for short Maynard Smith & Price, 1973), which, like subgame perfection, is another refinement of the Nash equilibrium. A strategy is evolutionarily stable if it cannot be invaded by a mutant, even when the mutant is neutral. Once it is fixated in a population, alternative strategies cannot increase in frequency: this occurs if and only if (a) the resident strategy is Nash, and (b) it performs strictly better against neutral mutants than neutral mutants do against themselves (unlike Tit-for-Tat,

[6] Complemented only with computer simulations in the case of chapter 4.

[7] In fact, there is a cycle, since Tit-for-Tat performs better against Defect. If we run evolutionary simulations with just these three strategies, we obtain oscillations consistent with cyclical invasion. Historically, similar oscillatory dynamics were put into evidence in simulations with noise and other strategies (e.g., those successively developed by Nowak & Sigmund, 1992, 1993; Wu & Axelrod, 1995). For a concise explanation, see Nowak, 2006, chapter 5; for a rapid review, see McCullough, 2008, chapter 5.

which does not do better against unconditional cooperation than unconditional cooperation against itself, allowing the latter to increase in frequency via neutral drift).

1.2.7 Side-stepping the folk theorem

The instability of Tit-for-Tat is not the only problem with the repeated prisoner's dilemma, however. Like all games that are repeated, the prisoner's dilemma admits many different Nash and subgame perfect equilibria; in fact, any level of cooperation can occur in equilibrium when the probability of repetition r is high enough (Fudenberg & Maskin, 1986). To see why, consider the strategy whereby players cooperate once every blue moon, playing according to the rules of Tit-for-Tat every n times, starting the first time, and agreeing to defect in all other interactions. If player 1 defects once, player 2 will defect the next time both players are supposed to cooperate, which is in n interactions, leading player 1 to defect in $2n$ interactions, and so on. We obtain the same type of back-and-forth as in the demonstration for Tit-for-Tat (see section 1.2.2). A similar argument shows that this strategy is Nash if $r^n \geq c/b$, that is, when r is sufficiently close to 1.

This general property of repeated game has been called the folk theorem. It arises because of the circular dimension of games like the repeated prisoner's dilemma: players can arbitrarily agree on a rule (e.g., ignore $n-1$ out of n interactions) and sustain cooperation based on that rule—once the arbitrary rule is in play, it doesn't make sense to cooperate the other times. It seems extremely unsatisfying: if everything is possible, what can game theory actually tell us about individual behaviors (Boyd, 2006)?

As it names suggests, the folk theorem has been around for a while (it originates in the informal tradition of the early game theory community, long before written proofs like Fudenberg and Maskin, 1986). Both economists (e.g., Fudenberg & Maskin, 1990) and evolutionary biologists (e.g., André & Day, 2007) have shown how the concept of evolutionary stability can drastically reduce the set of possible endpoints in (slight twists on) the repeated prisoner's dilemma, to only include the highest levels of reciprocal cooperation. More generally, another more biologically relevant equilibrium concept is needed (André, 2023). The ESS concept only tells us whether an endpoint is stable to mutation, assuming it has been reached; it does not tell us whether evolution can actually converge towards such an endpoint. By taking a more biological approach, we can eliminate unrealistic possibilities like the one just above, and better explain human social behaviors using evolutionary game theory.

In my thesis however, my advisors and I have adopted a different approach. We have not sought a general way to reduce the set of equilibria, based on a more biologically sound understanding of the optimization process of evolution. Instead, we have focused on a specific type of human decision, in an approach I am tempted to call psychological. How do we decide whom to trust (chapters 2, 5) and whom not to mess with (chapter 3)? Based on what information? What happens when people use imperfect proxies to make such decisions (chapter 4)? What makes up a person's *reputation*, and what are the minimal assumptions we can make to study reputation and reputation-based decisions? What do these models reveal about our social psychology?

Before I can introduce the projects I have worked on during my PhD, and detail their simplifying assumptions (and how these assumptions reduce the set of equilibria), we need to make one last detour, and see how game theorists define and model reputation.

1.3 Two views of reputation

1.3.1 Indirect reciprocity

Humans frequently help strangers, incurring costs with no hope of benefiting their kin, and no hope of reciprocation. To explain cooperation among such one-shot partners ($r = 0$), one approach is to extend the repeated prisoner's dilemma. Instead of modeling interactions between two privileged partners, who have a high chance of meeting again, we can model interactions between infinitely many different partners, occurring in infinitely many separate rounds (implicitly, we're assuming $r = 1$): in each round, an individual plays with a new partner, and her choice (cooperate or defect) is observed with probability p by the entire population, including her future partners. A similar argument to the one in section 1.2.2 shows that cooperation is possible when behavior is sufficiently likely to be observed—when $p \geq c/b$.

This approach has been called indirect reciprocity (Alexander, 1987). In equilibrium, cooperation must be indirectly reciprocal in the sense that individuals who help a partner today will be helped tomorrow, by another partner in some future round. Put differently, cooperation is possible as long as individuals abide by a simple maxim: 'help those who have helped others' (Nowak & Sigmund, 2005).

Of course, this maxim is unclear. What does it mean to have helped others? Does it mean to have never wavered, or only rarely, one day out of n? What succession of observed choices will lead an individual to be ascribed 'good reputation', and receive cooperation from her partners; what succession of choices will lead her to bad reputation? Just like the original model, this extension admits many different Nash equilibria, and therefore many different conventions through which individuals can decide to reward observed cooperative behavior. In response, scientists have investigated certain simple rules (e.g., Leimar & Hammerstein, 2001; Nowak & Sigmund, 1998; Panchanathan & Boyd, 2003) to test their evolutionary stability, and looked at the stability of simple reputation systems more systematically (Ohtsuki & Iwasa, 2006).

Why, however, should we even care about others' past behavior? Our simple model suffers from a more fundamental problem. Reputation is given a technical definition. Individuals are characterized by a list of binary choices, corresponding to each time they were observed cooperating or defecting; a stable reputation system is then a way in which people may agree to reward based on such lists. This definition is divorced from the intuitive, and individually functional, definition of reputation: a set of evaluations about the qualities of an individual (Giardini, Balliet, et al., 2021), whose purpose is to predict that individual's future behavior (Roberts et al., 2021).

1.3.2 Honest signaling

Why, then, should we care about others' past behavior? Why should we evaluate the qualities of our cooperative partners based on whether they cooperated with others? Why should their past predict our future?

The other approach to reputation is based on the theory of honest signaling. With origins in ethology (Grafen, 1990; Zahavi, 1975) and the social sciences (Spence, 1974; Veblen, 1899/1973), this theory explains otherwise puzzling behaviors through the information that they convey to observers. In signaling games like the one below, behaviors can come to inform onlookers about socially desirable qualities when they are more costly, or less beneficial (Számadó et al., 2023), for the less socially desirable. Honest signaling theory thus provides a natural framework for studying reputation-based decisions—the question then becomes: what is the relevant quality, and why does it influence payoffs?

To see how this approach works, let's take an example. (We'll delay answering the questions above one section.) Let's adapt a model developed by Gintis, Smith and Bowles (2001).[8] Individuals in an infinite population are characterized by a hidden quality q, which is continuously distributed in an interval $[q_L, q_H]$. We call these individuals signalers. Signalers decide whether to send a signal by paying a cost $c(q)$, where c is a decreasing function of the individual's quality q.

We assume that each signal is observed with a certain positive probability p by other individuals, which we call receivers. For simplicity of exposition, receivers are part of another infinite population. Receivers can chose one signaler to follow, in which case they obtain payoff $f(q')$, where f is an increasing function of the signaler's quality q' (if they do not choose anyone to follow, they obtain null payoff). We assume $\mathbf{E}(f) > 0$, such that on average, it is beneficial to follow an individual chosen at random. Receivers can follow based on observed signals (more precisely, they can either chose to follow no one, or chose to follow at random without using the signal, or chose to follow an individual chosen at random among all observed senders). Each time a receiver follows a signaler, the chosen signaler gains b.

An honest signaling equilibrium occurs when the signal discriminates between relatively high quality signalers and relatively low quality signalers; that is, when there exists a non-trivial threshold $\hat{q} \in (q_L, q_H)$ such that signalers send the signal if they are of relatively high quality ($q \geq \hat{q}$), and do not send the signal if they are of relatively low quality ($q < \hat{q}$). The signal is then informative: based on the signal, receivers can infer that a sender is of relatively high quality $q \geq \hat{q}$, and will thus on average be a relatively good ally to choose.

We can show that there exists a unique signaling ESS (a unique value for $\hat{q}$) as long as $c(q_L) > b$ (for the demonstration, see Lie-Panis & Dessalles, 2023). The picture is reversed with respect to the repeated prisoner's dilemma: while reciprocal cooperation requires that

[8] The two main differences with the original model: senders are observed with any positive probability p rather than with certainty, which we show has no effect; and individual quality is continuously distributed rather than binary, which we show allows for the evolution of signaling under less restrictive conditions. For the demonstration, see Lie-Panis and Dessalles, 2023.

the benefit of receiving cooperation be larger than the cost of cooperation ($b > c$), here, signaling is stable when it is prohibitively costly for individuals of minimum quality q_L, ensuring that these individuals are excluded in equilibrium—and that the signal be at least minimally informative.

Put differently: the picture is reversed when we allow individuals to chose their partners, rather than decide between cooperating and defecting with a partner we have chosen for them. Everything is as if individuals (the signalers) compete in a biological market to attract partners (the receivers) (Noë & Hammerstein, 1994). This market provides an incentive to signal at a higher level than others. Based on this model, we predict high levels of competitive helping, whereby individuals try to outdo each other by appearing more generous than others (Barclay, 2013; Roberts, 1998; Zahavi, 1995).

Note that in the above model, signaling and the ensuing social relationships are highly asymmetric. Signalers send the signal once, and may potentially attract an infinity of followers; in contrast, if receivers observe an individual signaling at a higher level, they can switch and chose to follow that individual instead without paying any cost. Its creators had in mind a public form of cooperation (e.g., contribution to a public good), and purely one-shot partnerships (Gintis et al., 2001). In contrast, when partnerships are more symmetric (Dessalles, 2014) and switching partners more costly (Geoffroy et al., 2019), the level of help is more constrained.

Put differently, the above model is best used to study another decision than the one we posed at the beginning of this section: should I follow this individual—not can I trust him or her to cooperate with me specifically? When deciding whom to trust, we pay less attention to public displays such as the one in the above model than to more subtle behaviors, which are better suited to inform about the individual's trustworthiness (Bliege Bird and Power, 2015; Bliege Bird et al., 2018; see also Dhaliwal et al., 2022). Accordingly, recent models show that dyadic help can function as a signal of commitment to a specific relationship (Quillien, 2020), or stake in a particular partner (Barclay et al., 2021).

1.4 Bridging the gap between reciprocity and signaling

During my PhD, I have relied on both of the classical frameworks for modeling reputation. I have relied on indirect reciprocity to study social behaviors; in particular, cooperation. I have relied on honest signaling to ground my models in reputation-based decisions. In particular, the decision to trust: when can we infer that an individual is worthy of our trust; and how, therefore, may individuals manage their cooperative reputation?

I have used this dual approach to explain: the variability of trust and cooperation (chapter 2); retaliatory punishment and its apparent quirks (chapter 3); seemingly dysfunctional public displays of commitment (chapter 4); and how institutions extend the reach of human cooperation (chapter 5).

We will delve into each of these questions later, at the beginning of each chapter. Below, I give a more technical introduction to each chapter, to give a little more detail about my approach, and how I have dealt with some of the issues we have discussed (the folk theorem,

the definition of reputation...).

1.4.1 Reputation-based trust and cooperative reputation

In chapter 2, Jean-Baptiste André and I use both frameworks at once: indirect reciprocity and honest signaling. We develop a model of reputation-based trust and dyadic cooperation. In lieu of the repeated prisoner's dilemma with varying partners that we introduced in section 1.3.1, we use a repeated trust game. The trust game is an asymmetric version of the prisoner's dilemma in which first, a 'chooser' decides whether to trust a 'signaler' based on the signaler's reputation, and second, the signaler decides whether to cooperate. Our model is thus as close as possible to the typical model for indirect reciprocity, whilst allowing for reputation-based partner choice.

Since we are interested in cooperation with many rare partners rather than with one privileged partner (in contrast to Barclay et al., 2021; Quillien, 2020), we assume that individuals vary in terms of their underlying time preferences (see also Roberts, 2020). This is because in such a context, cooperation involves a present-future trade-off: cooperative individuals help a partner to enhance their reputation today, and be trusted by other partners tomorrow. Future-oriented individuals gain more from investing in their reputation; these individuals can be thought of as individuals with high social capital (Jordan & Rand, 2017), making it more beneficial to maintain a good reputation; or as individuals with less pressing needs (Boon-Falleur et al., 2022; Mell et al., 2021), making it less costly to wait on reputational investments.

We show that cooperation emerges as a signal of an individual's time preferences. This allows us to (finally!) answer the questions we raised in section 1.3.2, when we introduced honest signaling. As long as individuals have differing time preferences and as long as these preferences are *sticky*, an individual's past behavior reveals her time preferences, and therefore informs about her propensity to cooperate tomorrow—the past then predicts the future (see also André, 2010; Leimar, 1997).

To limit the set of possible equilibria, we assume that choosers only retain the most recent piece of information. Our model applies when people focus on other's most recent cooperative acts when deciding whether to trust them—arguably, once again, this is relevant for cooperation with many rare partners and similar stakes, and not so much for cooperation with a privileged partner or variable stakes. Under this assumption, we show that our model admits two evolutionary endpoints: the cooperative ESS on which we focus, and one trivial ESS where choosers never trust and signalers never cooperate.

This last result is general, since we do not attempt to completely resolve the problem of multiple equilibra, and signaling games always have at least two equilibria due to their circular nature (if receivers do not use the signal, signalers have no reason to send, and vice-versa). In chapters 4 and 5, we also have an unwanted trivial equilibrium, in addition to the equilibria we are interested in; in chapter 3, we adopt a different tactic to 'side-step' the folk theorem, as explained just below.

1.4.2 Retaliatory reputation

In the model presented in chapter 2, individuals can invest in their cooperative reputation by helping a dyadic partner, prompting others to trust them. In chapter 3, Moshe Hoffman, Christian Hilbe, Bethany Burum and I look at the benefits of maintaining a reputation for retaliating against defectors, in order to deter others from defecting.

To do so, we consider a repeated version of the punishment game. Recall that threats to punish defectors were not credible in that game, preventing the establishment of an equilibrium with cooperation by player 1, and retaliatory punishment by player 2. By repeating the game between one individual in the role of player 2 and many different partners in the role of player 1, we once again introduce a present-future trade-off: player 2 invests in his retaliatory reputation by punishing a partner who defects, and deters future partners from defecting.

To model reputation in such a repeated game, we take a page from Mailath and Samuelson (2006).[9] Reputation and reputational benefits in future encounters only matter for the unique individual who takes on the role player 2, whom we call the actor. We assume that only the actor plays all rounds of the repeated game—in the language of Mailath and Samuelson, the actor is a long-run player. In contrast, we assume that all individuals who take on the role of player 1 are short-run players (we call these individuals partners): they only play one round of the game, in which they decide between cooperation and defection with the actor.

In contrast to the chapter just before, we do not assume that actors vary in an underlying quality, nor do we assume that partners only retain certain pieces of information. Partners decide to cooperate or defect based on the entire history of partner-actor interactions—reputation takes on the more 'technical' definition that is typical of indirect reciprocity models (see section 1.3.1).

Multiple equilibria are then possible. To side-step this issue, we look at the entire set of subgame perfect equilibria for which partners cooperate along the outcome path (i.e. in every round, as long as things go according to plan). We demonstrate properties that are valid for the entire set of these cooperative equilibria. More precisely, we show that, for cooperation to be enforced, the actor must punish unexpected transgressions, and such retaliatory punishment must serve to deter transgressions by future partners.

We also extend our model, by introducing new assumptions (e.g., allowing partners to apologize after a transgression) or relaxing other ones (e.g., assuming that transgressions are imperfectly observed). We show how these extensions impact our general results: by turning on and off certain assumptions, we turn on and off either of the features outlined above. This allows us to highlight other necessary features of retaliatory punishment.

For instance, we show that common knowledge is necessary. When a partner transgresses and the transgression is common knowledge, the actor punishes, and doing so deters future partners from transgressing. In contrast, when a partner transgresses but it looks like she did not transgress to future partners (but not to the actor, who has both

[9] This framework would have been useful for chapter 2 as well, but I discovered it afterwards.

pieces of information—see chapter 3), the actor does not punish even though she knows that a transgression did occur; doing so would be unnecessary.

This chapter highlights that even without honest signaling, indirect reciprocity remains a fruitful framework for reputation and reputation-enhancing behaviors. Even without grounding our model in a reputation-based decision (what quality should partners look for in the actor to infer the best course of action?), we are able to derive several qualitative predictions from general principles (e.g., common knowledge).

1.4.3 Second-order signaling

In chapter 4, Jean-Louis Dessalles and I investigate second-order signaling (Dessalles, 2018). Humans often communicate about the actions of others, to praise or blame them (Anderson et al., 2020). We start from the idea that such behaviors can be conceptualized as second-order signals. On the one hand, praise and blame are about praiseworthy and contemptible actions by definition, which are actions that are susceptible to reveal important information about the actor—signals (or at least cues) of the actor's moral qualities.[10] On the other hand, praise and blame are communicative actions—they're not just about passively making inferences about others, as a 'receiver'. They can entail social benefits (e.g., connecting with someone who admires the same people) and social costs (e.g., angering someone who sees our blame as unjustified); to the extent that these payoffs vary in a predictable manner, praise and blame can be informative to others.

In fact, directly related to blame, moral condemnation of others' immoral behavior has been shown to function as a signal of one's own moral behavior (Jordan et al., 2017). More indirectly related to praise, choosing cooperative allies over ones that are more able to provide immediate benefits has been shown to signal one's own cooperativeness (Dhaliwal et al., 2022).

In chapter 4, we investigate second-order signaling using the model introduced in section 1.3.2 (adapted from Gintis et al., 2001). We concentrate on condemnation of others' lack of signaling—we call this outrage—and look at what happens when onlookers can infer individuals' qualities from their investment in outrage. To do so, we simply assume that the cost of expressing outrage is prohibitively high for non-senders. Outrage is then automatically honest. In an extension to the model, we show that we can go one step further: outrage is honest when individuals who express outrage but do not invest in the signal—hypocrites—are preferential targets of outrage (but the question then becomes: why target hypocrites first?).

We look at the consequences of this assumption, namely, the consequences of assuming that individuals use imperfect proxies for reputation in this asymmetric model. Because the model only admits two types of players (signalers and receivers), the dynamics are a bit different than those described above: when receivers use such proxies, signalers compete to use first- and second-order signals.[11]

[10] For the distinction between cues and signals, and two discussions of how cues can evolve into signals, see Biernaskie et al., 2018; Pinsof, 2023.

[11] Another solution could have been to add another game played by receivers, in order to form more sym-

1.4.4 Extending cooperative reputation to two games

In chapter 5, Jean-Baptiste André, Nicolas Baumard, Léo Fitouchi and I look at two more closely related behaviors: dyadic help on the one hand, and investment in a collective action on the other. These are both cooperative actions. Following the model in chapter 2, we expect both of them to inform about an individual's time preferences, i.e. the individual's general propensity and ability to invest in her cooperative reputation.

We thus adapt the model presented in chapter 2, to include a collective action. We show that both dyadic help and investment in the collective action emerge as a signal of underlying time preferences. In equilibrium, choosers can then infer help from investment in the collective action—they have a reason to use the second behavior as proxy for trustworthiness. This echoes the longstanding observation that people's behavior in one cooperative interaction predicts their behavior in others (e.g., Peysakhovich et al., 2014), and a recent model and experiment in which investment in a collective action is explicitly shown to signal trustworthiness (Barclay & Barker, 2020).

metric and/or long-term partnerships. Receivers would then have an incentive to follow senders in the original game purely as a second-order signal, designed to attract partners in that second game. The questions, and problems to address, are then: why should this reveal important qualities, e.g., receiver's cooperativeness, and why use this indirect second-order signal rather than a more direct proof of cooperativeness? See Dhaliwal et al. (2022) for some elements of response.

Chapter 2

Cooperation as a signal of time preferences

Objectives and summary

Humans frequently help strangers, incurring costs with no hope of benefiting their kin, and no hope of reciprocation. Building on the repeated prisoner's dilemma, many models explain the *existence* of cooperation among such one-shot partners in terms of indirect reciprocity: help those who have helped others. When a population plays by the rules of indirect reciprocity, helping entails a present-future trade-off. Individuals who incur costs to help a one-shot partner acquire a good reputation today, and can hope to be helped tomorrow, if the game repeats with another one-shot partner. As long as the chance of repeat encounters is high enough, cooperation through indirect reciprocity is an equilibrium.

How, however, can we explain the *variability* of cooperation among strangers? In the paper below, published in 2022, we introduce a model of cooperation as a signal of time preferences. We assume that individuals vary in their underlying time preferences—the value they give to future payoffs relative to present ones.

We show the existence of an ESS in which future-oriented individuals cooperate, and present-oriented individuals defect. We use this result to explain some of the variability of cooperative behavior. In particular, our models helps explain why people in more affluent environments tend to help strangers more often in field studies (Nettle et al., 2011; Zwirner & Raihani, 2020), and report giving more to charity in surveys (Korndörfer et al., 2015). Indeed, in more affluent environments, individuals' most pressing needs are met, allowing them to explore other opportunities, like investing in their reputation or social network (Boon-Falleur et al., 2022; Mell et al., 2021). All other things being equal, these individuals should be more patient.

Our model also explains why we trust people based on proxies for self-control, including whether they indulge in victimless pleasures of the senses (Fitouchi et al., 2022), and why certain forms of cooperation lend themselves to more trust. In the model for instance, more

subtle forms of cooperation take longer to be observed on average, and therefore reveal higher preference for the future and higher cooperative motivation (see also Bliege Bird et al., 2018; Quillien, 2020).

The paper, printed below, is followed by a supplementary information, in which we detail the mathematical model and its results.

PROCEEDINGS B

royalsocietypublishing.org/journal/rspb

Research

Cite this article: Lie-Panis J, André J-B. 2022 Cooperation as a signal of time preferences. *Proc. R. Soc. B* **289**: 20212266. https://doi.org/10.1098/rspb.2021.2266

Received: 12 October 2021
Accepted: 25 March 2022

Subject Category:
Behaviour

Subject Areas:
behaviour, evolution, theoretical biology

Keywords:
cooperation, trust, time preferences, evolution, costly signalling

Author for correspondence:
Julien Lie-Panis
e-mail: jliep@protonmail.com

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.5923872.

THE ROYAL SOCIETY PUBLISHING

Cooperation as a signal of time preferences

Julien Lie-Panis[1,2,3] and Jean-Baptiste André[1]

[1]Institut Jean Nicod, Département d'études cognitives, Ecole normale supérieure, Université PSL, EHESS, CNRS, 75005 Paris, France
[2]LTCI, Télécom Paris, Institut Polytechnique de Paris, 91120 Palaiseau, France
[3]Université de Paris, EURIP Graduate School for Interdisciplinary Research, 75004 Paris, France

JL-P, 0000-0001-7273-7704; J-BA, 0000-0001-9069-447X

Many evolutionary models explain why we cooperate with non-kin, but few explain why cooperative behaviour and trust vary. Here, we introduce a model of cooperation as a signal of time preferences, which addresses this variability. At equilibrium in our model (i) future-oriented individuals are more motivated to cooperate, (ii) future-oriented populations have access to a wider range of cooperative opportunities, and (iii) spontaneous and inconspicuous cooperation reveal stronger preference for the future, and therefore inspire more trust. Our theory sheds light on the variability of cooperative behaviour and trust. Since affluence tends to align with time preferences, results (i) and (ii) explain why cooperation is often associated with affluence, in surveys and field studies. Time preferences also explain why we trust others based on proxies for impulsivity, and, following result (iii), why uncalculating, subtle and one-shot cooperators are deemed particularly trustworthy. Time preferences provide a powerful and parsimonious explanatory lens, through which we can better understand the variability of trust and cooperation.

1. Introduction

Human cooperation is inherently variable. Cooperation varies with the individual. We are not all equally likely to help an unrelated stranger in the field or in the laboratory, and report differing levels of cooperative behaviour in surveys [1–15]. Cooperation is also a function of historical and social context. Social trust tends to be lower in poorer countries, and in the aftermath of conflict or other dramatic events [16–21]. For the same interaction, the norm may even be to cooperate in one society, and defect in another [22,23]. Finally, the value of cooperation itself is variable. We place more trust in spontaneous and inconspicuous cooperators than we do in individuals who help others in deliberate or overt fashion [24–30].

Evolutionary biologists and game theoreticians explain the evolution of cooperation with non-kin based on the principle of reciprocity. We trust and help those who have helped us [31,32] or others, and have thus acquired a trustworthy reputation [33–36]. These approaches, however, are chiefly concerned with explaining the existence of cooperation, and rarely attend to its variable nature. In most models helpful behaviour varies because of exogenous noise [37–40]. Cooperative variability remains an open question: we are unable to predict who is more prone to help, where cooperation is more likely to emerge, and what determines its informational value.

The variable nature of cooperation may be studied following a framework introduced by Leimar [41]. His model is based on the assumption that individuals derive differing pay-offs from cooperation, and may thus be differentially motivated to help others (see also [42]). In line with honest signalling theory [43,44], an individual's behaviour in cooperative encounters will then reveal her private pay-offs, and therefore her future cooperative intentions—making it reasonable to trust others based on past behaviour [41,45,46].

Leimar's model provides the general framework for our study. At first glance however, his central assumption seems unrealistic. Virtually all the resources or

© 2022 The Author(s) Published by the Royal Society. All rights reserved.

services that we acquire on our own may be obtained via cooperative exchanges; it is therefore difficult to conceive that some of us could systematically benefit more from cooperation than others. In order to better understand the who, the where and the what of cooperation, we must first explain why individual pay-offs should vary *in general*.

One answer to these questions may lie in differences in individual time preferences. Laboratory and field experiments performed in a diversity of contexts reveal that individuals can be distinguished according to their level of preference for immediate versus future rewards [47–50]. These time preferences are stable in the short to medium term [51,52], and across similar decisions [53,54].

Interindividual differences could originate from adaptive phenotypic plasticity, as harsher environments make future rewards more uncertain and/or present needs more pressing, and select for stronger preference for the present [55–61]. At a fundamental level, cooperation entails paying immediate costs (to help others) and, following the principle of reciprocity, receiving delayed benefits (in the form of future help) [34,41,45,46,62]. In theory, an individual's time preferences should equivalently affect *all* the pay-offs she derives from cooperative encounters.

In this paper, we formally explore the hypothesis that time horizon is the underlying cause of the variability of human cooperation. We develop a mathematical model of cooperation in which individuals are characterized by a hidden discount rate, which remains constant throughout their life, and affects all future pay-offs. Individuals face strangers in a cooperative setting, and may use their reputation to discriminate between trustworthy and exploitative partners. Help emerges as an honest signal of time preferences in our model. Variation of time horizon ensures behavioural variability at evolutionary equilibrium, which stabilizes cooperation [63–67]. In addition, assuming that individual time preferences vary allows us to account for all three dimensions of cooperative variability.

First, we predict that more future-oriented individuals should be more prone to help. At equilibrium in our model, trustworthy partners are individuals whose time horizon surpasses a certain threshold. This result conforms with empirical data. Many studies report a positive correlation between individual time horizon and cooperation [68–72], although it should be noted that some of the evidence is inconclusive [9,62]. Our first result also helps explain interindividual cooperative variability. In surveys and field studies, individual cooperation is associated with environmental affluence [2,6,7,11,12,15]—a variable that closely aligns with time horizon [48–50,53,73–76]. Time preferences have been found to mediate the relationship between environmental affluence and individual investment in collective actions [12].

Second, we predict that more future-oriented populations should have access to a wider range of stable cooperative opportunities. In surveys and field studies, average cooperation and trust are associated with collective wealth [6,11,12,16,18]. Our model offers two complementary explanations for these observations. Following our first result, we expect higher aggregate cooperation when many individuals are future-oriented. Following our second result, we expect cooperation and trust to emerge in a wider range of contexts when the population distribution of time preferences shifts towards the future.

Third, we predict that cooperation should be a more informative signal of time preferences when observation is unlikely, or when the cost–benefit ratio is low. Our theory may explain why we place more trust in helpful partners who maintain a low profile or make impromptu decisions [24–30]. Inconspicuous cooperators are indeed less likely to be observed and, since spontaneous cooperators help more frequently [28,30,77], they stand to gain less from the average encounter. Both behaviours reveal strong preference for the future in our model, and therefore strong cooperative motivation.

Table 1. Pay-offs for the trust game.

		Signaller	
		cooperate	defect
Chooser	accept	$(b, r-c)$	$(-h, r)$
	reject	(0, 0)	(0, 0)

2. Cooperating with strangers

We model cooperative encounters following a trust game with two roles (adapted from [78]). The game consists in two stages: in the first, the 'Chooser' may either accept the 'Signaller' or reject partnership with that prospective partner, putting an early end to the interaction. Accepted Signallers reap reward r.

Partnership is only advantageous with trustworthy Signallers. In the second stage, the Signaller may cooperate with the Chooser, or opt to defect. Cooperation costs c and benefits the Chooser, who earns b. By contrast, defection is free and harms the Chooser, who loses h. We assume cooperation is net beneficial for Signallers: $r > c$. Pay-offs are summarized in table 1.

When in the role of Chooser, individuals always face a strange Signaller, with whom they have never interacted before, and of whom they possess no privileged information. Choosers may however condition their play on their partner's reputation. Signallers are observed with probability p, and error σ. Individuals form a trustworthy or exploitative image of Signallers based on the most recent observation (figure 1).

Signallers have varying time preferences. We assume that individuals engage in a large number of cooperative interactions throughout their life, and that lifetime pay-offs can be calculated following a discounted utility model [47]. A Signaller's time preference is represented by her discount rate δ: obtaining pay-off π at future time t is worth $(1/(1+\delta))^t \times \pi$ now. δ is positive and fixed at birth, by drawing in the population distribution of discount rates. The closer δ is to zero, the more an individual is future-oriented.

In the electronic supplementary material, we give a full description of the model, and provide a thorough equilibrium analysis. Below we focus on the conditional trust and trustworthiness (CTT) strategy profile, which is defined in relation to a threshold discount rate $\hat{\delta}$, and whereby, throughout their life, (i) Choosers accept strangers given trustworthy reputation, and reject them given exploitative reputation; and (ii) Signallers cooperate when their discount rate is smaller than $\hat{\delta}$, and defect when their discount rate is larger than $\hat{\delta}$. Demonstrations for this strategy profile are detailed in the Material and methods section.

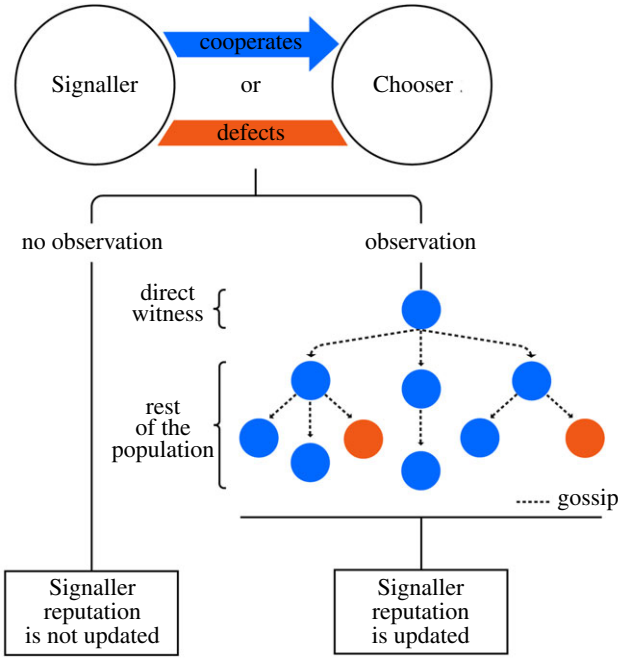


Figure 1. Reputation formation. Signaller behaviour is observed with probability p and error σ by the entire population in our model and $\left(0 < p < 1 \text{ and } 0 < \sigma < \frac{1}{2}\right)$. This may be interpreted to reflect direct observation by one or several witnesses, and rapid social transmission of information (gossip) [34,79,80]. Direct observers mention their observation to several acquaintances, who in turn inform their acquaintances, etc. When this process is rapid relative to social interactions, all individuals receive information by the next trust game. Error σ can thus be seen to reflect the noisiness of social transmission: when a Signaller is observed cooperating, $1-\sigma$ per cent of individuals form a trustworthy image of that Signaller, and σ per cent an exploitative image (and vice-versa with defection). We assume that new information replaces old information, and that individuals never forget. In future trust games, partners of that Signaller may condition their trust on (their private view of) her reputation. (Online version in colour.)

3. Results

(a) Cooperative equilibrium

We show that CTT is an evolutionarily stable strategy (ESS) if and only if [81]:

$$\hat{\delta} = p \times \left[(1-\sigma)\left(\frac{r}{c}-1\right) - \sigma\frac{r}{c}\right] \tag{3.1}$$

and

$$\frac{\sigma h}{\sigma h + (1-\sigma)b} < \mathbf{P}(\delta < \hat{\delta}) < 1 - \frac{\sigma b}{\sigma b + (1-\sigma)h}. \tag{3.2}$$

Equation (3.1) specifies the strategy profile under study, by specifying the value of the threshold discount rate. Since $\hat{\delta}$ must be positive for cooperation to actually occur, we deduce an upper bound on error σ:

$$\sigma < \frac{(r/c)-1}{2(r/c)-1}. \tag{3.3}$$

Cooperation is stabilized by variation of individual time preferences. Following equation (3.2), CTT is an ESS when at least $\sigma h/(\sigma h + (1-\sigma)b)$ per cent of individuals have a discount rate which is smaller than $\hat{\delta}$, and therefore cooperate when in the Signaller role; and at least $\sigma b/(\sigma b + (1-\sigma)h)$ individuals are above that threshold, and therefore defect. Both fractions are positive, increasing functions of error σ: cooperation is evolutionarily stable in our model when behaviour at equilibrium is sufficiently variable [63–67], and error sufficiently small [80].

(b) Who: cooperators are sufficiently future-oriented individuals

At equilibrium, trustworthy Signallers are individuals whose discount rate is inferior to $\hat{\delta}$. When individuals play CTT, Signallers who cooperate pay immediate cost c and increase their chances of facing well-disposed partners in the future, once they have been observed. The value of establishing and maintaining a trustworthy reputation $\hat{\rho}$ depends on the average delay Signallers have to wait before they are observed, which is proportional to $\Delta t = 1/p$, and on the benefit of consistently cooperating instead of defecting after observation, $\hat{\beta} = (1-\sigma)(r-c) - \sigma r$.

We can in fact write: $\hat{\rho} = p[(1-\sigma)(r-c) - \sigma r] = \hat{\beta}/\Delta t$. Since $\sum_{t=1}^{\infty}(1/(1+\delta))^t = 1/\delta$, an individual's social future may be represented by a single trust game whose pay-offs are discounted with rate $1/\delta$. Signallers cooperate at equilibrium if and only if the value they attach to gaining $\hat{\rho}$ their entire future social life exceeds the immediate cost of cooperation c—mathematically, $\delta < \hat{\delta} \Leftrightarrow 1/\delta \times \hat{\rho} > c$. Everything is as if trustworthy Signallers pay c to secure benefit $\hat{\beta}$ in a future trust game which occurs with probability p. (Note that $\hat{\rho}$ tends towards $r-c$ when p tends toward 1 and σ towards 0; when observation is highly faithful and certain, trustworthy Signallers pay c in order to gain approximately $r-c$ their entire future life, with quasi-certainty.)

(c) Where: future-oriented populations have access to a wider range of cooperative opportunities

When average discount rates are low, equation (3.2) is verified for a wide range of possible parameter values, including when $\hat{\delta}$ is small—i.e. when the cost–benefit ratio r/c of cooperation is low, and/or when observation is unlikely (small p) or unreliable (large σ). Even the most demanding forms of cooperation are stable in sufficiently future-oriented populations.

(d) What: cooperation reveals underlying time preferences

Cooperation evolves as a signal of time preferences. At equilibrium, when a Signaller cooperates, she reveals that her discount rate is under $\hat{\delta}$. What's more, cooperation emerges as a signal, and not merely a cue, of Signaller time preferences [82]. Cooperation is selected because it affects Choosers' behaviour: future-oriented Signallers cooperate in order to increase their chances of being trusted in the future, effectively paying c now in order to gain $\hat{\rho} > 0$ their entire future life. By contrast, cooperation cannot evolve in the absence of such an effect. If for instance Choosers accept whatever the information they are presented with, cooperative Signallers do not increase their relative chances of being trusted in the future; in such a case, they would pay c now to gain nothing later.

In addition, the informative value of cooperation increases when $\hat{\delta}$ decreases. When a Signaller helps given small cost–benefit ratio r/c or unlikely observation p, she reveals that

her temporal discount rate must be small—and that she could therefore potentially be trusted in a wide array of cooperative interactions.

4. Discussion

In this paper, we have shown that cooperation can be understood as a signal of time preferences, using a formal model. We derived three predictions from our model: (i) future-oriented individuals should be more motivated to cooperate, (ii) future-oriented populations should have access to a wider range of cooperative opportunities, and (iii) cooperators who reveal stronger preference for the future should inspire more trust. These results shed light on the variability of cooperative behaviour and trust.

(a) Environment and cooperation

Results (i) and (ii) help explain why individual and aggregate cooperation are associated with environmental affluence in large representative surveys [6,11,12,16,18], in field studies [2,7,15] and a natural experiment [8]—since people in more privileged circumstances tend to display stronger preferences for the future [48–50,53,73–76] (see also [83]).

Due to adaptive phenotypic plasticity, the environment in which we grow up and live may in fact directly fashion our time preferences; and therefore, fashion our cooperative inclinations [55–57]. Evolutionary models show that it is adaptive to be more present-oriented in adverse circumstances, i.e. when future rewards are uncertain [58,59], or when present needs are pressing [60,61]. Interindividual differences in time preferences and cooperation could thus arise from an adaptive plastic response to one's environment, for either of these reasons. In support of this hypothesis, a recent study finds that present biases partially mediate the relationship between affluence and investment in collective actions [12], while a meta-analytic review finds a negative correlation between early-life stress and self-reported cooperation [14].

It should be noted that the evidence from behavioural experiments is mixed. While some economic games have produced a positive association between affluence and cooperation [2,3,6,11,17,23], other laboratory experiments yield the opposite association [1,4,5,10], or no effect at all [9,13]. The previously mentioned meta-analysis finds no significant overall correlation [14]. In some instances, this discrepancy is attributable to small sample sizes [6,13]. More largely, the generalizability and ecological validity of many laboratory experiments can be questioned; in particular, when only one economic game is performed. Recent studies find that measures derived from a single economic game do not correlate with self-reported cooperation or real-life behaviour, but that a general factor based on several games does [84,85].

(b) Trust depends on revealed time preferences

Result (iii) helps explain why we infer trustworthiness from traits that appear unrelated to cooperation, but happen to predict time preferences. We trust known partners and strangers based on how impulsive we perceive them to be [86,87]; impulsivity being associated with both time preferences and cooperativeness in laboratory experiments [88–93]. Other studies show we infer cooperative motivation from a wide variety of proxies for partner self-control, including indicators of their indulgence in harmless sensual pleasures (for a review see [94]), as well as proxies for environmental affluence [95,96].

Time preferences further offer a parsimonious explanation for why different forms of cooperation inspire more trust than others. When probability of observation p or cost–benefit ratio r/c are small in our model, helpful behaviour reveals large time horizon—and cooperators may be perceived as relatively genuine or disinterested. We derive two different types of conclusion from this principle.

(c) Inconspicuous cooperation

First, time preferences explain why we trust our partners more when they cooperate in an inconspicuous manner (see also [26,29,97,98]). In our model, the average delay cooperators have to wait before help can be profitable varies like $\Delta t = 1/p$. Given smaller probability of observation p, helpful individuals literally reveal they are able to wait for a longer amount of time. By contrast, when immediate rewards are added (e.g. when blood donors are promised payment), help becomes much less informative; and less valuable to the more genuinely prosocial [99].

In particular, only acutely future-oriented individuals will help when observability p is tiny. Their cooperation is akin to a 'message in a bottle': a powerful demonstration of their intrinsic cooperativeness, which, so long as $p \neq 0$, will eventually be received by others. This could explain why some of us cooperate in economic games that are designed to make our help anonymous [100], so long as we assume that anonymity is never absolutely certain (see also [101]).

(d) Spontaneous cooperation

Second, time preferences explain why we trust our partners more when they cooperate spontaneously—when their behaviour appears more natural, unhesitant, intuitive, uncalculating or underlain by emotion [24,25,27,28,30]. Since they help their partners more frequently [28,30,77], including when defection is tempting, more spontaneous cooperators enjoy lower expected pay-offs in the typical encounter (see also [102]). Greater spontaneity could thus indicate willingness to help given smaller values of r/c; and therefore stronger preference for the future.

(e) Time preferences and other partner qualities

Our analysis has fixated on time preferences. This is somewhat arbitrary. Many other characteristics affect our cooperative interests, and are revealed by our social behaviour—underlying costs and benefits [28,78], revelation probability [97], and, when interacting with known associates, specific commitment to the shared relationship [29,62,98,103,104] (this latter dimension is absent in our model). These qualities shape our strategic interests in a given social context: we stand to gain more from cooperation when it involves a partner we know and are committed to; and when it occurs in a social network we value and are embedded in, where we should enjoy higher observability and pay-offs. Yet, context changes fast. We can help a close friend today, and donate anonymously tomorrow.

In contrast to other partner qualities, time preferences appear remarkably stable. Communication of time preferences is likely to be a fundamental element of human cooperation. It may even underlie other facets of our social

life. The larger our time horizon, the more likely we are to invest in our social surroundings, via dyadic help as well as collective actions or policing. Contribution to public goods [105] and prosocial punishment [78], which function as signals of cooperative intent, may also rely on communication of time preferences.

5. Material and methods

This section gives a sketch of the evidence regarding the conditional trust and trustworthiness strategy profile, in a simplified setting. For a full description of the model, and a thorough equilibrium analysis, see the electronic supplementary material.

Two types of players engage in a repeated trust game: Choosers and Signallers. In each round, a Chooser faces a Signaller she has never encountered before. She may first accept or reject the Signaller, putting an early end to the interaction. If accepted, the Signaller reaps reward r, and may then cooperate (play action C) or defect (play D). Cooperation involves the Signaller paying cost c for the Chooser to gain b; defection is free, and harms the Chooser, who loses h.

Choosers may condition their strategy on their private view of the Signaller's reputation. Each time a Signaller acts, she is observed with probability p. When a Signaller is observed cooperating, $1-\sigma$ per cent of Choosers receive information $\mathcal{T}$, correctly indicating that the Signaller behaved in a trustworthy manner; and the remaining σ per cent receive information $\mathcal{E}$, falsely indicating exploitative behaviour (and vice-versa with defection). We assume new information replaces old information.

Signallers may condition their strategy on their discount rate δ. To simplify things, we assume here that Signallers play a stationary strategy ('always cooperate', or 'always defect'), and that they are initially certain to be accepted (before the first observation). We relax both these assumptions in the electronic supplementary material, and obtain the same results. δ is fixed at birth, by drawing in a continuous probability distribution which characterizes the Signaller population. Signallers engage in a large number of rounds of the repeated trust game, a pay-off t rounds in the future being discounted by factor $(1/(1+\delta))^t$ now.

According to the conditional trust and trustworthiness (CTT) strategy profile, throughout their life, (i) Choosers accept given trustworthy reputation $\mathcal{T}$, and reject given exploitative reputation $\mathcal{E}$; and (ii) Signallers cooperate if their discount rate is smaller than a certain threshold value $\hat{\delta}$, and defect if their discount rate is larger than $\hat{\delta}$. We show that CTT is an evolutionarily stable strategy (ESS) [81] under the conditions set by equations (3.1)–(3.2), by computing equilibrium and deviation pay-offs for Signallers first, and Choosers second.

(a) Signaller equilibrium pay-offs

We consider a Signaller of discount rate δ. Let Π_C and Π_D be the lifetime discounted pay-off she can expect from playing always cooperate and always defect, respectively. We show that when the value of $\hat{\delta}$ is given by equation (3.1), the Signaller stands to strictly lose from deviation from CTT.

Let us first calculate Π_C. When the Signaller always cooperates, she gains $r-c$ every round she is accepted. She will eventually be observed, from which point she can expect to be accepted $1-\sigma$ per cent of the time in equilibrium, in rounds where she is paired with a Chooser who has (correctly) received information $\mathcal{T}$. In other words, she eventually gains pay-off $\Pi_C^\infty = \sum_{t=0}^{\infty}(1/(1+\delta))^t(1-\sigma)(r-c) = ((1+\delta)/\delta)(1-\sigma)(r-c)$, starting from the point of first observation.

In the initial round however, she is certain to be accepted, and gain $r-c$. Observation affects her pay-offs starting in the next round, which are discounted by factor $1/(1+\delta)$: if she is observed, she gains Π_C^∞ starting the next round, if not, she continues to gain pay-off Π_C. In other words, we have:

$$\Pi_C = r - c + \frac{p \times \Pi_C^\infty + (1-p) \times \Pi_C}{1+\delta}.$$

From which we deduce:

$$\Pi_C = \left(r - c + \frac{p \times \Pi_C^\infty}{1+\delta}\right) \times \frac{1+\delta}{p+\delta}.$$

We can apply an analogous reasoning to calculate Π_D. When the Signaller always defects, she gains r every round she is accepted. After the first observation, the Signaller can expect to be accepted σ per cent of the time, when paired with a Chooser who has (incorrectly) received information $\mathcal{T}$. She eventually gains: $\Pi_D^\infty = \sum_{t=0}^{\infty}(1/(1+\delta))^t \sigma r = ((1+\delta)/\delta)\sigma r$. Starting from the initial round, she therefore gains:

$$\Pi_D = r + \frac{p \times \Pi_D^\infty + (1-p) \times \Pi_D}{1+\delta}.$$

Which yields:

$$\Pi_D = \left(r + \frac{p \times \Pi_D^\infty}{1+\delta}\right) \times \frac{1+\delta}{p+\delta}.$$

By comparing both expressions, we deduce that the Signaller strictly benefits from cooperation if and only if the cost of cooperating now is smaller than the benefit of receiving Π_C^∞ instead of receiving Π_D^∞ in the future, with probability p:

$$\Pi_D < \Pi_C \Leftrightarrow c < p \times \frac{\Pi_C^\infty - \Pi_D^\infty}{1+\delta}.$$

And, by replacing Π_C^∞ and Π_D^∞ by their values, we deduce the logical equivalence:

$$\Pi_D < \Pi_C \Leftrightarrow \delta < p \times \left[(1-\sigma)\left(\frac{r}{c} - 1\right) - \sigma\frac{r}{c}\right].$$

Under condition (3.1), the Signaller therefore always stands to strictly lose from deviation from CTT. If her discount rate δ is smaller than $\hat{\delta}$, she strictly gains on average from cooperating her whole life instead of defecting her whole life; if conversely, $\delta > \hat{\delta}$, she strictly benefits from defecting. Note that CTT does not prescribe behaviour for the Signaller when her discount rate is precisely equal to the threshold. Here, we neglect this possibility, based on the fact that the population distribution of discount rates is continuous (we come back to this in the electronic supplementary material).

(b) Chooser equilibrium pay-offs

We show that in equilibrium, Choosers stand to strictly lose from deviation from CTT when equation (3.2) is verified. Let us first consider a Chooser faced with information $\mathcal{T}$. If the Chooser rejects the Signaller, she gains nothing; if she accepts, she gains b if the Signaller plays C and loses h if the Signaller plays D. Her expected benefit is then equal to: $\mathbf{P}(C|\mathcal{T}) \times b + \mathbf{P}(D|\mathcal{T}) \times (-h) = \mathbf{P}(C|\mathcal{T})(b+h) - h$. Accepting given $\mathcal{T}$ is therefore strictly beneficial iff:

$$\mathbf{P}(C|\mathcal{T}) > \frac{h}{b+h}.$$

Let $\tau = \mathbf{P}(C) = \mathbf{P}(\delta < \hat{\delta})$ be the equilibrium probability that the Signaller is trustworthy. Following Bayes' rule, $\mathbf{P}(C|\mathcal{T}) = \mathbf{P}(\mathcal{T}|C)/\mathbf{P}(\mathcal{T}) \times \tau$. The above inequality can be rewritten as:

$$\frac{1-\sigma}{\tau(1-\sigma) + (1-\tau)\sigma} \times \tau > \frac{h}{b+h}.$$

This is equivalent to:

$$\tau > \frac{\sigma h}{\sigma h + (1-\sigma)b}. \quad (5.1)$$

Let us now consider a Chooser faced with information $\mathcal{E}$. An analogous calculation shows that rejecting given $\mathcal{E}$ is strictly beneficial iff:

$$\mathbf{P}(C|\mathcal{E}) < \frac{h}{b+h}$$

Using Bayes' rule, we find: $\mathbf{P}(C|\mathcal{E}) = \mathbf{P}(\mathcal{E}|C)/\mathbf{P}(\mathcal{E}) \times \tau = \sigma/(\tau\sigma + (1-\tau)(1-\sigma)) \times \tau$. By replacing in the above inequality, we deduce that rejection given $\mathcal{E}$ is strictly beneficial iff:

$$\tau < 1 - \frac{\sigma b}{\sigma b + (1-\sigma)h}. \quad (5.2)$$

Combining equations (5.1) and (5.2), and using $\tau = \mathbf{P}(\delta < \hat{\delta})$, we deduce equation (3.2). Under that condition, Choosers therefore stand to strictly lose from deviation from CTT. We deduce that CTT is an ESS under the conditions set by equations (3.1) and (3.2): any mutant is strictly counter-selected. We show in the electronic supplementary material that we in fact have an equivalence; CTT is an ESS if and only if both equations are verified.

Data accessibility. The data are provided in electronic supplementary material [106].

Authors' contributions. J.L.-P.: conceptualization, investigation, methodology, validation, writing—original draft, writing—review and editing; J.-B.A.: conceptualization, methodology, supervision, validation, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. This study was supported by the EUR FrontCog grant no. ANR-17-EURE-0017 and funding from the EURIP Graduate School for Interdisciplinary Research.

Acknowledgements. We thank an anonymous reviewer, Pat Barclay, Léo Fitouchi and Moshe Hoffman for their insightful feedback, and Clara Lie for her graphic prowess.

References

1. Piff PK, Kraus MW, Côté S, Cheng B, Keltner D. 2010 Having less, giving more: the influence of social class on prosocial behavior. *J. Pers. Soc. Psychol.* **99**, 771. (doi:10.1037/a0020092)
2. Nettle D, Colléony A, Cockerill M. 2011 Variation in cooperative behaviour within a single city. *PLoS ONE* **6**, e26922. (doi:10.1371/journal.pone.0026922)
3. McCullough ME, Pedersen EJ, Schroder JM, Tabak BA, Carver CS. 2013 Harsh childhood environmental characteristics predict exploitation and retaliation in humans. *Proc. R. Soc. B* **280**, 20122104. (doi:10.1098/rspb.2012.2104)
4. Chen Y, Zhu L, Chen Z. 2013 Family income affects children's altruistic behavior in the dictator game. *PLoS ONE* **8**, e80419. (doi:10.1371/journal.pone.0080419)
5. Guinote A, Cotzia I, Sandhu S, Siwa P. 2015 Social status modulates prosocial behavior and egalitarianism in preschool children and adults. *Proc. Natl Acad. Sci. USA* **112**, 731–736. (doi:10.1073/pnas.1414550112)
6. Korndörfer M, Egloff B, Schmukle SC. 2015 A large scale test of the effect of social class on prosocial behavior. *PLoS ONE* **10**, e0133193. (doi:10.1371/journal.pone.0133193)
7. Andreoni J, Nikiforakis N, Stoop J. 2017 Are the rich more selfish than the poor, or do they just have more money? A natural field experiment [Working Paper]. (w23229):w23229. See http://www.nber.org/papers/w23229.pdf.
8. Akee R, Copeland W, Costello EJ, Simeonova E. 2018 How does household income affect child personality traits and behaviors? *Am. Econ. Rev.* **108**, 775–827. (doi:10.1257/aer.20160133)
9. Wu J, Balliet D, Tybur JM, Arai S, Van Lange PAM, Yamagishi T. 2017 Life history strategy and human cooperation in economic games. *Evol. Hum. Behav.* **38**, 496–505. (doi:10.1016/j.evolhumbehav.2017.03.002)
10. Amir D, Jordan MR, Rand DG. 2018 An uncertainty management perspective on long-run impacts of adversity: the influence of childhood socioeconomic status on risk, time, and social preferences. *J. Exp. Soc. Psychol.* **79**, 217–226. (doi:10.1016/j.jesp.2018.07.014)
11. Schmukle SC, Korndörfer M, Egloff B. 2019 No evidence that economic inequality moderates the effect of income on generosity. *Proc. Natl Acad. Sci. USA* **116**, 9790–9795. (doi:10.1073/pnas.1807942116)
12. Lettinga N, Jacquet PO, André JB, Baumand N, Chevallier C. 2020 Environmental adversity is associated with lower investment in collective actions. *PLoS ONE* **15**, e0236715. (doi:10.1371/journal.pone.0236715)
13. Stamos A, Lange F, Huang SC, Dewitte S. 2020 Having less, giving more? Two preregistered replications of the relationship between social class and prosocial behavior. *J. Res. Personal.* **84**, 103902. (doi:10.1016/j.jrp.2019.103902)
14. Wu J, Guo Z, Gao X, Kou Y. 2020 The relations between early-life stress and risk, time, and prosocial preferences in adulthood: a meta-analytic review. *Evol. Hum. Behav.* **41**, 557–572. (doi:10.1016/j.evolhumbehav.2020.09.001)
15. Zwirner E, Raihani N. 2020 Neighbourhood wealth, not urbanicity, predicts prosociality towards strangers. *Proc. R. Soc. B* **287**, 20201359. (doi:10.1098/rspb.2020.1359)
16. Nunn N, Wantchekon L. 2011 The slave trade and the origins of mistrust in Africa. *Am. Econ. Rev.* **101**, 3221–3252. (doi:10.1257/aer.101.7.3221)
17. Balliet D, Lange P. 2013 Trust, punishment, and cooperation across 18 societies a meta-analysis. *Perspect. Psychol. Sci.* **8**, 363–379. (doi:10.1177/1745691613488533)
18. Albanese G, de Blasio G. 2013 Who trusts others more? A cross-European study. *Empirica* **41**, 803–820. (doi:10.1007/s10663-013-9238-7)
19. Besley T, Reynal-Querol M. 2014 The legacy of historical conflict: evidence from Africa. *Am. Pol. Sci. Rev.* **108**, 319–336. (doi:10.1017/S0003055414000161)
20. Bjørnskov C. 2007 Determinants of generalized trust: a cross-country comparison. *Public Choice* **130**, 1–21. (doi:10.1007/s11127-006-9069-1)
21. Rohner D, Thoenig M, Zilibotti F. 2013 Seeds of distrust: conflict in Uganda. *J. Econ. Growth* **18**, 217–252. (doi:10.1007/s10887-013-9093-1)
22. Henrich J, Heine SJ, Norenzayan A. 2010 The weirdest people in the world? *Behav. Brain Sci.* **33**, 61–83. (doi:10.1017/S0140525X0999152X)
23. Henrich J *et al.* 2010 Markets, religion, community size, and the evolution of fairness and punishment. *Science* **327**, 1480–1484. (doi:10.1126/science.1182238)
24. Critcher CR, Inbar Y, Pizarro DA. 2013 How quick decisions illuminate moral character. *Soc. Psychol. Personal. Sci.* **4**, 308–315. (doi:10.1177/1948550612457688)
25. Gambetta D, Przepiorka W. 2014 Natural and strategic generosity as signals of trustworthiness. *PLoS ONE* **9**, e97533. (doi:10.1371/journal.pone.0097533)
26. Bird RB, Power EA. 2015 Prosocial signaling and cooperation among Martu hunters. *Evol. Hum. Behav.* **36**, 389–397. (doi:10.1016/j.evolhumbehav.2015.02.003)
27. Everett JAC, Pizarro DA, Crockett MJ. 2016 Inference of trustworthiness from intuitive moral judgments. *J. Exp. Psychol.: General.* **145**, 772–787. (doi:10.1037/xge0000165)
28. Jordan JJ, Hoffman M, Nowak MA, Rand DG. 2016 Uncalculating cooperation is used to signal trustworthiness. *Proc. Natl Acad. Sci. USA* **113**, 8658–8663. (doi:10.1073/pnas.1601280113)

29. Bird RB, Ready E, Power EA. 2018 The social significance of subtle signals. *Nat. Hum. Behav.* **2**, 452–457. (doi:10.1038/s41562-018-0298-3)
30. Levine EE, Barasch A, Rand D, Berman JZ, Small DA. 2018 Signaling emotion and reason in cooperation. *J. Exp. Psychol.: General* **147**, 702–719. (doi:10.1037/xge0000399)
31. Trivers RL. 1971 The evolution of reciprocal altruism. *Q Rev. Biol.* **46**, 35–57. (doi:10.1086/406755)
32. Axelrod R, Hamilton WD. 1981 The evolution of cooperation. *Science* **211**, 1390–1396. (doi:10.1126/science.7466396)
33. Alexander RD 1987 *The biology of moral systems*. New York, NY: Aldine de Gruyter.
34. Nowak MA, Sigmund K. 1998 Evolution of indirect reciprocity by image scoring. *Nature* **393**, 573–577. (doi:10.1038/31225)
35. Panchanathan K, Boyd R. 2003 A tale of two defectors: the importance of standing for evolution of indirect reciprocity. *J. Theor. Biol.* **224**, 115–126. (doi:10.1016/S0022-5193(03)00154-1)
36. Ohtsuki H, Iwasa Y. 2006 The leading eight: social norms that can maintain cooperation by indirect reciprocity. *J. Theor. Biol.* **239**, 435–444. (doi:10.1016/j.jtbi.2005.08.008)
37. Boyd R. 1989 Mistakes allow evolutionary stability in the repeated Prisoner's Dilemma game. *J. Theor. Biol.* **136**, 47–56. (doi:10.1016/S0022-5193(89)80188-2)
38. Nowak MA, Sigmund K. 1993 A strategy of win–stay, lose–shift that outperforms tit-for-tat in the Prisoner's Dilemma game. *Nature* **364**, 56–58. (doi:10.1038/364056a0)
39. McNamara JM, Barta Z, Houston AI. 2004 Variation in behaviour promotes cooperation in the Prisoner's Dilemma game. *Nature* **428**, 745–748. (doi:10.1038/nature02432)
40. McNamara JM, Barta Z, Fromhage L, Houston AI. 2008 The coevolution of choosiness and cooperation. *Nature* **451**, 189–192. (doi:10.1038/nature06455)
41. Leimar O. 1997 Reciprocity and communication of partner quality. *Proc. R. Soc. Lond. B* **264**, 1209–1215. (doi:10.1098/rspb.1997.0167)
42. Boyd R. 1992 The evolution of reciprocity when conditions vary. In *Coalitions and alliances in humans and other animals,* (eds A Harcourt, F De Waal, FBM de Waal). Oxford, UK: Oxford University Press.
43. Zahavi A. 1975 Mate selection—a selection for a handicap. *J. Theor. Biol.* **53**, 205–214. (doi:10.1016/0022-5193(75)90111-3)
44. Grafen A. 1990 Biological signals as handicaps. *J. Theor. Biol.* **144**, 517–546. (doi:10.1016/S0022-5193(05)80088-8)
45. Leimar O, Hammerstein P. 2001 Evolution of cooperation through indirect reciprocity. *Proc. R. Soc. Lond. B* **268**, 745–753. (doi:10.1098/rspb.2000.1573)
46. André J. 2010 The evolution of reciprocity: social types or social incentives?. *Am. Nat.* **175**, 197–210. (doi:10.1086/649597)
47. Frederick S, Loewenstein G, O'Donoghue T. 2002 Time discounting and time preference: a critical review. *J. Econ. Lit.* **40**, 351–401. (doi:10.1257/jel.40.2.351)
48. Kirby KN, Godoy R, Reyes-García V, Byron E, Apaza L, Leonard W, Perez E, Vadez V, Wilkie D. 2002 Correlates of delay-discount rates: evidence from Tsimane' Amerindians of the Bolivian rain forest. *J. Econ. Psychol.* **23**, 291–316. (doi:10.1016/S0167-4870(02)00078-8)
49. Tanaka T, Camerer CF, Nguyen Q. 2010 Risk and time preferences: linking experimental and household survey data from Vietnam. *Am. Econ. Rev.* **100**, 557–571. (doi:10.1257/aer.100.1.557)
50. Amir D, Jordan MR, McAuliffe K, Valeggia CR, Sugiyama LS, Bribiescas RG, Snodgrass JJ, Dunham Y. 2019 The developmental origins of risk and time preferences across diverse societies. *J. Exp. Psychol.: General* **149**, 650. (doi:10.1037/xge0000675)
51. Chuang Y, Schechter L. 2015 Stability of experimental and survey measures of risk, time, and social preferences: a review and some new results. *J. Dev. Econ.* **117**, 151–170. (doi:10.1016/j.jdeveco.2015.07.008)
52. Meier S, Sprenger CD. 2015 Temporal stability of time preferences. *Rev. Econ. Stat.* **97**, 273–286. (doi:10.1162/REST_a_00433)
53. Harrison GW, Lau MI, Williams MB. 2002 Estimating individual discount rates in Denmark: a field experiment. *Am. Econ. Rev.* **92**, 1606–1617. (doi:10.1257/000282802762024674)
54. Ubfal D. 2016 How general are time preferences? Eliciting good-specific discount rates. *J. Dev. Econ.* **118**, 150–170. (doi:10.1016/j.jdeveco.2015.07.007)
55. Ellis BJ, Figueredo AJ, Brumbach BH, Schlomer GL. 2009 Fundamental dimensions of environmental risk: the impact of harsh versus unpredictable environments on the evolution and development of life history strategies. *Hum. Nat.* **20**, 204–268. (doi:10.1007/s12110-009-9063-7)
56. Pepper GV, Nettle D. 2017 The behavioural constellation of deprivation: causes and consequences. *Behav. Brain Sci.* **40**, e314. (doi:10.1017/S0140525X1600234X)
57. Nettle D, Frankenhuis WE. 2020 Life-history theory in psychology and evolutionary biology: one research programme or two? *Phil. Trans. R. Soc. B* **375**, 20190490. (doi:10.1098/rstb.2019.0490)
58. Stevens JR, Stephens DW. 2010 The adaptive nature of impulsivity. In *Impulsivity: the behavioral and neurological science of discounting* (eds GJ Madden, WK Bickel), pp. 361–387. Washington, DC: American Psychological Association.
59. Fawcett TW, McNamara JM, Houston AI. 2012 When is it adaptive to be patient? A general framework for evaluating delayed rewards. *Behav. Processes* **89**, 128–136. (doi:10.1016/j.beproc.2011.08.015)
60. Chu CYC, Chien HK, Lee RD. 2010 The evolutionary theory of time preferences and intergenerational transfers. *J. Econ. Behav. Organ.* **76**, 451–464. (doi:10.1016/j.jebo.2010.09.011)
61. Mell H, Baumard N, André JB. 2021 Time is money. Waiting costs explain why selection favors steeper time discounting in deprived environments. *Evol. Hum. Behav.* **42**, 379–387. (doi:10.1016/j.evolhumbehav.2021.02.003)
62. Barclay P, Barker JL. 2020 Greener than thou: people who protect the environment are more cooperative, compete to be environmental, and benefit from reputation. *J. Environ. Psychol.* **72**, 101441. (doi:10.1016/j.jenvp.2020.101441)
63. Lotem A, Fishman MA, Stone L. 1999 Evolution of cooperation between individuals. *Nature* **400**, 226–227. (doi:10.1038/22247)
64. Sherratt TN. 2001 The importance of phenotypic defectors in stabilizing reciprocal altruism. *Behav. Ecol.* **12**, 313–317. (doi:10.1093/beheco/12.3.313)
65. Fishman MA, Lotem A, Stone L. 2001 Heterogeneity stabilizes reciprocal altruism interactions. *J. Theor. Biol.* **209**, 87–95. (doi:10.1006/jtbi.2000.2248)
66. Ferriere R, Bronstein JL, Rinaldi S, Law R, Gauduchon M. 2002 Cheating and the evolutionary stability of mutualisms. *Proc. R. Soc. Lond. B* **269**, 773–780. (doi:10.1098/rspb.2001.1900)
67. McNamara JM, Leimar O. 2010 Variation and the response to variation as a basis for successful cooperation. *Phil. Trans. R. Soc. B* **365**, 2627–2633. (doi:10.1098/rstb.2010.0159)
68. Harris AC, Madden GJ. 2002 Delay discounting and performance on the Prisoner's Dilemma game. *Psychol. Rec.* **52**, 429–440. (doi:10.1007/BF03395196)
69. Curry OS, Price ME, Price JG. 2008 Patience is a virtue: cooperative people have lower discount rates. *Personal. Individ. Differ.* **44**, 780–785. (doi:10.1016/j.paid.2007.09.023)
70. Fehr E, Leibbrandt A. 2011 A field study on cooperativeness and impatience in the Tragedy of the Commons. *J. Public Econ.* **95**, 1144–1155. (doi:10.1016/j.jpubeco.2011.05.013)
71. Kocher MG, Martinsson P, Myrseth KOR, Wollbrant CE, 2013 *Strong, bold, and kind: self-control and cooperation in social dilemmas*. Rochester, NY: Social Science Research Network.
72. Sjåstad H. 2019 Short-sighted greed? Focusing on the future promotes reputation-based generosity. *Judgm. Decis. Mak.* **14**, 15. (doi:10.38050/2078-3809-2019-11-4-46-73)
73. Adams J, White M. 2009 Time perspective in socioeconomic inequalities in smoking and body mass index. *Health Psychol.* **28**, 83–90. (doi:10.1037/0278-6133.28.1.83)
74. Reimers S, Maylor EA, Stewart N, Chater N. 2009 Associations between a one-shot delay discounting measure and age, income, education and real-world impulsive behavior. *Personal. Individ. Differ.* **47**, 973–978. (doi:10.1016/j.paid.2009.07.026)
75. Griskevicius V, Tybur JM, Delton AW, Robertson TE. 2011 The influence of mortality and socioeconomic status on risk and delayed rewards: a life history theory approach. *J. Pers. Soc. Psychol.* **100**, 1015–1026. (doi:10.1037/a0022403)
76. Bulley A, Pepper GV. 2017 Cross-country relationships between life expectancy, intertemporal choice and age at first birth. *Evol. Hum. Behav.* **38**, 652–658. (doi:10.1016/j.evolhumbehav.2017.05.002)

77. Rand DG, Greene JD, Nowak MA. 2012 Spontaneous giving and calculated greed. *Nature* **489**, 427–430. (doi:10.1038/nature11467)
78. Jordan JJ, Hoffman M, Bloom P, Rand DG. 2016 Third-party punishment as a costly signal of trustworthiness. *Nature* **530**, 473–476. (doi:10.1038/nature16981)
79. Nowak MA, Sigmund K. 2005 Evolution of indirect reciprocity. *Nature* **437**, 1291–1298. (doi:10.1038/nature04131)
80. Giardini F, Vilone D. 2016 Evolution of gossip-based indirect reciprocity on a bipartite network. *Sci. Rep.* **6**, 37931. (doi:10.1038/srep37931)
81. Maynard Smith J, Price GR. 1973 The logic of animal conflict. *Nature* **246**, 15–18. (doi:10.1038/246015a0)
82. Biernaskie JM, Perry JC, Grafen A. 2018 A general model of biological signals, from cues to handicaps. *Evol. Lett.* **2**, 201–209. (doi:10.1002/evl3.57)
83. de Courson B, Nettle D. 2021 Why do inequality and deprivation produce high crime and low trust? *Sci. Rep.* **11**, 1937. (doi:10.1038/s41598-020-80897-8)
84. Galizzi MM, Navarro-Martinez D. 2019 On the external validity of social preference games: a systematic lab-field study. *Manage. Sci.* **65**, 976–1002. (doi:10.1287/mnsc.2017.2908)
85. McAuliffe WHB, Forster DE, Pedersen EJ, McCullough ME. 2019 Does cooperation in the laboratory reflect the operation of a broad trait? *Eur. J. Personal.* **33**, 89–103. (doi:10.1002/per.2180)
86. Righetti F, Finkenauer C. 2011 If you are able to control yourself, I will trust you: the role of perceived self-control in interpersonal trust. *J. Pers. Soc. Psychol.* **100**, 874–886. (doi:10.1037/a0021827)
87. Peetz J, Kammrath L. 2013 Folk understandings of self regulation in relationships: recognizing the importance of self-regulatory ability for others, but not the self. *J. Exp. Soc. Psychol.* **49**, 712–718. (doi:10.1016/j.jesp.2013.02.007)
88. Aguilar-Pardo D, Martínez-Arias R, Colmenares F. 2013 The role of inhibition in young children's altruistic behaviour. *Cogn. Process* **14**, 301–307. (doi:10.1007/s10339-013-0552-6)
89. Burks SV, Carpenter JP, Goette L, Rustichini A. 2009 Cognitive skills affect economic preferences, strategic behavior, and job attachment. *Proc. Natl Acad. Sci. USA* **106**, 7745–7750. (doi:10.1073/pnas.0812360106)
90. Restubog SLD, Garcia PRJM, Wang L, Cheng D. 2010 It's all about control: the role of self-control in buffering the effects of negative reciprocity beliefs and trait anger on workplace deviance. *J. Res. Personal.* **44**, 655–660. (doi:10.1016/j.jrp.2010.06.007)
91. Cohen TR, Panter AT, Turan N, Morse L, Kim Y. 2014 Moral character in the workplace. *J. Pers. Soc. Psychol.* **107**, 943–963. (doi:10.1037/a0037245)
92. Martinsson P, Myrseth KOR, Wollbrant C. 2014 Social dilemmas: when self-control benefits cooperation. *J. Econ. Psychol.* **45**, 213–236. (doi:10.1016/j.joep.2014.09.004)
93. Myrseth KOR, Riener G, Wollbrant CE. 2015 Tangible temptation in the social dilemma: cash, cooperation, and self-control. *J. Neurosci. Psychol. Econ.* **8**, 61–77. (doi:10.1037/npe0000035)
94. Fitouchi L, André JB, Baumard N. 2021 Moral disciplining: the cognitive and evolutionary foundations of puritanical morality. See https://doi.org/10.31234/osf.io/2stcv.
95. Williams KEG, Sng O, Neuberg SL. 2016 Ecology-driven stereotypes override race stereotypes. *Proc. Natl Acad. Sci. USA* **113**, 310–315. (doi:10.1073/pnas.1519401113)
96. Moon JW, Krems JA, Cohen AB. 2018 Religious people are trusted because they are viewed as slow life-history strategists. *Psychol. Sci.* **29**, 947–960. (doi:10.1177/0956797617753606)
97. Hoffman M, Hilbe C, Nowak MA. 2018 The signal-burying game can explain why we obscure positive traits and good deeds. *Nat. Hum. Behav.* **2**, 397–404. (doi:10.1038/s41562-018-0354-z)
98. Quillien T. 2020 Evolution of conditional and unconditional commitment. *J. Theor. Biol.* **492**, 110204. (doi:10.1016/j.jtbi.2020.110204)
99. Benabou R, Tirole J. 2003 Intrinsic and extrinsic motivation. *Rev. Econ. Stud.* **70**, 489–520. (doi:10.1111/1467-937X.00253)
100. Raihani NJ, Bshary R. 2015 Why humans might help strangers. *Front. Behav. Neurosci.* **9**, 39. (doi:10.3389/fnbeh.2015.00039)
101. Delton AW, Krasnow MM, Cosmides L, Tooby J. 2011 Evolution of direct reciprocity under uncertainty can explain human generosity in one-shot encounters. *Proc. Natl Acad. Sci. USA* **95**, 13 335–13 340. (doi:10.1073/pnas.1102131108)
102. Hoffman M, Yoeli E, Nowak MA. 2018 Cooperate without looking: why we care what people think and not just what they do. *Proc. Natl Acad. Sci. USA* **112**, 1727–1732. (doi:10.1073/pnas.1417904112)
103. Barclay P. 2020 Reciprocity creates a stake in one's partner, or why you should cooperate even when anonymous. *Proc. R. Soc. B* **287**, 20200819. (doi:10.1098/rspb.2020.0819)
104. Barclay P, Bliege Bird R, Roberts G, Számadó S. 2021 Cooperating to show that you care: costly helping as an honest signal of fitness interdependence. *Phil. Trans. R. Soc. B* **376**, 20200292. (doi:10.1098/rstb.2020.0292)
105. Gintis H, Smith EA, Bowles S. 2001 Costly signaling and cooperation. *J. Theor. Biol.* **213**, 103–119. (doi:10.1006/jtbi.2001.2406)
106. Lie-Panis J, André JB. 2022 Cooperation as a signal of time preferences. Figshare. (https://doi.org/10.6084/m9.figshare.c.5923872)

Proceedings of the Royal Society B. 10.1098/rspb.2021.2266

Supplementary Information for

Cooperation as a signal of time preferences

Julien Lie-Panis and Jean-Baptiste André

Corresponding author: Julien Lie-Panis; Email: `jliep@protonmail.com`

Contents

A The Model . . . 1
A.1 Trust game . . . 1
A.2 Reputation formation . . . 2
A.3 Cooperating with strangers . . . 2
B Discussion of assumptions . . . 2
B.1 Individual discount rate . . . 2
B.2 Initial state . . . 2
B.3 Observation . . . 3
B.4 Binary reputation . . . 3
B.5 Private reputation . . . 3
B.6 Possible equilibria . . . 3
C Evolutionarily stable sets . . . 3
C.1 Methods . . . 4
C.2 Signaler optimal policy set when Choosers discriminate according to reputation . . . 4
C.3 Chooser use of information . . . 5
C.4 Cooperative equilibrium . . . 5
C.5 Other equilibria . . . 6

A. The Model.

A.1. Trust game. We consider a pairwise trust game with two roles, that of Chooser and that of Signaler. The game consists in an asymmetric prisoner's dilemma with two stages: in the first, the Chooser either rejects (R) the Signaler, in which case the interaction ends with both players earning null payoff; or she accepts (A) partnership with her. An accepted Signaler reaps reward r ($r > 0$). In the second stage, that Signaler can either defect (D), keeping r for herself; or prove worthy of the Chooser's trust, by paying cost c to cooperate (C). We assume: $0 < c < r$. When the Signaler defects, the Chooser is harmed, losing h ($h > 0$); and when the Signaler cooperates, she benefits, gaining b, the benefits of cooperation ($b > 0$).

Payoffs for the trust game are summarized below, in Table S1. A specific case to keep in mind is when $h = c$ and $b = r - c$. In such a case, payoffs are symmetric: everything is as if Choosers who play A pay cost c for the Signaler to gain b, thus losing c when the Signaler plays D; and gaining in turn b when the Signaler pays c to play C.

Table S1. Payoffs for the trust game.

		Signaler	
		Cooperate (C)	Defect (D)
Chooser	Accept (A)	$(b, r-c)$	$(-h, r)$
	Reject (R)	$(0, 0)$	$(0, 0)$

Individuals in an infinite population engage in cooperative encounters with strangers throughout their life, as represented by the trust game. Every T_S, an individual is paired up with two individuals she has never encountered before, with whom she plays the trust game; once in each role. The population is progressively renewed: individuals die after a given trust game with probability $\frac{T_S}{T_B}$, at which point they reproduce according to their accumulate payoffs. Individual expected life span is therefore equal to $T_S \times \sum_{i=0}^{\infty}(1 - \frac{T_S}{T_B})^i = T_B$. We assume social time step to be negligible in front of biological time step: $T_S \ll T_B$.

Individuals vary in a hidden quality: their temporal discount rate δ. Individuals are born with a certain δ, which is randomly selected in $]0, \infty[$, depending on a continuous probability distribution which characterizes the population. The probability that δ takes any single positive value is null. An individual's temporal discount rate remains constant throughout her life. At any point in time t, the total payoff an individual can expect to derive from cooperation is equal to her current payoffs in the trust games, plus $\frac{1}{1+\delta}$ times the payoffs she can expect at time $t + T_S$, plus $(\frac{1}{1+\delta})^2$ time the payoffs she can expect at time $t + 2T_S$, etc. (geometric discounting). Since $T_S \ll T_B$, this can be approximated using an infinite sum; individuals engage in a large number of cooperative interactions during their life span. To simplify future calculations, we measure time t in units of T_S from here on ($T_S = 1$).

A.2. Reputation formation. Signaler behavior is observed by the entire population with probability p, and error σ. We assume: $0 < p < 1$ and $0 < \sigma < \frac{1}{2}$. Information is private and binary: when a Signaler is observed, a fraction $1 - \sigma$ of the population makes the observation corresponding to her action (C or D), and a fraction σ wrongly observes the other action. We assume new information replaces old information and that individuals do not forget.

At any point in her life, a Signaler may be in one of three "reputational" *states*, depending on what action she was last observed undertaking (if any). Let $\mathcal{N}$ be the state Signalers are born in, and remain until they are observed for the first time; and let $\mathcal{C}$ ($\mathcal{D}$) be the state attained when a Signaler is last observed playing C (D).

Consider a Chooser-Signaler pair. When the Signaler is in state $\mathcal{N}$, the Chooser has no specific information on which to condition her play. We assume that in this case, there is an exogenous positive chance f that the Chooser accepts the Signaler (see section B.2).

The Chooser may otherwise face one of two informational *events* — indicating the Signaler has behaved in a trustworthy manner, by playing C after a previous partner accepted her (an event we note $\mathcal{T}$); or exploited the trust of that previous partner (event $\mathcal{E}$). Informational events faced by the Chooser correlate with the Signaler's state, but do not coincide given positive noise σ. When the latter is in state $\mathcal{C}$ ($\mathcal{D}$), the former has a $1 - \sigma$ (σ) chance of facing event $\mathcal{T}$, and otherwise faces event $\mathcal{E}$.

A.3. Cooperating with strangers. Mutual cooperation is net beneficial by assumption ($b > 0$ and $r > c$). Choosers stand to gain from accepting a Signaler who subsequently plays C, and to lose from partnering with a Signaler who subsequently plays D. Choosers may condition their play on specific information pertaining to the unfamiliar Signaler: event $\mathcal{T}$ or event $\mathcal{E}$.

When Chooser strategies differentiate between events $\mathcal{T}$ and $\mathcal{E}$, a Signaler's behavior in a given trust game may alter her future payoffs, by leading to a change in state (see section C.2). Signalers may condition their strategy on their temporal discount rate δ, as well as their state.

A fully specified strategy profile in this game therefore involves specifying, throughout an individual's life, whether to play A or R in the Chooser role under events $\mathcal{E}$ and $\mathcal{T}$; and whether to play C or D in the Signaler role, given own personal temporal discount rate δ and current reputational state. Note that we don't allow individuals to condition their play on arbitrary elements which are exogenous to the model (i.e. to change their strategy according to time t — see section B.6). For simplicity, we do not consider mixed strategies either (in which individuals behave probabilistically).

B. Discussion of assumptions.

B.1. Individual discount rate. Variation of time preferences inside a population can originate from two independent sources; evolutionary models show that it is adaptive to be more present-oriented (higher discount rate) when future rewards are more uncertain (1), and when present needs are more pressing (2). An individual's time preferences are thus susceptible to depend on a variety of factors, including mortality and accumulated biological and material capital: when individuals face higher probability of dying, future rewards are more uncertain; and when they have more capital, their present needs are less pressing. Since both mortality and capital tend to increase with age, age should therefore affect time preferences in a complex manner.

In our model, individuals are characterized by differing discount rates, which are exogenous and remain constant throughout their lifetime. We take between-individual differences as a given, and assume we do not need to consider within-individual temporal variability (because older age does not straightforwardly imply higher discounting). Individuals do not have access to any of the factors underlying their discount rate δ; in particular, mortality occurs without memory, and they cannot accumulate capital.

B.2. Initial state. We assumed that Choosers cooperate with Signalers for whom there is no specific information (state $\mathcal{N}$) with non-null probability f. This is a technical assumption, which allows reputation to be established: since a Signaler has a $f \times p$ chance of exiting state $\mathcal{N}$ at every social interaction (they have to be accepted for cooperation and to be observed), they spend on average $T_S \times \frac{1}{fp}$ in initial state $\mathcal{N}$ (an amount of time which is negligible with respect to their life span); after which they may alternate between states $\mathcal{C}$ and $\mathcal{D}$ depending on their strategy, and their partners face event $\mathcal{T}$ or $\mathcal{E}$.

$f > 0$ can be seen as a way to capture the existence of a cooperative past between certain players, outside of the interaction under study. While Choosers cannot gain any privileged information about an individual in the model (they always face new Signalers), they will have played other cooperative games with certain people beforehand and developed specific relations with trusted partners. In the most general sense, each individual Signaler i should face f_i depending on her past outside of the trust game — in any case, so long as $f_i > 0$, reputation for the trust game will be established, and the calculations conducted in section C.2 hold.

B.3. Observation. We rarely observe the cooperative or uncooperative behavior of strangers directly. We may however hear from third-party observers, or from individuals they talked to, etc. Our model can be seen to reflect rapid social transmission of information (gossip) (3). We would obtain the same results were we to assume that Signalers are observed by one or several witnesses with probability p, and that these witnesses are motivated to gossip about Signaler behavior to several acquaintances, who are also motivated to gossip to their acquaintances, etc. — leading all individuals to obtain new information by the next instantiation of the Trust game. σ can be seen to capture the noisiness of the entire social transmission process (even though the population is large, we still assume that $\sigma < \frac{1}{2}$).

Chooser decision in the absence of information arising from the trust game (which corresponds to a partner in state $\mathcal{N}$) is thus kept outside of the model. We only consider strategy given such information — given event $\mathcal{T}$ or given $\mathcal{E}$. Our simplified model allows us to focus on information reliability: section C.3 establishes the conditions under which reputation is reliable enough for cooperation to be established. In contrast, we are not concerned with information availability or consensus formation, which have been studied elsewhere (4–6).

B.4. Binary reputation. In the same spirit, we model trust and cooperation following an asymmetric game, where only Signalers may possess a reputation (and Choosers are only concerned with partner choice). This runs in contrast with most models of cooperation, which involve the framework of indirect reciprocity and a symmetric prisoner's dilemma (e.g. (4, 7, 8)).

As a result, reputation dynamics are particularly simple (for an exhaustive study in the symmetric case, see Ohtsuki and Iwasa (8)). There are only 2 possible states (vs. four in the symmetric case), and only $2^2 = 4$ possible Chooser strategies — which correspond to 4 ways of individually assigning reputation (Choosers need not play the same strategy in principle). In section C.3, we examine the conditions under which "discriminating according to reputation" (to last observation), i.e. playing A given $\mathcal{T}$ and R given $\mathcal{E}$ is advantageous to Choosers.

We can already note this is the only way of collectively assigning reputation which is conducive to cooperation. When Choosers all play one of the 3 other strategies, cooperation is reputation neutral or detrimental: a Signaler who is observed playing C is either as likely (when Choosers always accept or always reject) or less likely (when Choosers play A given $\mathcal{E}$ and R given $\mathcal{T}$) of being rewarded in the future, than a Signaler who is observed playing D. Since cooperation is costly, Signalers all benefit from playing D. We come back to this in Proposition P3, at the end of this document.

B.5. Private reputation. Reputation is also private in our model: every time a Signaler is observed (e.g. playing C), σ percent of Choosers end up with the conflicting piece of information (e.g. $\mathcal{E}$). We could have considered public reputation, whereby the entire population receives the same piece of information, which conflicts with actual Signaler behavior with probability σ.

Such a collective view of reputation would have led to certain technical simplifications: when reputation is public, Signaler state and Chooser information coincide perfectly. In particular, at a cooperative equilibrium where Choosers play play A given $\mathcal{T}$ and R given $\mathcal{E}$, Signalers who are assigned an exploitative reputation are simply never chosen again — which ends up being the case for all individuals under positive noise and infinite social interactions.

The main results should however remain the same. Since we can ignore state $\mathcal{D}$, calculations conducted in section C.2 are greatly simplified: a Signaler in good standing, can either play C and face probability $1 - p\sigma$ of being accepted again, or play D and face smaller probability $1 - p(1 - \sigma)$; as she will in future rounds (as long as she is not assigned exploitative reputation).

Hence she should face a trade-off between:

$$\Pi(C) = \sum_{t=0}^{\infty} (1 - p\sigma)^t \frac{r - c}{(1 + \delta)^t} = \frac{1 + \delta}{\delta + p\sigma}(r - c)$$

And:

$$\Pi(D) = \sum_{t=0}^{\infty} (1 - p(1 - \sigma))^t \frac{r}{(1 + \delta)^t} = \frac{1 + \delta}{\delta + p(1 - \sigma)} r$$

We find $\Pi(C) > \Pi(D) \iff \delta < \hat{\delta} = p[(1 - \sigma)(\frac{r}{c} - 1) - \sigma\frac{r}{c}]$, the same result we obtain below when reputation is private (see section C.2 for a complete demonstration in that case).

B.6. Possible equilibria. Our assumptions limit possible Chooser strategies, and therefore limit the set of possible equilibria. Repeated games are generally characterized by a "Folk theorem" (9), whereby numerous Nash equilibria are possible. Since Choosers can only hold one bit of information at a time, they can only engage in one of four simple strategies (they can't engage in a complex "Grim-trigger" strategy, whereby they pass on all Signalers who don't engage in a specific sequence of actions).

We also prevent players from conditioning their play on arbitrary exogenous elements. In theory, Chooser behavior could vary with time: they could cooperate given trustworthy reputation only when the weather is sunny, and reject given exploitative reputation on all days but those marked by the death of a famous pop star. Rejecting such arbitrary scenarios strongly limits the set of feasible outcomes, but does not affect our determination of the main cooperative equilibrium (the calculations conducted in section C.2 remain valid when we allow Signalers to condition their play on time, and the calculations conducted in section C.3 remain valid when we allow Choosers to condition their play on time).

C. Evolutionarily stable sets.

C.1. Methods. In this section, we investigate possible stable endpoints of evolution, by identifying evolutionary stable sets (ES sets) of strategies (10). We reason in terms of strategy sets in order to ignore the effect of meaningless deviations, which do not affect players' expected payoffs at equilibrium. Because calculations are heavy, we start by determining Signaler optimal policy when Choosers discriminate according to reputation (as they do at the main cooperative equilibrium). We then outline the conditions under which Choosers stand to benefit from this discrimination.

We end the section by identifying two strategy sets of interest, and the conditions under which they define a (strict) Nash equilibrium set. Since sets of strict Nash equilibria are ES sets, and since an evolutionary stable strategy (ESS) must be Nash, we are able to deduce the conditions under which cooperation may be favoured by an evolutionary process.

C.2. Signaler optimal policy set when Choosers discriminate according to reputation. Let us consider a Signaler of discount rate δ. Since $T_B \gg T_S$, her payoffs from any point in time t can be approximated using the infinite sum $\sum_{t'=t}^{\infty}(\frac{1}{1+\delta})^{\frac{t'-t}{T_S}}\pi(x'_t,a'_t)$ — where $\pi(x'_t,a'_t)$ is her expected payoff for the Trust game conducted at time t', in future state x'_t, when choosing action $a'_t = a(x'_t)$ (as per her strategy).

Let us assume that Choosers discriminate according to reputation: when in the role of Chooser, all individuals in the population play A given $\mathcal{T}$ and R given $\mathcal{E}$. The probability that a partner will place her trust in our Signaler is a function of state: $F_N = f$ percent of Choosers cooperate with a Signaler in state $\mathcal{N}$ by hypothesis; and that fraction jumps to $F_C = 1-\sigma$ for a Signaler in state $\mathcal{C}$, and to $F_D = \sigma$ for a Signaler in state $\mathcal{D}$.

In a given state $\mathcal{X}$, the Signaler can expect payoff $\pi(\mathcal{X},C) = F_X \times (r-c)$ when she plays C, and $\pi(\mathcal{X},D) = F_X \times r$ when she plays D. When she is not observed, her state remains the same in the next Trust game. When she is observed, with probability $F_X \times p$ (her partner first has to play A for observation to be possible), her state changes to $\mathcal{C}$ when she plays C and $\mathcal{D}$ when she plays D.

A Signaler's future state can therefore be described as a function of her current state and action, without reference to time t. Her optimal policy can be obtained following Bellmann's principle (11), by defining the value function V:

$$V(\mathcal{X}) = \max_{a\in\{C,D\}}\{\pi(\mathcal{X},a) + \frac{1}{1+\delta}V(\mathcal{X}')\}$$

$$V(\mathcal{X}) = \max\{F_X(r-c) + \frac{1}{1+\delta}[F_X pV(\mathcal{C}) + (1-F_X p)V(\mathcal{X})],\quad F_X r + \frac{1}{1+\delta}[F_X pV(\mathcal{D}) + (1-F_X p)V(\mathcal{X})]\}$$

$$V(\mathcal{X}) = F_X \times \max\{(r-c) + \frac{pV(\mathcal{C})}{1+\delta},\ r + \frac{pV(\mathcal{D})}{1+\delta}\} + \frac{(1-F_X p)V(\mathcal{X})}{1+\delta}$$

$$V(\mathcal{X}) = \frac{F_X(1+\delta)}{F_X p+\delta} \times \max\{(r-c) + \frac{pV(\mathcal{C})}{1+\delta},\ r + \frac{pV(\mathcal{D})}{1+\delta}\}$$

Since $F_X > 0$ for any state $\mathcal{X}$ in which the Signaler may find herself, her optimal policy in that state is determined by the comparison between two expressions which do not depend on $\mathcal{X}$. There are therefore two possibilities: either it pays more to play C now, in which case it will always pay more to play C (whatever the attained state) and $V(\mathcal{C})$ can be calculated assuming the Signaler always plays C and therefore remains in state $\mathcal{C}$:

$$V(\mathcal{C}) = \sum_{t'=t}^{\infty}(\frac{1}{1+\delta})^{\frac{t'-t}{T_S}}F_C(r-c) = \frac{1+\delta}{\delta}F_C(r-c)$$

Or it pays more to play D now, in which case the optimal policy is to always play D and:

$$V(\mathcal{D}) = \sum_{t'=t}^{\infty}(\frac{1}{1+\delta})^{\frac{t'-t}{T_S}}F_D r = \frac{1+\delta}{\delta}F_D r$$

Our Signaler's optimal policy is thus determined by the comparison:

$$(r-c) + \frac{pF_C(r-c)}{\delta} > r + \frac{pF_D r}{\delta}$$

$$\delta < \hat{\delta} = p[F_C(\frac{r}{c}-1) - F_D\frac{r}{c}] = \frac{p\times\hat{\beta}}{c} \qquad [1]$$

Where $\hat{\beta} = F_C(r-c) - F_D r$ is the benefit of consistently playing C instead of D. Optimal policy is to always cooperate if $\delta < \hat{\delta}$, and to always defect if $\delta > \hat{\delta}$. Note that restricting Signalers to stationary strategies is unnecessary here: when Choosers always discriminate according to reputation, Signaler future state is only a function of current state and action, and Bellman's principle can be applied.

Note also that the above formulation defines an optimal set of Signaler strategies. Signalers whose discount rate is precisely equal to $\hat{\delta}$ are indifferent between playing C and D following the above equation. (Since discount rates are continuously distributed in the population, this happens with null probability.) In addition, Signalers whose discount rate is smaller (larger) than $\hat{\delta}$ and who always cooperate (defect) never reach reputational state $\mathcal{D}$ ($\mathcal{C}$) — and are therefore indifferent between playing C and D given that unattained state.

A more precise definition of this optimal set is therefore: (i) if $\delta < \hat{\delta}$, play C in states $\mathcal{N}$ and $\mathcal{C}$, and C or D in unattained state $\mathcal{D}$; (ii) if $\delta = \hat{\delta}$, play any strategy; (iii) if $\delta > \hat{\delta}$, play D in states $\mathcal{N}$ and $\mathcal{D}$, and C or D in unattained state $\mathcal{C}$. We refer to this set as "throughout one's life, cooperate if $\delta < \hat{\delta}$ and defect if $\delta > \hat{\delta}$" from here on. When Choosers discriminate according to reputation, any two strategies strategies in this set yield identical payoffs on average, and any strategy not in the set can be expected to yield a strictly inferior payoff.

C.3. Chooser use of information. Let us consider a Chooser at a certain point in time t, who possesses specific information $\omega \in \{\mathcal{T}, \mathcal{E}\}$ on her prospective Signaler partner. (Since all Signalers are born in state $\mathcal{N}$, Choosers never possess specific information at time $t = 0$, when the population's first trust game occurs.) If she accepts, she can expect payoff $P_t(C|\omega) \times b + P_t(D|\omega) \times (-h) = P_t(C|\omega)(b + h) - h$. In contrast, passing given ω yields certain null payoff. Accepting the Signaler given informational event ω is beneficial on average if and only if the above expression is positive, which is equivalent to:

$$P_t(C|\omega) > \frac{h}{b+h}$$

Accepting a Signaler given ω is beneficial on average if and only if that event is a sufficiently good predictor of the Signaler's cooperation at time t. By assumption, Choosers cannot take time into account. When Signaler strategy only depends on individual discount rate δ (as it does following the above optimal policy), that assumption proves to be unnecessary: in such a case, the fraction $P(C)$ of trustworthy Signalers is constant, as are predictive values $P_t(C|\mathcal{T})$ and $P_t(C|\mathcal{E})$ (see under).

Let us therefore assume that Signalers play a stationary strategy, and consider the constant fraction $\tau = P(C)$. Following Bayes' rule, $P_t(C|\mathcal{T})$ is equal to $\frac{P_t(\mathcal{T}|C)}{P_t(\mathcal{T})} \times \tau$. Both events $\mathcal{T}|C$ and $\mathcal{T}$ require that the Signaler has exited state $\mathcal{N}$ (i.e. has been observed at least once), a possibility whose probability does not depend on Signaler strategy, and which is positive, and simplifies in the above expression*. We deduce that our Chooser stands to gain from accepting given $\mathcal{T}$ if and only if:

$$\frac{1-\sigma}{\tau(1-\sigma)+(1-\tau)\sigma} \times \tau > \frac{h}{b+h}$$

$$\tau > \underline{\tau} = \frac{\sigma h}{\sigma h + (1-\sigma)b} \quad [2a]$$

Note that, since $\sigma > 0$, the denominator of the above expression is positive whatever the fraction τ of cooperative Signalers (when $t > 0$, $\mathcal{T}$ is a non-trivial event). The obtained lower bound $\underline{\tau}$ is also positive: accepting given trustworthy reputation can only be (strictly) worthwhile if some individuals actually cooperate. In addition it is an increasing function of σ: the larger the error, the larger the minimum fraction of trustworthy Signalers. We can simplify the above expression when errors tend towards 0 while remaining positive:

$$\tau > \underline{\tau} = \frac{h}{b}\sigma + o(\sigma), \qquad \sigma \to 0^+ \quad [2a']$$

We proceed similarly for event $\mathcal{E}$. Following Bayes' rule, $P_t(C|\mathcal{E})$ is equal to $\frac{P_t(\mathcal{E}|C)}{P_t(\mathcal{E})} \times \tau$, which yields an expression which does not depend on time t, and whose denominator is positive. Choosers stand to lose from accepting given $\mathcal{E}$ if and only if:

$$\frac{\sigma}{\tau\sigma+(1-\tau)(1-\sigma)} \times \tau < \frac{h}{b+h}$$

$$\tau < \overline{\tau} = 1 - \frac{\sigma b}{\sigma b + (1-\sigma)h} \quad [2b]$$

The obtained upper bound is smaller than 1: rejecting given exploitative reputation can only be (strictly) worthwhile if some individuals actually defect. $\overline{\tau}$ is a decreasing function of σ: the larger the error, the larger the minimum fraction of exploitative Signalers. We can simplify the above expression when errors tend towards 0 while remaining positive:

$$\tau < \overline{\tau} = 1 - \frac{b}{h}\sigma + o(\sigma),\ \sigma \to 0^+ \quad [2b']$$

When Signalers play a stationary strategy such as the above optimal policy, discriminating according to reputation is strictly beneficial if and only if the fraction of trustworthy signalers τ verifies: $0 < \underline{\tau} < P(C) < \overline{\tau} < 1$. (A rapid calculation shows that $\underline{\tau}$ is always smaller than $\overline{\tau}$ when $\sigma < \frac{1}{2}$).

C.4. Cooperative equilibrium.

Proposition 1 (P1) *The strategy set in which, throughout their life, (i) Choosers accept given $\mathcal{T}$ and reject given $\mathcal{E}$, and (ii) Signalers of discount rate δ cooperate if $\delta < \overset{*}{\delta} = p[(1-\sigma)(\frac{r}{c} - 1) - \sigma\frac{r}{c}]$ and defect if $\delta > \overset{*}{\delta}$* (**Conditional Trust and Trustworthiness**)

a) is a set of strict Nash equilibria iff $\frac{\sigma h}{\sigma h+(1-\sigma)b} < P(\delta < \overset{*}{\delta}) < 1 - \frac{\sigma b}{\sigma b+(1-\sigma)h}$

* Any given Signaler of age t_S has a probability $1 - (1-fp)^{t_S}$ of having exited state $\mathcal{N}$. As long as $t > 0$, Choosers cannot be certain they will encounter a newborn Signaler, since the population is progressively renewed from time $t = 1$. The probability they face $\omega \in \{\mathcal{T}, \mathcal{E}\}$ is therefore positive, and may depend on time t (particularly for the first generation of Choosers).

b) is Nash iff $\frac{\sigma h}{\sigma h+(1-\sigma)b} \leq P(\delta < \overset{*}{\delta}) \leq 1 - \frac{\sigma b}{\sigma b+(1-\sigma)h}$

The value of trustworthy reputation is then: $\overset{*}{\rho} = p \times \overset{*}{\beta} = p \times [(1-\sigma)(r-c) - \sigma r]$.

Proof of P1-a): Let us assume individuals all play according to the above strategy set. We prove that all deviations available to Signalers and Choosers are detrimental if and only if $\frac{\sigma h}{\sigma h+(1-\sigma)b} < P(\delta < \overset{*}{\delta}) < 1 - \frac{\sigma b}{\sigma b+(1-\sigma)h}$.

i. The proportion of trustworthy Signalers (who play strategy C at a given point in time) is stable and equal to $\tau = P(\delta < \overset{*}{\delta})$. Choosers are therefore in the situation described in section C.3: deviation to playing R given $\mathcal{T}$ is detrimental iff $P(\delta < \overset{*}{\delta}) > \underline{\tau} = \frac{\sigma h}{\sigma h+(1-\sigma)b}$, and deviation to playing A given $\mathcal{E}$ is detrimental iff $P(\delta < \overset{*}{\delta}) < \overline{\tau} = 1 - \frac{\sigma b}{\sigma b+(1-\sigma)h}$.

ii. Choosers discriminate according to reputation: Signalers are in the situation described in section C.2. Following the previous calculations, Signalers are playing their optimal policy: deviation to any strategy outside the set is detrimental.

iii. Note that deviations inside the optimal policy set are meaningless. Whether a Signaler plays C or D given an unattained state does not affect her payoffs, or that of other players (her Chooser partners). Since $P(\delta = \overset{*}{\delta}) = 0$, the possibility of a Signaler being born with quality precisely equal to $\overset{*}{\delta}$ can be neglected; and Signaler behavior given that improbable eventuality does not affect other players' payoffs either.

iv. Signalers cooperate when $\delta < \overset{*}{\delta} \iff \frac{1}{\delta} \times (p \times \overset{*}{\beta}) > c$. Everything is as if trustworthy Signalers are those who can afford to pay c, the costs of cooperation, in order to gain $\overset{*}{\beta}$ their entire future life, with probability p. Indeed, since $\sum_{t'=t}^{\infty} (\frac{1}{1+\delta})^{\frac{t'-t}{T_S}} = \frac{1}{\delta}$, an individual's social future may be represented by a single trust game whose payoffs are discounted with rate $\frac{1}{\delta}$. Since optimal Signaler policy does not depend on (attained) state, the value of establishing and maintaining a trustworthy reputation appear equal, and can be captured by $\overset{*}{\rho}$.

Proof of P1-b): following the calculations conducted in section C.3, Choosers stand to gain from deviation to playing $R|\mathcal{T}$ iff $P(\delta < \overset{*}{\delta}) < \frac{\sigma h}{\sigma h+(1-\sigma)b}$, and to playing $A|\mathcal{E}$ iff $P(\delta < \overset{*}{\delta}) > 1 - \frac{\sigma b}{\sigma b+(1-\sigma)h}$. There are no profitable deviations available to Signalers.

In our model, cooperation is therefore stabilized by variation of individual time preferences. Following proposition P1, cooperation is strict Nash and therefore ESS if the fraction of individuals whose discount rate is inferior to $\overset{*}{\delta}$ exceeds $\underline{\tau} > 0$ and the fraction of individuals whose discount rate is superior to $\overset{*}{\delta}$ exceeds $1 - \overline{\tau} > 0$. When in contrast $P(\overset{*}{\delta}) < \underline{\tau}$ or $P(\overset{*}{\delta}) > \overline{\tau}$, the above strategy set is not Nash, and therefore not ESS. (In either equality case, it is also not ESS, since rare mutants playing $R|\mathcal{T}$ or $A|\mathcal{E}$ when in the Chooser role perform as well the resident against the resident, and against themselves).

To take an extreme illustrative example, were all individuals to possess the same discount rate δ_0, $P(\delta < \overset{*}{\delta})$ would have to be equal to 0 or 1 — meaning that Choosers would stand to gain from acceptance given exploitative reputation or rejection given trustworthy reputation. Cooperation is also impossible if $\overset{*}{\delta} \leq 0 \iff \overset{*}{\rho} \leq 0$, in which case $P(\delta < \overset{*}{\delta})$ has to be null. A necessary condition for the above equilibrium is therefore:

$$\overset{*}{\rho} > 0 \iff \sigma < \frac{\frac{r}{c} - 1}{2\frac{r}{c} - 1} < \frac{1}{2} \qquad [2]$$

We deduce an upper bound on error σ, which is more restrictive than our initial assumption ($\sigma < \frac{1}{2}$). The above equation underscores the importance of Chooser discrimination according to reputation: were Choosers to treat Signalers identically whatever their reputation (e.g. always accept), cooperation would yield no relative benefit to trustworthy Signalers in the future. Mathematically, if we assume $F_C = F_D$ in the calculations performed in section C.2, we obtain $\hat{\beta} = 0$, and therefore $\hat{\rho} = 0$. (Even better, if Choosers accept given $\mathcal{E}$ and reject given $\mathcal{T}$, we obtain $\hat{\beta} < 0$). The two below propositions show that our model admits only one other Nash equilibrium, in which individuals trivially do not engage in cooperation.

C.5. Other equilibria.

Proposition 2 (P2) *The strategy set in which, throughout their lives, (i) Choosers reject given $\mathcal{E}$ and $\mathcal{T}$ and (ii) Signalers defect* (**Pooling with Rejection**) *is always a set of strict Nash equilibria.*
The value of trustworthy reputation in such a situation is: $\overset{*}{\rho}_0 = 0$.

Proof of P2-a): Let us assume that individuals all play according to the above strategy profile. We prove that there are no profitable deviations for either role.

i. Since Signalers always exploit their partners, acceptance given either $\mathcal{T}$ or $\mathcal{E}$ is always detrimental to Choosers (both events occur since $\sigma > 0$).

ii. Since Choosers do not use past behavior, a Signaler's future payoffs are unaffected by her current actions, precluding any profitable deviation to a more cooperative strategy outside of the set. (The set of Signaler strategies defined by (ii) includes playing either strategy C or D given unattained state $\mathcal{C}$.)

iii. A trustworthy Signaler would pay c to gain no future benefits: $\overset{*}{\rho}_0 = 0$.

Pooling with Rejection is always an ES set in our model. The evolution of cooperation from non-cooperation raises a bootstrapping problem (12). A rare mutant of initial frequency $\mu \ll 1$ who plays a strategy in the Conditional Trust and Trustworthiness set loses: (i) h with probability $1-\mu$, every time she faces event $\mathcal{T}$ when in the Chooser role (which occurs with probability $\sigma + O(\mu)$), and (ii) is eventually accepted in the Signaler role (after an average of $\frac{1}{f}$ iterations of the trust game), at which point everything is as if she pays c to gain $\mu \times \overset{*}{\rho}$ her whole life if her discount rate is smaller than $\overset{*}{\delta}$.

Going beyond the confines of our model, one may find reasons to (moderately) put into doubt the evolutionary stability of this trivial non-cooperative equilibrium. To begin, when error σ is sufficiently small (so that losing h with probability σ one's whole life is not overly costly), the above equilibrium may be invaded (by mutants playing Conditional Trust and Trustworthiness) due to stochastic effects (13).

Another possibility is to add an additional, unrealized informational event $\mathcal{G}$ (or, equivalently to assume that σ is in fact null in this situation). When a trust game is never played, there may be no reason to consider positive reputation for that game. We can imagine the following scenario (which, once again, cannot occur in our model as it stands): (i) first, Choosers may deviate to playing A given null event $\mathcal{G}$ (or null event $\mathcal{T}$ when $\sigma = 0$), without any impact on their payoff. This would not be a meaningless deviation: when a fraction μ of Choosers play $A|\mathcal{G}$, (ii) Signalers benefit from deviation to C given sufficiently small discount rate $\delta < \mu \times \overset{*}{\delta}$ (Mathematically, F_C and F_D are multiplied by μ in the demonstration of section C.2.) As long as $P(\delta < \overset{*}{\delta}) \geq \underline{\tau}$, (iii) this second advantageous Signaler deviation does not make the original Chooser deviation disadvantageous as long as the fraction of trustworthy Signalers exceeds $\underline{\tau}$. Hence, following this hypothetical scenario, Pooling with Rejection is subject to indirect invasion arising from neutral mutations when $P(\delta < \overset{*}{\delta} \geq \underline{\tau})$ (as per the definition of indirect invasion introduced by Jordan et al. (14)).

Proposition 3 (P3) *A Nash equilibrium of this game is either Conditional Trust and Trustworthiness or Pooling with Rejection.*

Proof of P3): Let us assume we are at a Nash equilibrium. We prove we are either at Conditional Trust and Trustworthiness or Pooling with Rejection.

i. Since Choosers are at equilibrium, Signalers' prospects depend solely on their state. We can introduce F_C (F_D), a Signaler's chance of facing a cooperative partner when in state $\mathcal{C}$ ($\mathcal{D}$), which remains constant. (F_C and F_D depend on the strategy or strategies played by the Chooser population.) Signaler optimal policy is thus obtained as in section C.2; optimal policy will be to always play C if one's discount rate δ is smaller than $\frac{\hat{\rho}}{c}$, and to always play D when δ is larger than $\frac{\hat{\rho}}{c}$, $\hat{\rho}$ being a function of F_C and F_D and the game parameters (one of these two conditions may be impossible).

ii. If $\hat{\rho} \leq 0$, optimal Signaler policy is to always play D, hence optimal Chooser strategy is to always reject prospective partners. We are thus in the Pooling with Rejection equilibrium.

iii. If $\hat{\rho} > 0$, optimal policy is to always play C for a fraction $f(C) = P(\delta < \frac{\hat{\rho}}{c})$ of Signalers, and to always defect for others. We show by contraposition that $\underline{\tau} \leq f(C) \leq \overline{\tau}$. Indeed, let us assume that $f(C) < \underline{\tau}$. In such a case, Choosers earn greater payoff when they reject given $\mathcal{T}$ than when they cooperate given that event: at a Nash equilibrium, Choosers would therefore play the former, and $\hat{\rho}$ would have to be negative, which is not the case (replace F_C with 0 in equation (1) to see this). An analogous reasoning can be made when that fraction exceeds $\overline{\tau}$: Choosers must therefore discriminate according to reputation, and we must be in the Conditional Trust and Trustworthiness equilibrium (with $\hat{\rho} = \overset{*}{\rho}$).

References

1. TW Fawcett, JM McNamara, AI Houston, When is it adaptive to be patient? A general framework for evaluating delayed rewards. *Behav. Process.* **89**, 128–136 (2012).
2. H Mell, N Baumard, JB André, Time is money. Waiting costs explain why selection favors steeper time discounting in deprived environments. *Evol. Hum. Behav.* **42**, 379–387 (2021).
3. F Giardini, D Vilone, Evolution of gossip-based indirect reciprocity on a bipartite network. *Sci. Reports* **6**, 37931 (2016) Number: 1 Publisher: Nature Publishing Group.
4. MA Nowak, K Sigmund, Evolution of indirect reciprocity by image scoring. *Nature* **393**, 573–577 (1998) Number: 6685 Publisher: Nature Publishing Group.
5. M Hoffman, E Yoeli, CD Navarrete, Game Theory and Morality in *The Evolution of Morality*, Evolutionary Psychology, eds. TK Shackelford, RD Hansen. (Springer International Publishing, Cham), pp. 289–316 (2016).
6. B Burum, MA Nowak, M Hoffman, An evolutionary explanation for ineffective altruism. *Nat. Hum. Behav.* **4**, 1245–1257 (2020).

7. O Leimar, P Hammerstein, Evolution of cooperation through indirect reciprocity. *Proc. Royal Soc. London. Ser. B: Biol. Sci.* **268**, 745–753 (2001) Publisher: Royal Society.
8. H Ohtsuki, Y Iwasa, The leading eight: Social norms that can maintain cooperation by indirect reciprocity. *J. Theor. Biol.* **239**, 435–444 (2006).
9. D Fudenberg, E Maskin, The Folk Theorem in Repeated Games with Discounting or with Incomplete Information. *Econometrica* **54**, 533–554 (1986) Publisher: [Wiley, Econometric Society].
10. B Thomas, On evolutionarily stable sets. *J. Math. Biol.* **22**, 105–115 (1985).
11. R Bellman, *Dynamic programming.* (Dover Publications, Mineola, N.Y), Dover ed edition, (2003).
12. JB André, Contingency in the Evolutionary Emergence of Reciprocal Cooperation. *The Am. Nat.* **185**, 303–316 (2015).
13. MA Nowak, A Sasaki, C Taylor, D Fudenberg, Emergence of cooperation and evolutionary stability in finite populations. *Nature* **428**, 646–650 (2004).
14. JJ Jordan, M Hoffman, P Bloom, DG Rand, Third-party punishment as a costly signal of trustworthiness. *Nature* **530**, 473–476 (2016).

Chapter 3

The repeated punishment game explains why, and when, we seek revenge

Objectives and summary

A prominent explanation for revenge is the deterrence of future transgressions (McCullough et al., 2013). By seeking to punish those who cause us harm—that is, seeking revenge on transgressors—we may acquire a retaliatory reputation, and deter others from causing us harm in the future.

In the draft project below, we study such dynamics using the repeated punishment game, where we consider interactions between one actor any many different partners. Each partner has one opportunity to transgress on the actor, who may engage in costly retaliatory punishment. We show that, in order for partner cooperation to be enforced, revenge must serve a deterrence function: the actor punishes following a transgression, and as a result it is less likely that future partners will transgress as well.

We show that our model can be extended to explain the finer features of human revenge. First feature: the cost of apologizing. As is commonly noted (e.g., Fitouchi & Singh, 2023), revenge does not always deter future transgressions: retaliation can spark counter-retaliation, and lead to the deterioration of mutually beneficial relationships. For this reason, revenge is often envisioned along with forgiveness, whose function is to preserve such valuable relationships (McCullough et al., 2013).

Apologies from offenders then play an important role. Yet it is unclear when offenders must bear costs, and when mere expressions of regret are enough. Experiments offer contrasting evidence (Abeler et al., 2010; Ohtsubo & Watanabe, 2009), or lump costly and cost-free apologies together (McCullough et al., 2014). We show that our model makes a simple prediction on this matter: apologies must be sufficiently costly to deter future transgressions, so they must be costly when transgressions are beneficial to the offender, and they can be cost-free when the offender does not benefit (and, e.g., the transgression is accidental).

The second feature of human revenge that our model helps explain is why, when deciding whether someone deserves punishment, we sometimes overlook important information, or rely on information that should be irrelevant. An emblematic example is moral luck (Nagel, 1979) (for other examples, see Carlsmith et al., 2002). The same harmful behavior (e.g., reckless driving) is judged less severely when, by pure chance, no one was there to be hurt. In this example, we overlook intention (recklessness) and rely instead on outcome (whether or not someone was there to be hurt)—even though it is only the former that impacts the expected level of harm caused by the behavior (for an experimental test, see Cushman et al., 2009).

To explain moral luck and other similar phenomenons, we extend our model. We show that the actor exacts revenge conditional on information that is common knowledge with partners, and not based on private knowledge. Since it is in general easier to agree on the outcome of an action that on the intentions of an individual, this may be why we observe phenomenons like moral luck. While a system designed to optimize deterrence *stricto sensu* should take into account intention rather than outcome, our individual retaliatory sentiments are best understood as designed to protect our retaliatory reputation.

The draft manuscript, printed below, is followed by a supplementary information, in which we detail the mathematical model and its results. In its private version, the document is then followed by a presentation of the experimental stimuli we have run ('baseline' series of vignettes), and plan to run ('apologies' and 'private vs. common knowledge' series of vignettes). This experimental material (16 pages) is taken out of the online version of this dissertation.

The repeated punishment game explains why, and when, we seek revenge

Julien Lie-Panis Bethany Burum Christian Hilbe Moshe Hoffman

November 23, 2023

Abstract

A prominent explanation for revenge is the deterrence of future transgressions. Yet revenge often fails to enact optimal deterrence. We forgive others' dangerous behavior based on variables out of their control, such as a lucky positive outcome. Minor transgressions can lead to full-blown conflict, rather than a proportionate response. In addition, the role of costly apologies remains unclear—when do apologies need to involve costs, and when are mere words enough? Here, we address these gaps in our knowledge of revenge using a mathematical model. This model—the repeated punishment game—involves one actor and many successive partners. Each partner has one opportunity to transgress on the actor, while the actor may decide to engage in retaliatory punishment; that is, exact revenge. We show that revenge serves to deter future transgressions. By extending our model, we show that apologies must be costly when transgressions are beneficial, and that they can be cost-free when transgressions are non-beneficial. Finally, our model suggests that revenge should be overly sensitive to information that is likely to be common knowledge—such as a lucky positive outcome—and relatively insensitive to information that is likely to be privately held—such as the exact severity of a transgression.

We are currently testing the predictions from our mathematical model using a series of vignette studies.

Keywords: *revenge — justice — game theory*

Introduction

When we are wronged, we often seek revenge. Across social contexts and cultures (Anderson, 2000; Boehm, 2008; Daly & Wilson, 1988; Ericksen & Horton, 1992; Gambetta, 2009; Nisbett & Cohen, 1996), people are motivated to impose retaliatory costs on wrongdoers. These vengeful motivations, and the ensuing vengeful behavior—revenge—are not only costly to the target but also to the perpetrator. Perpetrators of revenge assume immediate risks, potentially instigate counter-retaliation, or worse, kindle a destructive cycle of revenge (Fitouchi & Singh, 2023; Glowacki, 2022). These considerable risks beg the question—why do individuals seek revenge when it would be safer to forgive and forget?

A prominent explanation for revenge is the deterrence of future transgressions (McCullough et al., 2013). In lab experiments, subjects who employ retaliatory punishment are more often helped by their partners (Molm, 1997; although see Raihani and Bshary, 2019). What's more, subjects use retaliatory punishment for offenses committed against others strategically, to deter future transgressions against themselves (Krasnow et al., 2012; Krasnow et al., 2016), and close allies (Delton & Krasnow, 2017). Outside of the lab, revenge is also fueled by reputational concerns (Anderson, 2000; Brezina et al., 2004; Crombag et al., 2003; Ericksen & Horton, 1992; Gambetta, 2009; IJzerman et al., 2007; Nisbett & Cohen, 1996). By enacting revenge, individuals demonstrate that they are not to be trifled with.

Revenge does not always enforce future cooperation, however. Retaliatory punishment can instead spark counter-retaliation, and lead to the deterioration of mutually beneficial relationships (Fitouchi & Singh, 2023; Raihani & Bshary, 2019). For this reason, revenge is usually envisioned along with forgiveness, which serves to restore valuable relationships

(McCauley et al., 2022; McCullough et al., 2013). To communicate relationship value and seek forgiveness, offenders can then use conciliatory gestures, such as an apology or an offer of compensation (Burnette et al., 2012; Forster et al., 2021; McCullough et al., 2014, see also Fehr and Gelfand, 2010).

Will mere words suffice? Apologies have been shown to be more effective when they are costly for offenders (Ohtsubo and Watanabe, 2009; although see Abeler et al., 2010). Yet it remains unclear when offenders need to bear costs, and when cost-free apologies are enough to seek forgiveness.

Another issue: even when revenge acts as a deterrent, it very often fails to enact optimal deterrence. In their punishment decisions, people overlook key pieces of information—the very details that an optimal deterrence system would use to calibrate the level of punishment. For instance, when dealing with offenses of the same category, people overlook the severity of the offense (up to one order of magnitude, Burum et al., in prep.; although see Molho et al., 2020), even though it would be optimal to be less punishing when offenses are smaller, to avoid unnecessary costs.

Similarly, people overlook the likelihood of detection (Carlsmith et al., 2002), even though it would be optimal to be more punishing when offenders are less likely to get caught. On the other hand, people are sensitive to factors that are irrelevant for optimal deterrence, such as a lucky positive outcome (Cushman et al., 2009). The same harmful behavior is judged less severely when, by pure chance, no one was there to be hurt, a phenomenon known as moral luck (Nagel, 1979).

To address these gaps in our understanding of revenge, we develop a model called the repeated punishment game. We consider interactions between one actor and many different partners. Partners each have one opportunity to transgress on the actor, and the actor may subsequently engage in retaliatory punishment (revenge).

Revenge serves to deter transgressions from future partners in our model. In any cooperative subgame perfect equilibrium, partners do not transgress on the actor, and this cooperative behavior is maintained by the threat of punishment. If a transgression is to occur (e.g., by accident), the actor pays an immediate cost to punish, and recoups that cost by deterring transgressions from future partners (whereas failing to punish leads to future exploitation).

We extend our model to look at the cost of apologies. We show that apologies must deter transgression to be effective. When transgressions benefit the partner, apologies must then be (sufficiently) costly. In contrast, when transgressions do not benefit the partner, a cost-free apology is sufficient. Accidental transgressions do not need to be punished—in that case, forgiveness can be granted on the basis of words alone (see also Martinez-Vaquero et al., 2015).

We also extend our model to look at apparently sub-optimal deterrence. In one extension, we allow for imperfect monitoring of a partner's action. We show that the actor punishes conditional on publicly shared information, even when the actor privately knows that information to be false. Since intentions and likelihoods are hard to observe, and when observe unlikely to be shared, our model can explain why revenge tends to overlook intention to harm and likelihood of detection (Carlsmith et al., 2002; Cushman et al., 2009). Conversely, our model can explain why we appear oversensitive to outcomes (even when they are the result of pure luck), since outcomes are more likely to constitute shared information.

In another extension of our model, we show that the actor will ignore the exact severity of a transgression, when this severity is not common knowledge—as long as it is observed with any positive noise by future partners. This may explain why revenge overlooks the severity of categorically similar transgressions (Burum et al., in prep.).

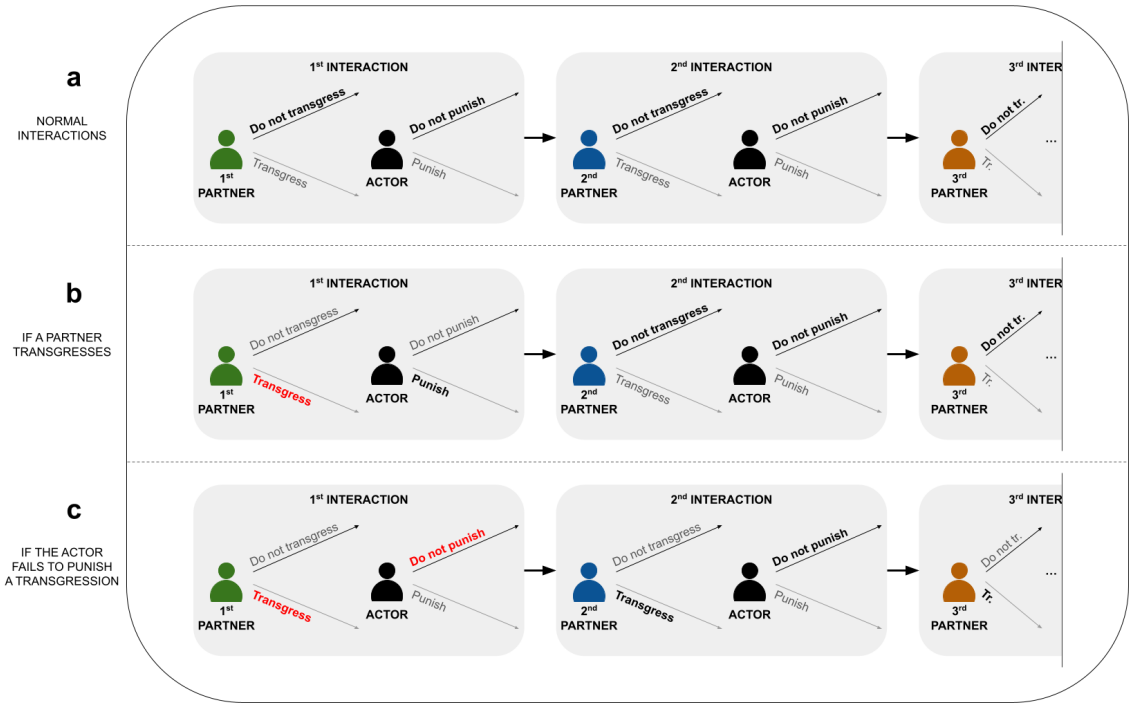


Figure 1: **Results for the baseline model.** We consider successive interactions between one actor and many different partners. In each interaction, a partner first decides whether or not to transgress on the actor, who then decides whether or not to punish that partner. With probability δ, another interaction ensues. We demonstrate properties that are shared by all cooperative equilibria. To illustrate, we consider here a specific cooperative equilibrium called maximal deterrence. **a:** In any cooperative subgame perfect equilibrium, when things go according to plan, partners do not transgress and the actor does not punish. **b:** If a partner transgresses however, the actor punishes that partner. Some level of cooperation is then restored; here, with maximal deterrence, every future partner can be expected not to transgress. **c:** If the actor fails to punish this transgression, she is exploited in the future: strictly more future partners will transgress than had she punished. Here, with maximal deterrence, every future partner can be expected to transgress.

We are currently testing all the predictions from our model using a series of vignette studies.

The repeated punishment game

Baseline model and results

The repeated punishment game involves two types of individuals: one actor, and an infinite pool of different partners. The actor interacts with partners one at a time, in different rounds. In each round, a partner begins by deciding whether or not to transgress on the actor. Then, the actor decides whether or not to punish that partner.

Once a round is over, the game repeats with probability δ (with probability $1-\delta$, the game stops). A new round ensues, pitting the same actor against a new partner, who interact following the rules outlined just above. Once that round is over, yet another round occurs with probability δ—and so on. The actor can thus engage in many different rounds, while partners only engage in one round (in the language of repeated games, the actor is a long-run player, while partners are short-run players, Mailath and Samuelson, 2006).

Transgressions are beneficial to partners, but harmful to the actor: every time a partner

transgress on the actor, he gains $b > 0$ and she loses $c > 0$.[1] We will also refer to not transgressing as cooperating: by opting not to transgress, a partner renounces a benefit to avoid harming the actor. Punishment is harmful to both types of individuals: every time the actor punishes a partner, she loses $\gamma_a > 0$ and he loses $\gamma_p > 0$.

A common issue with repeated games is that they admit multiple equilibria (e.g., Fudenberg & Maskin, 1986). To get around this issue, we do two things. First, we concentrate on subgame perfect equilibria: these are equilibria which are evolutionarily stable assuming the existence of noise; that is, assuming that individuals mistakenly play the action not recommended by their strategy with a small positive probability (Selten, 1983). Second, we consider the entire set of cooperative subgame perfect equilibria (from here on: of cooperative equilibria). We look for features that are shared by all cooperative equilibria.

We show that cooperative equilibria all share two features: our approach reveals two necessary features for cooperation to occur in an evolutionary endpoint, without having to pinpoint which one. To illustrate, we consider a specific strategy profile, which we call maximal deterrence, whereby:

- In each round, the partner does not transgress, so long as no prior transgression has gone unpunished; otherwise he transgresses.
- The actor punishes if a partner has transgressed in this round and no prior transgression has gone unpunished; otherwise she does not punish.

First feature: transgressions are punished. For cooperation to occur, partners have to be incentivized by the threat of punishment. In any cooperative equilibrium, when things go according to plan (in the language of repeated games, along the outcome path), partners do not transgress and the actor does not have to punish—normal play is represented on panel a of Figure 1.

As represented on panel b of Figure 1, the first partner to transgress is punished by the actor. Compared to normal play, that partner gains b but then loses γ_p. Thus, for partners not to transgress, a necessary condition is that the cost of being punished outweigh the benefit of transgression, that is:

$$\gamma_p \geq b \tag{1}$$

Second feature: not punishing leads to exploitation. For punishment to occur, the actor has to be incentivized by the threat of exploitation. In any cooperative equilibrium, if a partner transgresses along the outcome path, the actor punishes. Some level of cooperation is then restored: at least some of the actor's future partners will abstain from transgression—in the case of maximal deterrence for instance, every single future partner can then be expected not to transgress, as represented on panel b of Figure 1.

As represented on panel c Figure 1, if the actor fails to punish that transgression, she can expect to be exploited more often—in the case of maximal deterrence, every single future partner can then be expected to transgress. Maximal deterrence thus represents the case in which punishment is maximally incentivized. Comparing to the case in which she punishes, not punishing leads the actor to save on the cost γ_a. However, she is then exploited in all future encounters, losing $\delta c + \delta^2 c + ... = (\delta c)/(1-\delta)$. For actors to punish transgressions, a necessary condition is therefore that the cost of being exploited by all future partners outweigh the cost of punishing, that is:

$$\frac{\delta}{1-\delta} c \geq \gamma_a \tag{2}$$

Put differently, for cooperation to be sustained in our model, transgressions have to be punished, and retaliatory punishment—revenge—has to serve a deterrence function. The

actor benefits from punishing an exploitative partner because doing so sets a precedent, deterring exploitation by future partners—at best, punishment leads to all future partners switching from transgression to cooperation (maximal deterrence).

Apologies

How can partners escape punishment? Will a mere apology do? We extend our model to allow partners to apologize after a transgression, by paying a positive or null cost $s \geq 0$. We look at the conditions under which apologies are sustained in a cooperative equilibrium, by once again considering the entire set of these equilibria, and assuming that (a) partners apologize following an unexpected transgression, and (b) the actor subsequently does not punish. We show that a necessary condition is (see Supplementary section 3):

$$s \geq b \tag{3}$$

For apologies to be effective, they must deter transgression: their cost must exceed the benefit of transgression. When transgressions are beneficial ($b > 0$), it follows that apologies must be costly ($s > 0$). Conversely, we show that when transgressions aren't beneficial ($b \leq 0$), apologies can be cost-free ($s = 0$), by showing the existence of a cooperative equilibrium in which cost-free apologies are sustained when $b \leq 0$.

Private vs. common knowledge

When might the actor ignore valid information? Implicitly, up until now, we have been making an assumption called perfect monitoring. We have been assuming that every individual's behavior is observed without error by others. As a result, each time a partner transgresses, the actor and all her future partners have this information, and the actor has no reason to ignore it—in the language of game theory, the fact that the partner transgressed is common knowledge: in particular, the actor knows it, and knows that future partners know it.

In another extension of our model, we allow for noisy observation of partner's behavior. Each time a partner acts, two signals are created, to represent private observation by the actor on the one hand, and public observation by all individuals on the other (see Figure 2). First, the actor receives a private signal which always reflects the actual action, e.g., 'transgress' if the partner did in fact transgress. Second, to reflect the fact that individuals outside of the interaction may have more imperfect knowledge of what transpired, the entire population receives a noisy public signal. The public signal reflects the actual action with probability $1-\varepsilon$, and the other action with probability ε, e.g. 'did not transgress' even though the partner did in fact transgress ($0 < \varepsilon < 1$).

We show that the actor ignores the private signal in every cooperative equilibrium, and instead punishes conditionally on the public signal—even when the latter gives false information (see Supplementary section 5). Our model suggests that revenge should be sensitive to information that, like the public signal, is likely to be commonly known—such as the outcome of dangerous behavior (lucky or not)— and relatively insensitive to information that is unlikely to be commonly known—such as harmful intentions or likelihood of detection.

Other extensions to the baseline model

When might the actor overlook the severity of the transgression? In another extension of the model, we allow the magnitude of transgression to vary in the first round, by multi-

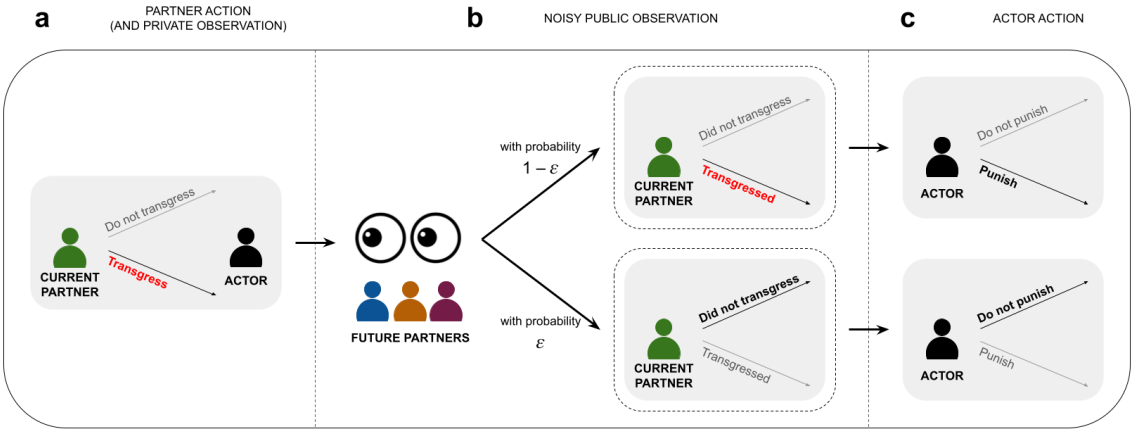


Figure 2: **Results for the model with imperfect monitoring.** We extend our model to include noisy observation of partner's behavior. **a**: Each time a partner acts, his action is observed privately by the actor, without any error. **b**: In addition, the partner's action is observed publicly—by the actor as well as future partners. Public observation is noisy: here, the public mistakenly observes cooperation with probability ε, even though the current partner transgressed. **c**: The actor punishes based on the result of public observation: here, when the public mistakenly observes cooperation, the actor does not punish, even though she knows that a transgression did occur (and vice-versa when the public mistakenly observes transgression—this case is not depicted in the figure).

plying by λ both the benefit of transgression for the first partner and its cost for the actor. We assume that λ can take any real value, and that future partners do not observe the exact value of λ, but instead receive an imperfect signal of its value (see Supplementary Information section 6, and Yoeli et al. 2022 for a more general model). We show that as long as this signal is noisy; that is, as long as the precise value of λ is not commonly known, the actor ignores this value, punishing minuscule transgressions and larger transgressions alike.

Lastly, through a further extension of our model, we demonstrate that for revenge to serve as a deterrent, administering punishment must come with a cost (see Supplementary section 7). Indeed, when in contrast $\gamma_a \leq 0$, retaliatory punishment does not necessarily deter transgressions: there exists a cooperative equilibrium in which failing to punish a transgression does not lead to future exploitation.

Experimental tests of the repeated punishment game

Baseline vignettes

In a series of vignette studies, we test the predictions arising from our model of revenge. First, in our baseline model, we showed that retaliatory punishment serves a deterrence function. If a transgression does occur, the actor sets a precedent by punishing (or failing to), thus deterring transgression by future partners. Based on our model, we make two predictions: compared to not punishing, punishing a transgressor should lead to (a) lower change of a future transgression, and (b) higher chance that a future transgression is punished.

We test these two predictions in our baseline vignette study. To do so, we recruited 231 English-speaking subjects living in the United States of America using Prolific. Subjects' age ranged from 19 to 80 years old, with a median at 35. 142 subjects identified as female, and 79 subjects identified as male (10 answered 'Other / Prefer not to say'). In terms of

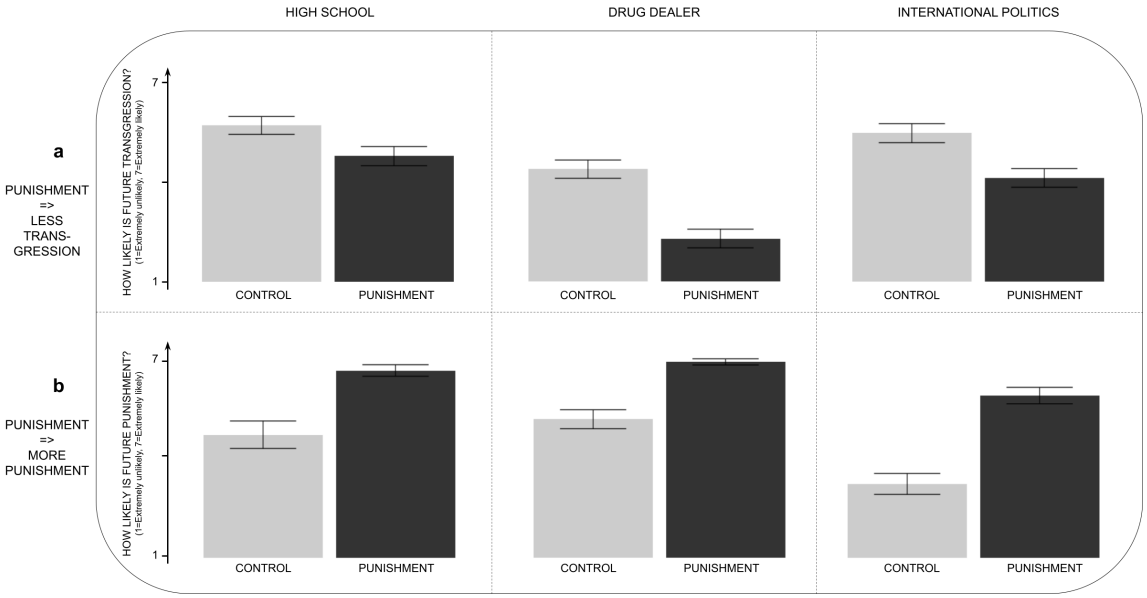


Figure 3: **Results for the baseline vignettes**. Subjects were presented with three vignettes, called high school, drug dealer, and international politics (left to right). In each case, a transgression occurs. Subjects either see the control condition, or the punishment condition, in which the transgression is subsequently punished (between-subject design). We collected two dependent variables (one per prediction), by asking subjects to rate on a 1-7 scale the likelihood of a similar transgression occurring in the future, and, were it to occur, of such a transgression being punished. For both of our (directional) predictions, we conducted a unidirectional independent samples t-test, obtaining significant results in all six cases. Error bars represent the 95% confidence interval. **a**: Our first prediction is that the likelihood of future transgression is lower in the punishment condition. High school: mean of control = 5.63, mean of punishment = 4.71, $p < 0.001$. Drug dealer: mean of control = 4.31, mean of punishment = 2.22, $p < 0.001$. International politics: mean of control = 5.40, mean of punishment = 4.05, $p < 0.001$. **b:** Our second prediction is that the likelihood of future punishment is higher in the punishment condition. High school: mean of control = 4.63, mean of punishment = 6.57, $p < 0.001$. Drug dealer: mean of control = 5.10, mean of punishment = 6.82, $p < 0.001$. International politics: mean of control = 3.15, mean of punishment = 5.81, $p < 0.001$.

highest attained level of education, 54 subjects reported attending or finishing graduate school, 143 reported attending or finishing college, and 34 reported finishing high school. They were £1 to complete the study; the median completion time was 8 minutes and 48 seconds.

We present subjects with three naturalistic vignettes called high school, drug dealer, and international politics. In each case, a transgression occurs in a cooperative relationship, respectively: a high-school student kisses her friend's crush, a drug dealer is robbed by inhabitants of the same neighborhood, and a member of an international trade alliance invades a neighboring country. Subjects are presented with one of two conditions: either a control condition or a punishment condition, where the latter features a response to the transgression—either the silent treatment, a punitive expedition, or economic sanctions, as the case may be. We ask subjects about the likelihood of a similar transgression occurring in the future, and, were it to occur, of such a transgression being punished. As shown in Figure 3, this series of vignettes validate both of our predictions.

Ongoing: apologies

In an extension to our baseline model, we showed that effective apologies also had to deter transgressions. When transgressions are beneficial, apologies must be sufficiently costly. We deduce our first prediction: (a) compared to cost-free apologies, sufficiently costly apologies should lower the likelihood of future transgression.

In contrast, when transgressions don't provide benefits, apologies can be cost-free. We deduce our second prediction: (b) there should be a negative interaction of whether apologies are sufficiently costly vs. cost-free and whether transgressions are beneficial vs. non-beneficial.

We will test these two predictions in another vignette study. We will present subjects with two naturalistic vignettes called high school and drug dealer. In both cases, a transgression occurs in a cooperative relationship, and is followed by an apology. Subjects will see one of four conditions, obtained by crossing between cost-free vs. (sufficiently) costly apology, and beneficial vs. non-beneficial transgression. We will ask them about the likelihood of a similar transgression occurring in the future.

Ongoing: private vs. common knowledge

In another extension to our baseline model, we showed that under imperfect monitoring, the actor ignores private information, and instead punishes conditionally on shared information—even when the former is true and the latter is false. Based on this extension, we predict that: (a) when a transgression occurs, retaliatory punishment should be more likely when the transgression is common knowledge compared to privately observed, and (b) when a transgression does not occur but an action looks suspicious, retaliatory punishment should be more likely when the fact that the action was not a transgression is common knowledge, compared to privately observed.

This study is ongoing as well. We will test our first prediction using two vignettes called bully and infidelity. Bully: a student is bullied at school in an empty hall (privately observed) vs. a packed hallway (publicly observed[2]). Infidelity: a spouse's affair from a long-time ago is discovered via a WhatsApp message that pops onto the computer while the spouse is in the shower (privately observed) vs. while the couple is watching Netflix (common knowledge).

We will test our second prediction using a vignette called Donnie Brasco. In this vignette, a notoriously foolish member of a crime family vouches for a new recruit, who is revealed to be an undercover FBI agent. Leaders of the family must decide whether to kill the foolish member, whom they all privately suspect was duped and was not working for the FBI (privately observed). In one condition, video surveillance footage reveals to all the leaders assembled in the same room that the member was most certainly duped (common knowledge).

Discussion

Using a mathematical model, we show that revenge can serve a deterrence function. In any cooperative equilibrium, individuals punish transgressions against them, and this vengeful behavior serves to deter future transgressions. We extend our model to include apologies, imperfect observation, and variable magnitude of transgression.

Our model extensions show that, to successfully deter, apologies should be sufficiently costly to the initial transgressor, and that the victim of the transgression should retaliate based on information that is common knowledge; they should, for instance, refrain from

punishing transgressions that appear benign, and fully punish minuscule transgressions when the magnitude of transgression is not commonly known. Individual deterrence may then explain why we believe that offenders should suffer to pay a debt to their victim (Miller, 2005), and why retaliatory punishment and retributive intuitions tend to be insensitive to intentions (Cushman et al., 2009; Nagel, 1979), magnitudes (Burum et al., 2023), frequencies and probabilities (Carlsmith et al., 2002)—cues which, while relevant to the cost and benefit of transgression, are rarely common knowledge, and therefore rarely relevant to the cost and benefit of individual deterrence.

Insensitivity to such cues is often seen as a reason to argue against a deterrence account for revenge (K. M. Carlsmith & Darley, 2008; Fitouchi et al., 2023; Wenzel & Thielmann, 2006). In order to be successfully deterred, small and unintentional transgressions, for instance, appear to require less punishment. Our model shows that this is in fact not always the case. Intentions and magnitude only affect the expected cost of future transgressions when they are common knowledge. This suggests that the explanatory power of deterrence has been underestimated, and that it may in fact explain many of the seemingly quirky features of revenge.

References

Abeler, J., Calaki, J., Andree, K., & Basek, C. (2010). The power of apology. *Economics Letters*, *107*(2), 233–235. https://doi.org/10.1016/j.econlet.2010.01.033

Anderson, E. (2000). *Code of the street: Decency, violence, and the moral life of the inner city* (1. ed). Norton.

Boehm, C. (2008). Purposive social selection and the evolution of human altruism. *Cross-cultural Research - CROSS-CULT RES*, *42*, 319–352. https://doi.org/10.1177/1069397108320422

Brezina, T., Agnew, R., Cullen, F., & Wright, J. (2004). The code of the street: A quantitative assessment of elijah anderson's subculture of violence thesis and its contribution to youth violence research. *Youth Violence and Juvenile Justice*, *2*, 303–328. https://doi.org/10.1177/1541204004267780

Burnette, J. L., McCullough, M. E., Van Tongeren, D. R., & Davis, D. E. (2012). Forgiveness results from integrating information about relationship value and exploitation risk. *Personality & Social Psychology Bulletin*, *38*(3), 345–356. https://doi.org/10.1177/0146167211424582

Burum, B., Dalkiran, A., Yoeli, E., & Hoffman, H. (2023). Why norms are categorical [Joint first authorship]. *Under review*.

Carlsmith, Darley, & Robinson. (2002). Why do we punish? deterrence and just deserts as motives for punishment [Publisher: J Pers Soc Psychol]. *Journal of personality and social psychology*, *83*(2). https://doi.org/10.1037/0022-3514.83.2.284

Carlsmith, K. M., & Darley, J. M. (2008). Psychological aspects of retributive justice [Publisher: Elsevier]. *Advances in Experimental Social Psychology*, *40*, 193–236. https://doi.org/10.1016/s0065-2601(07)00004-4

Crombag, H., Rassin, E., & Horselenberg, R. (2003). On vengeance [Publisher: Routledge _eprint: https://doi.org/10.1080/1068316031000068647]. *Psychology, Crime & Law*, *9*(4), 333–344. https://doi.org/10.1080/1068316031000068647

Cushman, F., Dreber, A., Wang, Y., & Costa, J. (2009). Accidental outcomes guide punishment in a "trembling hand" game (L. Santos, Ed.). *PLoS ONE*, *4*(8), e6699. https://doi.org/10.1371/journal.pone.0006699

Daly, M., & Wilson, M. (1988). *Homicide*. A. de Gruyter.

Delton, A. W., & Krasnow, M. M. (2017). The psychology of deterrence explains why group membership matters for third-party punishment. *Evolution and Human Behavior*, *38*(6), 734–743. https://doi.org/10.1016/j.evolhumbehav.2017.07.003

Ericksen, K. P., & Horton, H. (1992). "blood feuds": Cross-cultural variations in kin group vengeance [Publisher: SAGE Publications]. *Behavior Science Research*, *26*(1), 57–85. https://doi.org/10.1177/106939719202600103

Fehr, R., & Gelfand, M. J. (2010). When apologies work: How matching apology components to victims' self-construals facilitates forgiveness. *Organizational Behavior and Human Decision Processes*, *113*(1), 37–50. https://doi.org/10.1016/j.obhdp.2010.04.002

Fitouchi, L., Baumard, N., & André, J.-B. (2023). Just desert: Contractualist computations explain moralistic punishment. *In preparation*.

Fitouchi, L., & Singh, M. (2023). Punitive justice serves to restore reciprocal cooperation in three small-scale societies. *Evolution and Human Behavior*. https://doi.org/10.1016/j.evolhumbehav.2023.03.001

Forster, D. E., Billingsley, J., Burnette, J. L., Lieberman, D., Ohtsubo, Y., & McCullough, M. E. (2021). Experimental evidence that apologies promote forgiveness by com-

municating relationship value [Number: 1 Publisher: Nature Publishing Group]. *Scientific Reports*, *11*(1), 13107. https://doi.org/10.1038/s41598-021-92373-y

Fudenberg, D., & Maskin, E. (1986). The folk theorem in repeated games with discounting or with incomplete information [Publisher: [Wiley, Econometric Society]]. *Econometrica*, *54*(3), 533–554. https://doi.org/10.2307/1911307

Gambetta, D. (2009). *Codes of the underworld: How criminals communicate*. Princeton University Press. Retrieved April 28, 2021, from https://www.jstor.org/stable/j.ctt7tbn3

Glowacki, L. (2022). The evolution of peace. *Behavioral and Brain Sciences*, 1–100. https://doi.org/10.1017/S0140525X22002862

IJzerman, H., van Dijk, W. W., & Gallucci, M. (2007). A bumpy train ride: A field experiment on insult, honor, and emotional reactions [Place: US Publisher: American Psychological Association]. *Emotion*, *7*(4), 869–875. https://doi.org/10.1037/1528-3542.7.4.869

Krasnow, M. M., Cosmides, L., Pedersen, E. J., Tooby, J., & Zalla, T. (2012). What are punishment and reputation for? [Publisher: Public Library of Science]. *PLoS ONE*, *7*(9), 1–9. https://doi.org/10.1371/journal.pone.0045662

Krasnow, M. M., Delton, A. W., Cosmides, L., & Tooby, J. (2016). Looking under the hood of third-party punishment reveals design for personal benefit [Publisher: SAGE Publications Inc]. *Psychological Science*, *27*(3), 405–418. https://doi.org/10.1177/0956797615624469

Mailath, G. J., & Samuelson, L. (2006). *Repeated games and reputations: Long-run relationships* [OCLC: ocm61821903]. Oxford University Press.

Martinez-Vaquero, L. A., Han, T. A., Pereira, L. M., & Lenaerts, T. (2015). Apology and forgiveness evolve to resolve failures in cooperative agreements [Number: 1 Publisher: Nature Publishing Group]. *Scientific Reports*, *5*(1), 10639. https://doi.org/10.1038/srep10639

McCauley, T. G., Billingsley, J., & McCullough, M. E. (2022). An evolutionary psychology view of forgiveness: Individuals, groups, and culture. *Current Opinion in Psychology*, *44*, 275–280. https://doi.org/10.1016/j.copsyc.2021.09.021

McCullough, M. E., Kurzban, R., & Tabak, B. A. (2013). Cognitive systems for revenge and forgiveness. *Behavioral and Brain Sciences*, *36*(1), 1–15. https://doi.org/10.1017/S0140525X11002160

McCullough, M. E., Pedersen, E. J., Tabak, B. A., & Carter, E. C. (2014). Conciliatory gestures promote forgiveness and reduce anger in humans [Publisher: Proceedings of the National Academy of Sciences]. *Proceedings of the National Academy of Sciences*, *111*(30), 11211–11216. https://doi.org/10.1073/pnas.1405072111

Miller, W. I. (2005, December 19). *Eye for an eye* [Google-Books-ID: _RMiOXoLnncC]. Cambridge University Press.

Molho, C., Tybur, J. M., Van Lange, P. A. M., & Balliet, D. (2020). Direct and indirect punishment of norm violations in daily life [Number: 1 Publisher: Nature Publishing Group]. *Nature Communications*, *11*(1), 3432. https://doi.org/10.1038/s41467-020-17286-2

Molm, L. D. (1997). *Coercive power in social exchange* [Pages: xii, 316]. Cambridge University Press. https://doi.org/10.1017/CBO9780511570919

Nagel, T. (1979). *Mortal questions*. New York: Cambridge University Press.

Nisbett, R. E., & Cohen, D. (1996). *Culture of honor: The psychology of violence in the south* [Pages: xviii, 119]. Westview Press.

Ohtsubo, Y., & Watanabe, E. (2009). Do sincere apologies need to be costly? test of a costly signaling model of apology. *Evolution and Human Behavior*, *30*(2), 114–123. https://doi.org/10.1016/j.evolhumbehav.2008.09.004

Raihani, N. J., & Bshary, R. (2019). Punishment: One tool, many uses. *Evolutionary Human Sciences*, *1*, e12. https://doi.org/10.1017/ehs.2019.12

Selten, R. (1983). Evolutionary stability in extensive two-person games. *Mathematical Social Sciences*, *5*(3), 269–363. https://doi.org/10.1016/0165-4896(83)90012-4

Wenzel, M., & Thielmann, I. (2006). Why we punish in the name of justice: Just desert versus value restoration and the role of social identity. *Social Justice Research*, *19*(4), 450–470. https://doi.org/10.1007/s11211-006-0028-2

Notes

1. To avoid lengthy repetitions of the terms actor and partner, we have assigned a gender to each type of individual based on the result of a coin toss; throughout this text, we will use feminine pronouns (she/her) to refer to the actor, and masculine pronouns to refer to partners (he/him/his).

2. Technically, this is broader than just testing common vs. private knowledge, since we vary whether an audience observes the bullying, rather than whether the bullying is common knowledge between the concerned parties.

Supplementary information for:

The repeated punishment game explains why, and when, we seek revenge

November 23, 2023

List of main results

In this document, we develop several instantiations of a repeated game involving two players, player 1 and player 2. For ease of reading, we have assigned a gender to each player based on the result of a coin toss—we will use masculine pronouns to refer to player 1 (he/him/his), and feminine pronouns (she/her) to refer to player 2.

In each round, player 1 first decides whether to transgress on player 2, thereby imposing a cost on her (by default, transgressions benefit him). Then, player 2 decides whether to punish player 1, thus imposing a cost on him (by default, punishment is costly to her).

We analyze the necessary features of cooperation in our repeated game, and show how changing certain assumptions affects these features. To do so, we demonstrate properties shared by all strategy profiles that are: (i) subgame perfect equilibria of the instantiation of the repeated game under consideration, and (ii) cooperative, in that they induce non-transgression from player 1 in each period. We show that, along the outcome path:

R.1 If player 1 transgresses, he is punished in that round (section 1.2);

R.2 If player 2 does not punish following a player 1 transgression, she is exploited in at least one future round (section 1.2);

R.3 When transgressions do not benefit player 1, they need not be punished by player 2 (section 2.2);

R.4 Apologies from player 1 must be costly, and their cost must exceed the benefit of transgression (section 3.3);

R.5 When transgressions do not benefit player 1, apologies can be cost-free (section 4.2);

R.6 If player 1 appears not to transgress, player 2 does not punish, even when she knows that the transgression did in fact occur (section 5.3);

R.7 If player 1 appears to transgress, player 2 punishes, even when she knows that the transgression did not occur (section 5.3);

R.8 When the magnitude of transgression is not commonly known, players cannot ignore small transgressions and concentrate on large transgressions. If player 1 engages in a small transgression in the initial round, player 2 punishes in that round (proven in section 6.3);

R.9 When players behave according to a cooperative Nash equilibrium rather than a cooperative subgame perfect equilibrium, failing to punish need not lead to exploitation (section 7.1).

'Results' **R.1-R.9** are formulated as more mathematically precise 'propositions', and proven, below. We also prove various subsidiary results, formulated as 'lemmas'.

Contents

1 Baseline model **3**
- 1.1 Set up . . . 3
- 1.2 Proof of results R.1 and R.2 . . . 6
- 1.3 Subsidiary result: parameter space for which $\mathcal{S} \neq \emptyset$ . . . 8

2 Model with non-beneficial transgressions **10**
- 2.1 Changes to the baseline model . . . 10
- 2.2 Proof of result R.3 . . . 10

3 Model with apologies **10**
- 3.1 Changes to the baseline model . . . 10
- 3.2 Subsidiary result: parameter space for which $\mathcal{S}^{apo} \neq \emptyset$ . . . 12
- 3.3 Proof of result R.4 . . . 13

4 Model with apologies and non-beneficial transgressions **13**
- 4.1 Changes to the previous model . . . 13
- 4.2 Proof of result R.5 . . . 13

5 Model with imperfect monitoring **14**
- 5.1 Changes to the baseline model . . . 14
- 5.2 Subsidiary results . . . 15
- 5.3 Proof of Results R.6 and R.7 . . . 16

6 Model with variable transgression **17**
- 6.1 Changes to the baseline model . . . 17
- 6.2 Subsidiary result . . . 18
- 6.3 Proof of Result R.8 . . . 18

7 Baseline model with Nash equilibria **19**
7.1 Proof of result R.9 19

1 Baseline model

1.1 Set up

1.1.1 Basic set up (common to all models)

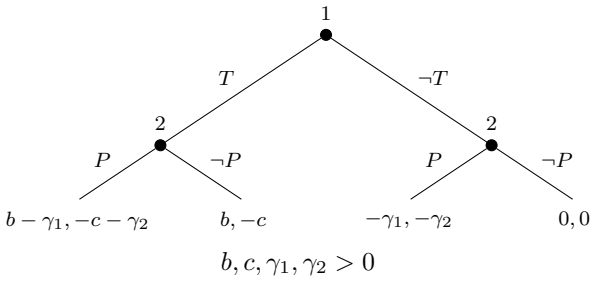


Figure 1: Stage game payoffs.

We build a model using Mailath and Samuelson's (2006) repeated game framework. Many short-run players and one long-run player engage in a repeated game. Short-run players are replaced every round, and the long-run player is the same every round. The stage game described below is infinitely repeated, for each of the rounds $t \in \mathbb{N}$.

Each stage proceeds as follows (see Figure 1). First, the short-run player, player 1, either transgresses (plays T) or does not transgress (plays $\neg T$). Second, the long-run player, player 2, either punishes (P) or does not punish ($\neg P$). We note $A_1 \equiv \{T, \neg T\}$ and $A_2 \equiv \{P, \neg P\}$ the set of actions that players 1 and 2 may respectively undertake; and $A \equiv A_1 \times A_2$ the set of stage game action profiles.

In each stage, players receive payoffs as follows. If player 1 transgresses, he gains benefit $b > 0$, and player 2 incurs a cost $c > 0$; if he does not transgress, neither players' payoffs are affected. If player 2 punishes, she pays $\gamma_2 > 0$, and player 1 incurs a cost $\gamma_1 > 0$; if she does not punish, neither players' payoffs are affected. We note u_1 and u_2 the payoffs accrued by players 1 and 2 respectively in the stage game, as a function of the chosen action profile $a \in A$ (e.g. $u_1(T, P) = b - \gamma_1$ and $u_2(T, P) = -c - \gamma_2$).

1.1.2 Histories for the baseline model

In the baseline model, we assume perfect monitoring. In each period, both players observe the action chosen by player 1 first, and by player 2 second.

A history of the repeated game is a finite sequence of alternating actions for both players, starting with player 1. The initial history is the empty sequence

$\emptyset$; play begins in period 0 with player 1 playing an action $a_1^0 \in A_1$ at history $\emptyset$. Play continues with player 2 playing an action $a_2 \in A_2$ at history (a_1^0). We then switch to period 1, with (another individual in the role of) player 1 playing an action a_1^1 at history (a_1^0, a_2^0), followed by player 2, who plays at history (a_1^0, a_2^0, a_1^1), etc.

A period t history for player 1 $h_1^t \equiv (a_1^0, a_2^0, ..., a_1^{t-1}, a^{t-1})$ identifies the $2t$ actions played by both players in periods 0 through $t-1$. We note $\mathscr{H}_1^t \equiv A^t$ the set of period t histories for player 1, A^t being the t-fold product of the set of stage game action profiles A.

The addition of a player 1 action a_1^t to a period t player 1 history h_1^t yields a period t player 2 history $h_2^t \equiv (h_1^t, a_1)$, which identifies the $2t$ actions played by both players in periods 0 through $t-1$, and the action played by player 1 in period t. We note $\mathscr{H}_2^t \equiv \mathscr{H}_1^t \times A_1$ the set of period t histories for player 2. The addition of a player 2 action a_2^{t+1} to a period t player 2 history h_2^t yields a period $t+1$ player 1 history, which we note (h_2^t, a_2^{t+1}).

The set of all possible histories for player i ($i = 1$ or $i = 2$) is: $\mathscr{H}_i \equiv \bigcup_{t=0}^{\infty} \mathscr{H}_i^t$. We note $\mathscr{H} \equiv \mathscr{H}_1 \cup \mathscr{H}_2$ the set of all possible histories for the repeated game.

1.1.3 Notations and definitions (common to all models)

We use notations from Mailath and Samuelson (2006), which we adapt to our asynchronous game. A pure strategy for player i is a mapping from the set of all possible histories he may face into the set of pure actions he may undertake:

$$\sigma_i : \mathscr{H}_i \to A_i.$$

We restrict our analysis to pure strategies. A pure strategy profile $\sigma \equiv (\sigma_1, \sigma_2)$ comprises two maps, and specifies game play given any possible history for the repeated game; when a history $\hbar_i \in \mathscr{H}_i$ is reached, player i plays action $\sigma_i(\hbar_i)$. We further note $\sigma \mid_{h_i} \equiv (\sigma_i \mid_{h_i}, \sigma_{-i} \mid_{(h_i, \sigma_i(h_i))})$ the strategy profile for the subgame that follows history h_i. For any strategy profile $\sigma = (\sigma_1, \sigma_2)$, any player $i \in \{1, 2\}$, and any history $h_i \in \mathscr{H}_i$, we note $\sigma_i \mid_{h_i}$ the continuation strategy of player i induced by the history h_i.

(Note that the subgame that follows a player 1 history h_1 is the continuation game, as it is commonly defined. A continuation strategy for player 1 given a history h_1 is then a mapping from $\mathscr{H}_1$ to A_1. Following a player 2 history h_2, we obtain, however, a subgame starting with a player 2 action. A continuation strategy for player 2 given a history h_2 is then a mapping from a slightly different set $\mathscr{H}_2'$ to A_2, $\mathscr{H}_2' \equiv \mathscr{H}_2 \mid_{h_2} = \{\emptyset\} \cup (A_2 \times \mathscr{H}_2)$ being the set of lists h_2' of even length, identifying the actions played by player 2, first, and player 1 second, between the player 2 histories h_2 and (h_2, h_2').)

We note $\mathbf{a}(\sigma) \equiv (a^0(\sigma), a^1(\sigma), a^2(\sigma), ...) \in A^\infty$ the outcome path induced by a strategy profile σ, which is an infinite sequence of action profiles; for any t, $a^t(\sigma) \equiv (a_1^t(\sigma), a_2^t(\sigma))$ designates the action profile induced by σ in round t. The first t periods of the outcome path induced by σ are noted $\mathbf{a}^t(\sigma) \equiv (a^0, a^1, ..., a^{t-1})$.

Player 1 is short-run, and participates in a single round of the repeated game. Given strategy profile $\sigma = (\sigma_1, \sigma_2)$ and history $h_1 \in \mathscr{H}_1$, player 1 earns payoff:

$$u_1(\sigma \mid_{h_1}) = u_1(\sigma_1(h_1), \sigma_2(h_1, \sigma_1(h_1))).$$

Player 2 is long-run, and participates in all rounds of the repeated game, which are infinite. We assume player 2 is characterized by a fixed discount factor $\delta \in (0, 1)$; payoffs obtained t rounds in the future are discounted by factor δ^t in the current round. δ can be interpreted to represent player 2's patience, or the probability that the game repeats after a given round, i.e. the probability that player 2 will in fact face other individuals in the future, after having played an action—even though these are conceptually different objects, using one instead of the other does not affect the results.

Given strategy profile $\sigma = (\sigma_1, \sigma_2)$ and history $h_1 \in \mathscr{H}_1$, player 2 earns continuation payoff:

$$U_2(\sigma \mid_{h_1}) \equiv (1 - \delta) \sum_{t=0}^{\infty} \delta^t u_2(a^t(\sigma \mid_{h_1})).$$

We define player 2's continuation payoff following a player 2 history using the above definition. Given strategy profile $\sigma = (\sigma_1, \sigma_2)$ and history $h_2 \in \mathscr{H}_2$, player 2 earns continuation payoff:

$$U_2(\sigma \mid_{h_2}) \equiv \begin{cases} -(1-\delta)\gamma_1 + \delta U_2(\sigma \mid_{(h_2, P)}) & \text{if } \sigma_2(h_2) = P \\ \delta U_2(\sigma \mid_{(h_2, \neg P)}) & \text{if } \sigma_2(h_2) = \neg P \end{cases}$$

Note that we rely on this complex formula because player 2 histories occur 'at the middle' of a round, and because we do not have a more economic way of noting the payoffs accrued from just considering the second half of the round (i.e. just from whether player 2 punishes). The formula above simply states that player 2's continuation payoff h_2 is equal to the payoff accrued by punishing or not punishing in this round, plus the discounted future payoff obtained starting in the next round, which is a player 1 history (which is either (h_2, P) or $(h_2, \neg P)$), and can therefore be calculated using the simpler formula before.

A strategy profile $\sigma = (\sigma_1, \sigma_2)$ is a Nash equilibrium of the repeated game if for any player 1 strategy σ_1', any player 2 strategy σ_2', and any player 1 history of the form $\mathbf{a}^t(\sigma)$ (i.e. any player 1 history along the outcome path):

$$u_1(\sigma \mid_{\mathbf{a}^t(\sigma)}) \geq u_1(\sigma_1', \sigma_2 \mid_{\mathbf{a}^t(\sigma)})$$
$$U_2(\sigma) \geq U_2(\sigma_1, \sigma_2')$$

A strategy profile σ is a subgame perfect equilibrium of the repeated game if for any history $\hbar \in \mathscr{H}$, $\sigma \mid \hbar$ is a Nash equilibrium of the repeated game.

A one-shot deviation for player i from strategy σ_i is a strategy $\hat{\sigma}_i \neq \sigma_i$, with the property that there exists a unique history $\tilde{h}_i \in \mathscr{H}_i$ such that, for all other player i histories $h_i \in \mathscr{H}_i \setminus \{\tilde{h}_i\}$: $\sigma_i(\hbar_i) = \hat{\sigma}_i(\hbar_i)$.

Let $\sigma = (\sigma_1, \sigma_2)$ be a strategy profile. A player 1 one-shot deviation $\hat{\sigma}_1$ is profitable if, at the history $\tilde{h}_1$ for which $\hat{\sigma}_1(\tilde{h}_1) \neq \sigma_1(\tilde{h}_1)$, $u_1(\hat{\sigma}_1, \sigma_2 \mid_{\tilde{h}_1}) > u_1(\sigma_1, \sigma_2 \mid_{\tilde{h}_1})$. A player 2 one-shot deviation $\hat{\sigma}_2$ is profitable if, at the history $\tilde{h}_2$ for which $\hat{\sigma}_2(\tilde{h}_2) \neq \sigma_2(\tilde{h}_2)$, $U_2(\sigma_1, \hat{\sigma}_2 \mid_{\tilde{h}_2}) > U_2(\sigma_1, \sigma_2 \mid_{\tilde{h}_1})$. Following the one-shot deviation principle, a strategy profile is subgame perfect if and only if there are no profitable one-shot deviations.

1.1.4 Objective (common to most models)

A strategy profile σ is said to be *cooperative* if it induces non-transgression from player 1 in each period, i.e. if:

$$\forall t \in \mathbb{N}, \qquad a_1^t(\sigma) = \neg T.$$

Let $\mathcal{S}$ be the set of strategy profiles which are (i) subgame perfect equilibria of the repeated game, and (ii) cooperative. Our objective is to study this set.

1.2 Proof of results R.1 and R.2

In this section, we derive two general properties, which apply when players behave according to any strategy profile $\sigma \in \mathcal{S}$. We show first that, along the outcome path, if player 1 transgresses, then he is punished by player 2 (Proposition 1.1). We show second that, following a transgression occurring along the outcome path, not punishing is followed by transgression by player 1 in at least one future round (Proposition 1.2).

Note that these results can be made more general—this is throughout this document. The first result applies to all player 1 histories h_1 for which σ prescribes not transgressing—including along the outcome path, i.e. for a history of the form $h_1 = \mathbf{a}^t(\sigma)$. The second result applies to all player 2 histories h_2 for which σ prescribes punishment—including following a transgression occurring along the outcome path, i.e. for a history of the form $h_2 = (\mathbf{a}^t(\sigma), T)$. We chose to focus on such particular histories throughout this document.

1.2.1 If player 1 transgresses along the outcome path, he is punished

Proposition 1.1

$$\forall \sigma = (\sigma_1, \sigma_2) \in \mathcal{S}, \forall t \in \mathbb{N}, \sigma_2(\mathbf{a}^t(\sigma), T) = P$$

Along the outcome path of a cooperative subgame perfect equilibrium, if player 1 transgresses, he is punished in that round.

Proof: let us consider $\sigma = (\sigma_1, \sigma_2) \in \mathcal{S}$, and a player 1 history along the outcome path $h_1 = \mathbf{a}^t(\sigma)$. When players play as prescribed by σ given h_1, player 1 does not transgress (since σ is cooperative), obtaining no benefit. He thus obtains a payoff this round that is negative or null, depending on whether player 2 subsequently punishes: $u_1(\sigma \mid_{h_1}) \leq 0$.

We prove that $\sigma_2(h_1, T) = P$ by contraposition, by showing that σ cannot be subgame perfect if player 2 does not punish given (h_1, T). Let us therefore assume $\sigma_2(h_1, T) = \neg P$ and consider the one-shot deviation $\hat{\sigma_1}$ from strategy σ_1, whereby player 1 transgresses given h_1 (and otherwise plays as prescribed by σ_1). We note $\hat{\sigma} = (\hat{\sigma_1}, \sigma_2)$ the resulting strategy profile. If player 1 unilaterally deviates to $\hat{\sigma_1}$, he gains a strictly positive payoff, since player 2 does not then punish: $u_1(\hat{\sigma} \mid_{h_1}) = b - 0 > 0$.

This deviation is then profitable: we must have $u_1(\hat{\sigma} \mid_{h_1}) > u_1(\sigma \mid_{h_1})$ since the first term is strictly positive, and the second negative or null. Since σ is subgame perfect this is impossible; by contraposition, it must in fact be that $\sigma_2(h_1, T) = P$.

1.2.2 If player 2 does not punish after player 1 transgresses along the outcome path, she is exploited

Proposition 1.2

$$\forall \sigma = (\sigma_1, \sigma_2) \in \mathcal{S}, \forall t \in \mathbb{N}, \exists t' \in \mathbb{N}, a_1^{t'}(\sigma \mid_{(\mathbf{a}^t(\sigma), T, \neg P)}) = T$$

Along the outcome path of a cooperative subgame perfect equilibrium, if player 2 *does not punish following a transgression, she is exploited in at least one future round.*

Proof: let us consider $\sigma = (\sigma_1, \sigma_2) \in \mathcal{S}$, and the player 2 history $h_2 = (\mathbf{a}^t(\sigma), T)$ obtained after player 1 transgresses at a history $\mathbf{a}^t(\sigma)$ along the outcome path.

Following the previous proposition, player 2 punishes given history h_2, losing $-(1-\delta)\gamma_2$ at this history. Her continuation payoff given h_2 is then:

$$U_2(\sigma \mid_{h_2}) = -(1-\delta)\gamma_2 + \delta U_2(\sigma \mid_{(h_2, P)})$$

Let us consider the one-shot deviation $\hat{\sigma}_2$ from strategy σ_2, whereby player 2 does not punish given h_2 (and otherwise plays as prescribed by σ_2). We note $\hat{\sigma} = (\sigma_1, \hat{\sigma_2})$ the resulting strategy profile. If player 2 unilaterally deviates to $\hat{\sigma}_2$, she does not punish, losing nothing at this history. Her continuation payoff given h_2 is then:

$$U_2(\hat{\sigma} \mid_{h_2}) = \delta U_2(\sigma \mid_{(h_2, \neg P)})$$

Since σ is subgame perfect, this deviation cannot be profitable; we must have:

$$U_2(\hat{\sigma} \mid_{h_2}) \leq U_2(\sigma \mid_{h_2})$$

Replacing, we deduce:

$$\delta(U_2(\sigma \mid_{(h_2, \neg P)}) \leq -(1-\delta)\gamma_2 + \delta U_2(\sigma \mid_{(h_2, P)}))$$

And therefore:

$$(1-\delta)\gamma_2 \leq \delta \times [U_2(\sigma \mid_{(h_2, P)}) - U_2(\sigma \mid_{(h_2, \neg P)})]$$

It follows that the difference in continuation payoffs must be strictly positive (since $\gamma_2 > 0$ and $0 < \delta < 1$); we must have:

$$U_2(\sigma \mid_{(h_2,\neg P)}) < U_2(\sigma \mid_{(h_2,P)})$$

By deviating from σ_2 to $\hat{\sigma}_2$, player 2 saves on the (strictly positive) cost of punishment at history h_2, but players subsequently reach history $(h_2, \neg P)$ instead of history (h_2, P). Since σ is subgame perfect, it must be that player 2's continuation payoff at history $(h_2, \neg P)$ is strictly smaller than the continuation payoff at history (h_2, P)—this is what we have just shown algebraically.

Note that player 2's maximum payoff is 0—at best, player 2 never incurs the cost of punishing or transgression by player 1. This is true for any continuation payoff as well: it follows that $U_2(\sigma \mid_{(h_2,P)}) \leq 0$. We deduce, using the previous inequality, that we must have: $U_2(\sigma \mid_{(h_2,\neg P)}) < 0$. After saving on the cost of punishment, player 2's continuation payoff must be strictly negative. Given history $(h_2, \neg P)$, when players play according to σ player 2 will incur a cost at least once, in some future period $t' \in \mathbb{N}$.

There are two possibilities. Either player 1 will transgress in that period $(a_1^{t'}(\sigma \mid_{(h_2,\neg P)}) = T)$, in which case the proposition is proven; or player 2 will punish in that period $(a_2^{t'}(\sigma \mid_{(h_2,\neg P)}) = P)$. We conclude by nothing that in this latter case player 1 must have transgressed just before, in the same period. The reasoning is analogous to the one detailed in the proof of Proposition 1.1. By contraposition, were we to have $a_1^{t'}(\sigma \mid_{(h_2,\neg P)}) = \neg T$, then, in this period player 1 would gain: $-\gamma_1 < b - \gamma_1$—deviation to playing T would then be strictly profitable. This is impossible, since σ is subgame perfect. It follows that our initial assumption was false: by contraposition, player 1 will transgress in this period. This proves the proposition: if player 2 does not punish following a transgression along the outcome path (at history $h_2 = (\mathbf{a}^t(\sigma), T)$), then she is exploited in some future round—player 1 will then transgress in t' periods, for some $t' \in \mathbb{N}$.

1.3 Subsidiary result: parameter space for which $\mathcal{S} \neq \emptyset$

Lemma 1.1 *There exists a cooperative subgame perfect equilibrium if and only if:*

$$b \leq \gamma_1 \tag{1.1}$$

$$(1-\delta)\gamma_2 \leq \delta c \tag{1.2}$$

Proof: let us assume $\mathcal{S} \neq \emptyset$, and consider a strategy profile $\sigma = (\sigma_1, \sigma_2) \in \mathcal{S}$. We immediately deduce condition (1.1) by considering the player 1 one-shot deviation to playing T given any history where he is expected not to transgress (as in the demonstration of Proposition 1.1). We deduce condition (1.2) by considering a player 2 deviation to a strategy σ_2' given a history h_2 where she is expected to punish, such that $\sigma' = (\sigma_1, \sigma_2') \mid_{h_2}$ induces non-punishment from player 2 in every period (e.g. deviation to playing $\neg P$ for any history $h \in$

$\mathscr{H}_2 \times \mathscr{H}'_2$, i.e. any history h which can be written in the form $h = (h_2, h'_2)$ with $h'_2 \in \mathscr{H}'_2$; and otherwise playing as prescribed by σ_2). Were player 2 to unilaterally deviate to σ'_2, she would earn:

$$U_2(\sigma/' \mid_{h_2}) = \delta U_2(\sigma' \mid_{h_2,\neg P})$$

Since σ is subgame perfect, we deduce (just as in the proof of Proposition 1.2):

$$\delta(U_2(\sigma \mid_{h_2,P}) - U_2(\sigma' \mid_{h_2,\neg P})) \geq (1-\delta)\gamma_2$$

Since $U_2(\sigma \mid_{h_2,P}) \leq 0$ (maximum possible payoff) and $U_2(\sigma' \mid_{h_2,\neg P})) \geq -c$ (worst possible payoff given that, when player 2 deviates to σ'_2, she never punishes following history h_2), we deduce the proposed inequality. This proves the implication.

To prove that we have an equivalence, let us assume the above inequalities hold. We then construct a strategy profile which we prove to sustain cooperation and be subgame perfect. Let σ be the strategy profile whereby: (i) player 1 plays T if a transgression has gone unpunished (given any history of the form $(a^0, a^1, ...T, \neg P, ...a^{t-1}) \in \mathscr{H}_1$), and otherwise plays $\neg T$, and (ii) player 2 plays $\neg P$ if a transgression has gone unpunished (given any history of the form $(a^0, a^1, ...T, \neg P, ...a^{t-1}, a^t_1) \in \mathscr{H}_2$), and otherwise plays P if and only if player 1 transgresses in the current round.

σ is cooperative. We show that σ is subgame perfect, by considering all possible one-shot deviations from σ, and showing that none of them are profitable.

Let us first consider a player 1 history $\hbar_1 \in \mathscr{H}_1$. If no transgression has gone unpunished, σ prescribes playing $\neg T$. A one-shot deviation to playing T given $\hbar_1$ allows player 1 to gain b, and leads her to be punished; under condition (1.1), this deviation isn't beneficial. If in contrast a transgression has gone unpunished, σ prescribes playing T. In this case, player 1 immediately does not benefit from deviation to playing $\neg T$, since doing so would entail losing benefit b without affecting player 2's behavior.

Let us now consider a player 2 history $\hbar_2 = (h_1, a_1) \in \mathscr{H}_2$ for which transgression has never gone unpunished. If $a_1 = \neg T$, player 2 immediately does not benefit from deviation to playing P since $\gamma_2 > 0$. If $a_1 = T$, player 2 can deviate to not punishing, in which case she saves on the cost of punishing in the current round, earning $(1-\delta)\gamma_2$. In future rounds, player 1 will then play T, leading to an increase in the future continuation cost of cooperation equal to δc. Under condition (1.2), this one-shot deviation isn't beneficial.

Finally, given a player 2 history $\hbar_2 \in \mathscr{H}_2$ for which a transgression has gone unpunished, player 2 immediately does not benefit from deviation to playing P, since $\gamma_2 > 0$. We have considered a partition of all possible histories: σ is subgame perfect. This proves the proposed equivalence.

2 Model with non-beneficial transgressions

2.1 Changes to the baseline model

We make only one change to the baseline model, and assume that the benefit of transgressing is negative or null, i.e. that:

$$b \leq 0.$$

Histories, (continuation) strategies, outcomes and payoffs are defined as in the baseline model, using the same notations. We conserve the same objective, and continue to note $\mathcal{S}$ the set of strategy profiles which are subgame perfect equilibria of the repeated game, and which sustain cooperation.

2.2 Proof of result R.3

In section 1.2.1, we showed that, in our baseline model where transgressions are beneficial ($b > 0$), unexpected transgressions by player 1 are punished. Below we show that $b > 0$ is necessary for this to occur, by showing that Proposition 1.1 does not hold in this modified model in which $b \leq 0$.

Proposition 2.1

$$\exists \sigma = (\sigma_1, \sigma_2) \in \mathcal{S}, \forall h_1 \in \mathscr{H}_1, \\ \sigma_1(h_1) = \neg T \wedge \sigma_2(h_1, T) = \neg P$$

Non-beneficial transgressions need not be punished. There exists a cooperative subgame perfect equilibrium in which unexpected transgressions are never punished.

Proof: let σ be the strategy profile obtained when player 1 plays $\neg T$ whatever the history $h_1 \in \mathscr{H}_1$, and player 2 plays $\neg P$ whatever the history $h_2 \in \mathscr{H}_2$. σ is cooperative. When players play according to σ and player 1 plays T instead of $\neg T$, he isn't subsequently punished by player 2.

We conclude by showing that σ is subgame perfect in this modified model. Indeed, there are no profitable one-shot deviations for player 1, since, given any history h_1, playing T entails losing $b \leq 0$, without affecting play by player 2. Similarly, there are no profitable one-shot deviations for player 2, since $\gamma_2 > 0$.

3 Model with apologies

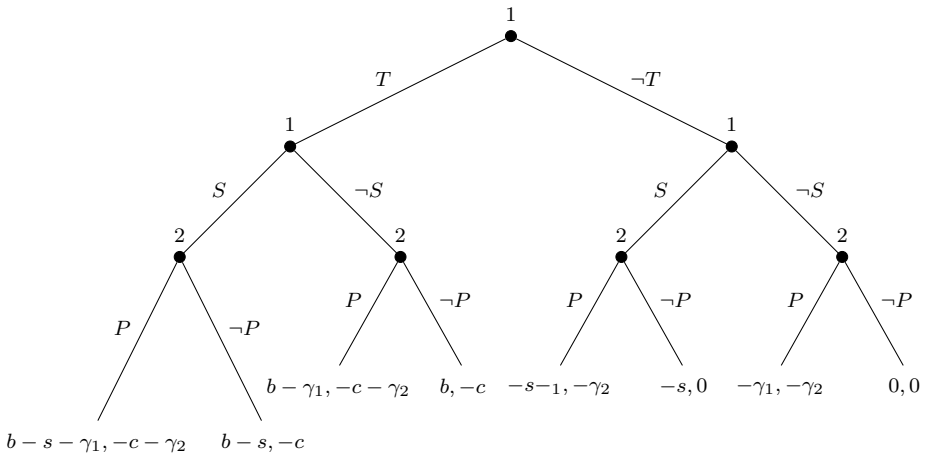
3.1 Changes to the baseline model

3.1.1 Modified stage game and histories

We modify the baseline model, by adding the possibility for player 1 to 'apologize', before player 2 has a chance to punish (see Figure 2). In each stage,

play now occurs in three steps. First, player 1 may opt to transgress or not, as before.

Second, player 1 can now either send an apology (play S), at cost $s \geq 0$; or opt not to send the apology (play $\neg S$), at no cost. The apology does not affect player 2's payoff. Third, player 2 may punish or not punish, as before.



$$b, c, \gamma_1, \gamma_2 > 0; \quad s \geq 0$$

Figure 2: Payoffs for the modified stage game (with apologies).

We note $A_{1'} = \{S, \neg S\}$, and $A \equiv A_1 \times A_{1'} \times A_2$ the set of stage game action profiles for this new model. We continue to assume perfect monitoring. A history for the repeated game is a finite sequence of alternating actions for both players, with player 1 now playing twice before player 2 plays. We note $\mathscr{H}_1$ the set of histories for player 1's first action, $\mathscr{H}_{1'}$ the set of histories for player 1's second action, and $\mathscr{H}_2$ the set of histories for player 2.

A pure strategy profile σ can now be written as a list of three maps $(\sigma_1, \sigma_1', \sigma_2)$, respectively specifying player 1's first action given a history $h_1 \in \mathscr{H}_1$, player 1's second action given a history $h_1' \in \mathscr{H}_{1'}$, and player 2's action given a history $h_2 \in \mathscr{H}_2$. We use the same notations as in our baseline model to refer to (continuation) strategies, outcomes and payoffs, which are defined using these newly defined histories.

3.1.2 New definitions and objective

Let $\sigma = (\sigma_1, \sigma_1', \sigma_2)$ be a cooperative strategy profile, i.e. one which induces non-transgression from player 1 in each period. When players play according to σ, player 1 is said to *use apologies* when he sends following an unexpected transgression, i.e. when, given any history h_1 at which he is not expected to

transgress ($\sigma_1(h_1) = \neg T$), we have: $\sigma_1'(h_1, T) = S$. Player 2 is said to *accept apologies* when she does not punish following an unexpected transgression and a subsequent apology, i.e. when, given any history h_1 at which player 1 is not expected to transgress, $\sigma_2(h_1, T, S) = \neg P$. Apologies are said to *occur* when player 1 uses apologies and player 2 accepts apologies.

Our objective is to study the set $\mathcal{S}^{apo} \subset \mathcal{S}$ of strategy profiles which (i) are coopertive subgame perfect equilibria of the repeated game, and, for which (ii) apologies occur.

3.2 Subsidiary result: parameter space for which $\mathcal{S}^{apo} \neq \emptyset$

Lemma 3.1 *$\mathcal{S}^{apo}$ is non-empty if and only if:*

$$b \leq s \leq \gamma_1 \tag{3.1}$$

$$(1-\delta)\gamma_2 \leq \delta c \tag{1.2}$$

Proof: let us assume $\mathcal{S}^{apo} \neq \emptyset$, and consider a strategy profile $\sigma = (\sigma_1, \sigma_1', \sigma_2)$ in that set. Let h_1 be a history at which player 1 is expected not to transgress. If player 1 engages in the one-shot deviation $\hat{\sigma}_1$ consisting in playing T given h_1, he subsequently apologizes, and gains $b - s$; otherwise, if he does not deviate, he gains 0. Since σ is subgame perfect, we must have: $0 \geq b - s \implies s \geq b$.

Now, we consider the one-shot deviation $\hat{\sigma}_1'$, whereby player 1 does not apologize given history (h_1, T). By engaging in this one-shot deviation, player 1 saves on the cost of apologizing s—for this deviation not to be beneficial, it must be that this he is then punished (he exists the game after player 2's move). In other words, by engaging in this deviation, player 1 gains $s - \gamma_1$—in a subgame perfect equilibrium, we must have: $0 \geq s - \gamma_1 \implies \gamma_1 \geq s$. Using the inequality obtained in the previous paragraph, we deduce condition (3.1).

Condition (1.2) can be proven by using the same reasoning as in the proof of lemma 1.1—by considering a unilateral deviation σ_2', whereby player 2 does not punish given $h_2 = (h_1, T, \neg S)$, and $(\sigma_1, \sigma_2, \sigma_2') \mid_{h_2}$ induces non-punishment in all subsequent rounds (we do not repeat the full reasoning here).

We have thus shown the implication: if $\mathcal{S}^{apo} \neg \emptyset$, then conditions (3.1) and (3.1) follow.

We show that we in fact have an equivalence by finding an element of the set $\mathcal{S}^{apo}$ assuming that these two conditions hold. Let σ be the strategy profile whereby: (i) player 1 transgresses given any history of the form $(a^0, ...T, \neg S, \neg P, ..., a^{t-1})$, i.e. when a transgression has gone 'unapologized for' and unpunished, (ii) player 1 does not send given any history of the form $(a^0, ...T, \neg S, \neg P, ..., a^{t-1}, a_0^t)$, and otherwise plays S if and only if he played T in the current round, and (iii) player 2 does not punish given any history of the form $(a^0, ...T, \neg S, \neg P, ..., a^{t-1}, a_1^t, a_{1'}^t)$, and otherwise plays P if and only if player 1 played T then played $\neg S$ in the current round.

σ is cooperative. Since b, s and γ_2 are positive or null, to check that σ is subgame perfect, we only need to check at histories at which transgressions have not gone unapologized for and unpunished. At such a history h_1, player

1 does not transgress, and deviation to playing T isn't beneficial, player 1 will then apologize. At a history $h_{1'} = (h_1, T)$, player 1 plays S, and deviation to playing $\neg S$ isn't beneficial because player 2 will then punish. At a history $h_{1'} = (h_1, \neg T)$, player 1 plays $\neg S$; deviation to playing S is non-beneficial because it is costly and does not affect player 2's future action.

Finally, at a history $h_2 = (h_1, T, \neg S)$, player 2 punishes, and gains $U_2(\sigma \mid_{h_2}) = -(1-\delta)\gamma_2$; deviation to not punishing isn't beneficial, because it leads to the break down of cooperation, and to losing $-\delta c$. At other histories of the form $h_2 = (h_1, a_1, a_{1'})$ where $(a_1, a_{1'})\neg(T, \neg S)$, player 2 does not punish, and deviation to playing P is immediately detrimental because it does not affect future play by player 1. This proves the proposed proposition.

3.3 Proof of result R.4

Proposition 3.1

$$\mathcal{S}^{apo} \neq \emptyset \implies b \leq s$$

In a cooperative subgame perfect equilibrium in which apologies occur, apologies are costly, and their cost exceeds the benefit of transgression.

Proof: immediate consequence of the above lemma.

4 Model with apologies and non-beneficial transgressions

4.1 Changes to the previous model

We make only two changes to the previous model, detailed in section 3, and assume that transgressions are non-beneficial (as in section 2), and that the cost of apologizing is null, i.e. that:

$$b \leq 0$$
$$s = 0.$$

We use the same notations as in the previous section, including for the strategy profile set $\mathcal{S}^{apo}$.

4.2 Proof of result R.5

Proposition 4.1

$$\mathcal{S}^{apo} \neq \emptyset$$

Apologies can be costless. There exists a cooperative subgame perfect in which apologies occur, at no cost to player 1.

Proof: let σ be the strategy profile whereby: (i) player 1 always plays $\neg T$, whatever the history $h_1 \in \mathcal{H}_1$; (ii) player 1 always plays S, whatever the history $h_1' \in \mathcal{H}_1'$; and (iii) player 2 always plays $\neg P$, whatever the history $h_2 \in \mathcal{H}_2$. σ sustains cooperation and, when players behave according to σ, player 1 sends given any history of the form $h_1' = (h_1, T)$ (hence utilizes apologies), and player 2 does not punish given any history of the form $h_2 = (h_1, T, S)$ (hence accepts apologies).

Since $b \leq 0$, $s = 0$ and $\gamma_2 > 0$, there are not profitable one-shot deviations for either player; σ is subgame perfect. This proves the proposition.

5 Model with imperfect monitoring

5.1 Changes to the baseline model

5.1.1 Imperfect monitoring and player i histories

In this section, we consider another modification to the baseline model, whereby the short-run player's action in a given round t is imperfectly observed by the short-run player in subsequent rounds $t' \geq t+1$. We interpet this to reflect interactions between one long-run player (player 2) and many different individuals, who each take on the role of player 1 once, in a given period. These individuals do not observe the behavior of individuals who have interacted with player 2 in the past; instead, they receive a noisy public signal.

More precisely, we assume that, in any given round, player 2 receives two signals right after player 1 plays, and before her own play. First, she receives a public signal y, which is observed by both players (i.e. will be available to player 1 in future rounds). y can take one of two values in the set $Y \equiv \{\mathcal{T}, \neg\mathcal{T}\}$. Errors are symmetric and positive: when player 1 transgresses (does not transgress), $y = \mathcal{T}$ ($y = \neg\mathcal{T}$) with probability $1 - \varepsilon$, and $y = \neg\mathcal{T}$ ($y = \mathcal{T}$) with probability ε (we assume: $0 < \varepsilon < \frac{1}{2}$).

Second, player 2 receives a private signal z, which player 1 does not observe. z can take one of two values in the set $Z \equiv \{T, \neg T\}$. We assume no errors for the private signal: when player 1 transgresses (does not transgress), $z = T$ ($z = \neg T$) with certainty.

We make no other changes to the baseline model, and continue to use the same notations. In particular, player 2's actions are observed without error by player 1.

Player 1 and player 2 now face histories of a different nature. A period t player 1 history $h_1^t \in \mathscr{H}_1^t$ is a sequence of the form $h_1^t \equiv (y^0, a_2^0, ...y^{t-1}, a_2^{t-1})$, specifying values taken by the public signal, and player 2's chosen action, in previous rounds. A player 1 history contains public information only, which is available to all players. As before, we note $\mathscr{H}_1 = \bigcup_{t\in\mathbb{N}} \mathscr{H}_1^t$ the set of player 1 (public) histories; a pure strategy for player 1 is a mapping: $\sigma_1 : \mathscr{H}_1 \to A_1$.

A period t player 2 history $h_2^t \in \mathscr{H}_2^t$ is a sequence of the form $h_2^t \equiv (y^0, z^0, a_2^0, ...y^{t-1}, z^{t-1}, a_2^{t-1}, y^t, z^t)$, specifying values taken by the public and private signals, and player 2's chosen action, in previous rounds; as well as the

current value of the public and private signals. A player 2 history contains public and private information. As before, we note $\mathscr{H}_2 = \bigcup_{t\in\mathbb{N}} \mathscr{H}_2^t$ the set of player 2 histories; a pure strategy for player 2 is a mapping: $\sigma_2 : \mathscr{H}_2 \to A_2$.

In contrast to the baseline model, a pure strategy profile σ induces a non-deterministic *path of play*, rather than a specific outcome. For instance, for a cooperative strategy profile σ, player 1's induced action in a given round t is $a_1^t(\sigma) = \neg T$, and the value of the public signal is either $y^t(\sigma) = \neg\mathcal{T}$ with probability $1-\varepsilon$, or $y^t(\sigma) = \mathcal{T}$ with probability ε. We continue to note $\mathcal{S}$ the set of such strategy profiles.

5.2 Subsidiary results

Lemma 5.1

$$\forall \sigma = (\sigma_1, \sigma_2) \in \mathcal{S}, \forall h_1 = (y^0, a_2^0, ...y^{t-1}, a_2^{t-1}) \in \mathscr{H}_1, \forall (z^0, ...z^{t-1}, z) \in \{T, \neg T\}^{t+1},$$
$$\sigma_1(h_1) = \neg T \implies \begin{cases} \sigma_2(y^0, a_2^0, z^0, ...y^{t-1}, z^{t-1}, a_2^{t-1}, \mathcal{T}, z) = P \\ \sigma_2(y^0, a_2^0, z^0, ...y^{t-1}, z^{t-1}, a_2^{t-1}, \neg\mathcal{T}, z) = \neg P \end{cases}$$

In a cooperative subgame perfect equilibrium, player 2 *punishes unexpected transgression conditional on the public signal.*

Proof: since $b > 0$ and $\varepsilon < \frac{1}{2}$, it must be that, along the path of play, player 2 punishes unexpected transgressions conditional on the public or the private signal, following an analogous reasoning to the one detailed in the demonstration of Proposition 1.1.

Following an analogous reasoning to the one detailed in the demonstration of Proposition 1.2, since $\gamma_2 > 0$, it must then be that not punishing a transgression will lead to exploitation in a future round. Since player 1 does not observe the private signal, it must be that punishing conditional on the value of y will be incentivized (i.e. that not punishing when $y = \mathcal{T}$ along the path of play will lead to a decrease in player 2's continuation payoff due to sufficient exploitation by player 1 in subsequent rounds), and that punishing conditional on the value of z won't be. This proves the above lemma.

Lemma 5.2 *$\mathcal{S}$ is non-empty if and only if:*

$$b \leq (1-2\varepsilon)\gamma_1 \tag{5.1}$$
$$(1-\delta)\gamma_2 + \varepsilon\delta\gamma_2 \leq \delta c \tag{5.2}$$

Proof: let us assume $\mathcal{S} \neq \emptyset$, and consider $\sigma = (\sigma_1, \sigma_2) \in \mathcal{S}$. We prove the first condition using the above lemma: since player 2 punishes conditional on the public signal, along the path of play, player 1 does not transgress, and is punished with probability ε. If he deviates, he gains b but is punished with probability $1-\varepsilon$; by comparing both payoffs we deduce condition (5.1).

We deduce condition (5.2) by consider a player 2 history h_2 at which she is expected to punish. Given h_2, if player 2 does not deviate from σ_2, she

punishes in the current round, and loses $(1-\delta)\gamma_2$. At best, cooperation will be fully restored in future rounds, and she will only have to punish in the future with probability ε; she earns: $U_2(\sigma \mid_{h_2}) \leq -(1-\delta)\gamma_2 - \delta\varepsilon\gamma_2$.

If she deviates to never punishing, a strategy we note σ_2', then, at worst she is exploited in all future rounds. Noting $\sigma' = (\sigma_1, \sigma_2')$, she earns: $U_2(\sigma' \mid_{h_2}) \geq 0 - \delta c$. We obtain condition (5.2) by comparing both continuation payoffs.

Conversely, let us assume both conditions hold. We consider the strategy profile σ whereby: (i) player 1 plays T given any history of the form $(y^0, a_2^0, ...\mathcal{T}, \neg P, ...y^{t-1}, a_2^{t-1}) \in \mathscr{H}_1)$, and (ii) players 2 plays $\neg P$ given any history of the form $(y^0, z^0, a_2^0, ...\mathcal{T}, z, \neg P, ...y^{t-1}, z^{t-1}, a_2^{t-1}, y^t, z^t) \in \mathscr{H}_2)$, and otherwise plays P if and only if $y = \mathcal{T}$ in the current round.

σ is cooperative. Since punishment is costly and transgression beneficial, we only need to consider one-shot deviations occurring along the path of play for both players (i.e. before what publicly looks like a transgression goes unpunished), when σ prescribes $\neg T$ for player 1 or P for player 2.

Along the path of play, deviation to playing T leads to being punished with probability $1-\varepsilon$; given condition (5.1), this one-shot deviation isn't beneficial. Deviation to playing $\neg P$ when P is prescribed (i.e. when $y = \mathcal{T}$), leads to saving on the immediate cost of punishment, and a full break-down of cooperation; instead, punishing as prescribed allows full cooperation to continue. We are in the extreme case of the above calculation; given condition (5.2), this one-shot deviation isn't beneficial. This proves the proposed lemma.

5.3 Proof of Results R.6 and R.7

Proposition 5.1

$$\forall \sigma = (\sigma_1, \sigma_2) \in \mathcal{S}, \forall h_1 = (y^0, a_2^0, ...y^{t-1}, a_2^{t-1}) \in \mathscr{H}_1, \forall (z^0, ...z^{t-1}) \in \{T, \neg T\}^t,$$
$$\sigma_1(h_1) = \neg T \implies \sigma_2(y^0, a_2^0, z^0, ...y^{t-1}, z^{t-1}, a_2^{t-1}, \neg\mathcal{T}, T) = \neg P$$

In a cooperative subgame perfect equilibrium, if player 1 *transgresses without it being commonly known, player* 2 *does not punish.*

Proposition 5.2

$$\forall \sigma = (\sigma_1, \sigma_2) \in \mathcal{S}, \forall h_1 = (y^0, a_2^0, ...y^{t-1}, a_2^{t-1}) \in \mathscr{H}_1, \forall (z^0, ...z^{t-1}) \in \{T, \neg T\}^1,$$
$$\sigma_1(h_1) = \neg T \implies \Big\{\sigma_2(y^0, a_2^0, z^0, ...y^{t-1}, z^{t-1}, a_2^{t-1}, \mathcal{T}, \neg T) = P$$

In a cooperative subgame perfect equilibrium, if player 1 *appears to transgress when he is expected not to, player* 2 *punishes, even when knows that the transgression did not occur.*

Proof: these two propositions follow from Lemma 5.1. Since player 2 punishes based on the public signal and not on her private signal, then she does not punish transgressions that appear not to have occurred (Proposition 5.1), and does punish when it looks like a transgression occurred even though she knows that not to be the case (Proposition 5.2).

6 Model with variable transgression

6.1 Changes to the baseline model

In this section, we consider another modification to the baseline model, whereby the magnitude of transgression in the initial round is variable, and imperfectly observed.

Before the initial round, nature draws a number $\omega \in \Omega$. To avoid edge cases, we assume that ω is randomly drawn from the real line $\Omega = \mathbb{R}$, according to the (improper) uniform distribution over Ω. The state of the world ω determines the magnitude of transgression in that round $\lambda(\omega) = \frac{e^{\omega}}{1+e^{\omega}}$: when player 1 transgresses in the initial round, he gains $\lambda(\omega) \times b$, and player 2 loses $\lambda(\omega) \times c$. Note that the function λ is an increasing bijection from $\mathbb{R}$ to $(0, 1)$: to each state of the world ω corresponds a unique value of the magnitude of the transgression which is between 0 and 1.

Neither player observes ω directly; instead they receive a signal, which is at most from $\varepsilon \geq 0$ of the value of the state of nature. After ω is drawn, player i ($i = 1, 2$) receives a signal s_i, which is uniformly drawn from the interval $[\omega - \varepsilon, \omega + \varepsilon]$. We note S_1 and S_2 the random variables representing each players' signal; and $T_1 = T_2 = \Omega = \mathbb{R}$ the set of possible values for s_1 and s_2.

We make no other changes to the baseline model, and continue to use the same notations. From round 1, the magnitude of transgressions is equal to 1. A player i strategy is now a mapping $\sigma_i : T_i \times \mathscr{H}_i \to A_i$, where the set of player i histories $\mathscr{H}_i$ and the set of player i actions A_i are defined as in the baseline model.

Our goal is to show that players cannot coordinate to ignore small transgressions occurring in the initial round. To do that we introduce the concept of threshold strategy profile, and extend our previous definition of cooperation, to include all strategy profiles whereby players ignore small transgressions (but not all transgressions) occurring in the initial round.

More specifically, let $\overline{s} \in \mathbb{R} \cup \{-\infty, +\infty\}$. A *threshold strategy profile for* $\overline{s}$ $\sigma^{\overline{s}}$ is defined as a strategy profile whereby: in the initial round, (i) player 1 transgresses if and only if he receives a signal $s_1 > \overline{s}$, and (ii) player 2 punishes if and only if player 1 transgresses and she receives a signal $s_2 > \overline{s}$. A *cooperative strategy profile* σ is defined as a strategy profile whereby: (i) there exists $\overline{s} \in \mathbb{R} \cup \{-\infty, +\infty\}$ such that σ is a threshold strategy profile for $\overline{s}$, (ii) cooperation occurs with non-null probability in the initial round, and (iii) given $s_1 > \overline{s}$ and given $s_2 > \overline{s}$, or given $s_1 \leq \overline{s}$ and $s_2 \leq \overline{s}$, $\sigma^{\overline{s}}$ induces non-transgression from player 1 in all rounds $t \geq 1$.

We note $\mathcal{T}^{\overline{s}}$ the set of threshold strategy profiles for a certain threshold $\overline{s}$. We continue to designate by $\mathcal{S}$ the set of strategy profiles which are: (i) cooperative, according the definition adopted in this section, and (ii) subgame perfect. In this section, cooperative strategy profiles are strategy profiles which induce cooperation from round 1 as long as there is 'agreement' on the signal value, in the sense that both players get a signal above or below the threshold. We include in the definition that cooperation must occur with non-null probability

in the initial round to exclude the case $\overline{s} = +\infty$ (players can always coordinate to ignore the initial round).

6.2 Subsidiary result

Below, we show that players cannot coordinate to ignore transgressions under a specific magnitude occurring in the initial round when assessments are noisy; if $\varepsilon > 0$, threshold strategies for $\overline{s} \neq -\infty$ are not subgame perfect.

Lemma 6.1 *Assume $\varepsilon > 0$. $\forall \overline{s} \in \mathbb{R}, \mathcal{T}^{\overline{s}} \cap \mathcal{S} = \emptyset$. Given any positive noise, any threshold strategy profile for a real number $\overline{s}$ cannot be a subgame perfect equilibrium of the repeated game.*

Proof: we prove this lemma by contraposition. Let us assume $\varepsilon > 0$, and consider a cooperative threshold subgame perfect strategy profile $\sigma^{\overline{s}}$, for a real number $\overline{s}$.

Let us assume that players play according to this strategy profile. We consider a player 2 history for round 0 of the form $h_2 = (T, s_2)$, with $s_2 \in (\overline{s}, \overline{s}+\varepsilon)$—in other words, we consider the case where player 1 transgresses, and player 2 receives a signal just above the threshold. The prescribed behavior is for player 2 to punish, paying γ_2. Following this, player 1 never transgresses and player 2 never punishes—player 2's continuation payoff is thus $U_2(\sigma \mid_{h_2}) = -\gamma_2$.

Yet, given that a transgression occurred in round 0 and that player 2's signal is sufficiently close to the threshold, it must be that player 1 faced $s_1 \leq \overline{s}$. Mathematically:

$$\mathbf{P}(s_1 \leq \overline{s} \mid T, s_2 \in (\overline{s} - \varepsilon, \overline{s})) = 1$$

In other words, given that player 1 is playing according to the strategy profile, and that the event $s_2 \in (\overline{s} - \varepsilon, \overline{s}) \cap s_1 \leq \overline{s}$ is non-null (since $\varepsilon > 0$), player 2 can deduce that player 1 must have faced a signal under the threshold at any history h_2 of the proposed form.

It follows that there is a beneficial one-shot deviation for player 2: if player 2 unilaterally deviates to not punishing given h_2, she saves on the cost of punishing without affecting future outcomes. Nothing the resulting strategy profile $\hat{\sigma}$, player 2 then obtains $U_2(\hat{\sigma} \mid_{h_2}) = 0 > U_2(\sigma \mid_{h_2})$, since, like σ, $\hat{\sigma}$ induces $(\neg T, \neg P)$ in all ulterior rounds.

Therefore, σ cannot be subgame perfect. Since this is true for any real value $\overline{s}$ and any cooperative threshold strategy profile for $\overline{s}$, this proves the proposed lemma.

6.3 Proof of Result R.8

Proposition 6.1 *If $\varepsilon > 0$, then $\mathcal{S} \subset \mathcal{T}^{-\infty}$. Any cooperative threshold strategy profile for $\overline{s}$ which is subgame perfect must verify: $\overline{s} = -\infty$.*

In a cooperative subgame perfect equilibrium, when the magnitude of transgression is not commonly known, players cannot coordinate to ignore small transgressions; except in the specific parameter case $b \neq \frac{\gamma_1}{2}$.

Proof: immediate consequence of the above lemma.

7 Baseline model with Nash equilibria

In this section, we return to the baseline model, formulated in section 1. Rather than the set $\mathcal{S}$ of cooperative subgame perfect equilibria, we consider the wider set $\mathcal{N} \supset \mathcal{S}$ of strategy profiles which are: (i) Nash equilibria of the repeated game, and (ii) cooperative.

In addition, we assume (without which both punishment cannot deter transgression, and both sets $\mathcal{S}$ and $\mathcal{N}$ are empty; see Lemma 1.1):

$$b \leq \gamma_1 \tag{1.1}$$

7.1 Proof of result R.9

Proposition 7.1

$$\exists \sigma = (\sigma_1, \sigma_2) \in \mathcal{N}, \forall h_2 \in \mathscr{H}_2 \ s.t. \ \sigma_2(h_2) = P$$
$$\forall t \in \mathbb{N}, a_1^t(\sigma \mid_{h_2, \neg P}) = \neg T$$

In a cooperative Nash equilibrium, failing to punish need not lead to future exploitation. There exists a cooperative Nash equilibrium in which player 1 *never transgresses, even if player* 2 *does not punish when expected to.*

Proof: let us consider the strategy profile σ whereby: (i) player 1 never transgresses, and (ii) player 2 punishes if and only if player 1 transgresses in this round round. σ is cooperative. By construction, σ verifies the above property.

We conclude by showing that $\sigma \in \mathcal{N}$. Since $b \leq \gamma_1$, player 1 does not stand to benefit from unilateral deviation to T given any history along the outcome path. Since c and γ_2 are negative and $U_2(\sigma) = 0$, player 2 has not beneficial deviations either; σ is a Nash equilibrium. This proves the proposed proposition.

Élément sous droit, diffusion non autorisée

Élément sous droit, diffusion non autorisée

Élément sous droit, diffusion non autorisée

Élément sous droit, diffusion non autorisée

Élément sous droit, diffusion non autorisée

Élément sous droit, diffusion non autorisée

Élément sous droit, diffusion non autorisée

Élément sous droit, diffusion non autorisée

Élément sous droit, diffusion non autorisée

Élément sous droit, diffusion non autorisée

Élément sous droit, diffusion non autorisée

Élément sous droit, diffusion non autorisée

Élément sous droit, diffusion non autorisée

Élément sous droit, diffusion non autorisée

Élément sous droit, diffusion non autorisée

Élément sous droit, diffusion non autorisée

Chapter 4

Runaway signals: Exaggerated displays of commitment may result from second-order signaling

Objectives and summary

Sometimes, people engage in seemingly dysfunctional displays, such as when all members of a group undergo the same initiation (Cimino, 2011; Densley, 2012) or feel compelled to engage in the same public ritual (Gelfand et al., 2020; Whitehouse & Lanman, 2014). These displays seem dysfunctional because they seem uniform. When everyone sends the same signal, receivers gain no new information; in theory, such a dishonest signal should be abandoned (Gintis et al., 2001).

Seemingly uniform signals may nevertheless be explained by adding something to the model, such as comparison to non-members or underlying variation in another trait. Uniform signals may function to screen individuals out of optional groups, thus serving to inform by comparison with non-members (Cimino, 2011; Densley, 2012). Alternatively, seemingly uniform signals can be informative to onlookers who possess other important information (in which case the signal only appears uniform to those who don't have this information). When the payoffs of signaling depend on a hidden quality (e.g., commitment to the group) and another trait (e.g., social capital), individuals can make inferences about quality based on seemingly uniform investment and cues of the other trait (Barker et al., 2019; Dumas et al., 2021).

In the article below (which is at the stage of journal pre-proof), we take an alternative route. We show that a signaling system can become truly dysfunctional through what we call *second-order signaling*. In our model, adapted from the signaling game introduced by Gintis et al. (2001), signalers can engage in two types of displays: they can invest in a costly display (signaling), and express outrage at non-senders (second-order signaling).

The purpose of outrage in our model is to advertise one's investment in the original

signal. When outrage is prohibitively costly for non-senders (e.g., because hypocrites are preferential targets of outrage), receivers can infer signaling behavior from second-order signaling: all individuals who express outrage have invested in the display. As a result, outrage can allow senders to reach a broader audience. In the model, we capture this by assuming that outrage increases a sender's chances of being observed: by default, signaling behavior is observed with probability p_1; expressing outrage increases the chances that an individual's signal (or lack thereof) is observed to $p_2 > p_1$.

When senders invest in outrage, everyone's incentive to send increases. Using our mathematical model, we study the conditions under which dishonest signaling may emerge. Using a computer simulation with several possible levels of investment in the signal, we show that outrage can lead signal costs to escalate.

The journal pre-proof, printed below, is followed by a supplementary information, in which we detail the model and simulation, as well as their results. A companion website for the simulation can be accessed at https://evolife.telecom-paris.fr/outrage/.

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Runaway signals: Exaggerated displays of commitment may result from second-order signaling

Julien Lie-Panis [a,b,*], Jean-Louis Dessalles [b]

[a] Institut Jean Nicod, Département d'études cognitives, Ecole normale supérieure, Université PSL, EHESS, CNRS, 75005 Paris, France
[b] LTCI, Télécom Paris, Institut Polytechnique de Paris, 91120 Palaiseau, France

ARTICLE INFO

Keywords:
Signaling
Commitment displays
Ritual
Game theory
Outrage

ABSTRACT

To demonstrate their commitment, for instance during wartime, members of a group will sometimes all engage in the same ruinous display. Such uniform, high-cost signals are hard to reconcile with standard models of signaling. For signals to be stable, they should honestly inform their audience; yet, uniform signals are trivially uninformative. To explain this phenomenon, we design a simple model, which we call the signal runaway game. In this game, senders can express outrage at non-senders. Outrage functions as a *second-order signal*. By expressing outrage at non-senders, senders draw attention to their own signal, and benefit from its increased visibility. Using our model and a simulation, we show that outrage can stabilize uniform signals, and can lead signal costs to run away. Second-order signaling may explain why groups sometimes demand displays of commitment from all their members, and why these displays can entail extreme costs.

1. Uniform investment in high-cost displays

Membership in human groups often involves ritual behaviors which appear arbitrary and wasteful to non-members, ranging from the embarrassment of hazing and the time constraints of religious practice to the emotional and physical scarring of certain rites or recruitment devices (Sosis et al., 2007; Atran and Henrich, 2010; Cimino, 2011; Densley, 2012; Whitehouse and Lanman, 2014). Drawing on honest signaling theory (Zahavi, 1975-09; Spence, 1974-03-01; Veblen, 1973; Grafen, 1990), these behaviors have been explained as displays of prosocial commitment (Irons, 2001; Sosis, 2003; Gambetta, 2009; Bulbulia and Sosis, 2011).

Yet, some commitment displays seem uniform, in direct contradiction to the predictions of honest signaling theory. Displays of commitment are often binary. Individuals decide whether or not to participate in a rite, or whether or not to comply with a prescription. When in addition investment is universal, that is when all group members engage in the binary display, the resulting signal is uniform (at least in first approximation, see also: Barker et al., 2019). Uniform signals are trivially dishonest. In theory, they should not be stable.

In the next section, we introduce an explanation for uniform displays, based on understanding outrage as a second-order signal of commitment. To formally investigate our theory, we adapt Gintis et al. (2001) multi-player model. When outrage is absent, signaling occurs only at an honest, non-uniform equilibrium, as shown in Section 3. In Section 4, we show that outrage can destabilize the honest signaling equilibrium, and lead to uniform signaling. In Section 5, we introduce a simulation of our model, and show that outrage can also lead to high-cost displays, through a step-by-step runaway process.[1] We discuss the scope of our model in Section 6.

2. Outrage as a second-order signal

Our aim here is to reconcile the existence of uniform displays with honest signal theory, based on a formal model. Mathematical signaling games have helped clarify the logic of a wide range of animal behaviors, pertaining for instance to mate choice (Grafen, 1990), cooperation (Leimar, 1997-08), aggression (Enquist, 1985-11-01), parent–offspring conflict (Godfray, 1991-07), and predator–prey interactions (Smith and Harper, 2003-11-06). In these models, interactions are most often dyadic, or involve one receiver and many signalers.

In contrast, the ritual behaviors we have mentioned occur in the context of an entire group. To model commitment displays, we adapt a model introduced by Gintis et al. (2001). This model is distinctive in applying to group interactions, involving many signalers and many receivers. Crucially, signalers compete for asymmetric affiliations (from here on: for followers). Optimal signaler behavior depends on the

* Corresponding author at: Institut Jean Nicod, Département d'études cognitives, Ecole normale supérieure, Université PSL, EHESS, CNRS, 75005 Paris, France.
E-mail address: jliep@protonmail.com (J. Lie-Panis).

[1] We borrow the term runaway from Fisher (1915) The mechanism we have in mind is, however, entirely different from Fisher's. In our simulation, individuals gradually invest in higher levels of signaling due to social pressures—in order to attract partners, and avoid others' outrage.

https://doi.org/10.1016/j.jtbi.2023.111586
Received 27 December 2022; Received in revised form 14 July 2023; Accepted 20 July 2023

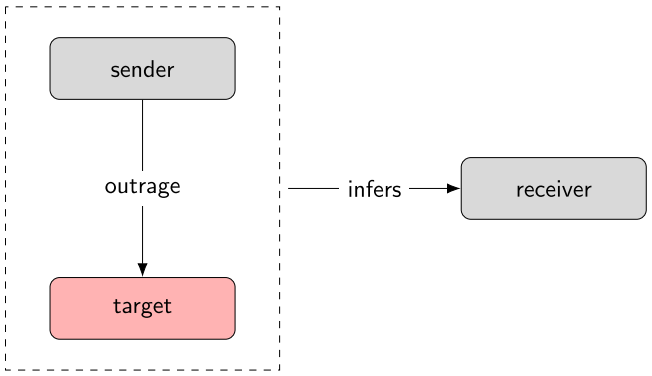


Fig. 1. Outrage as a second-order signal. A sender can express outrage at a target who does not invest in the signal. When outrage is honest, receivers can infer that the sender has invested in the signal, even without having observed the sender's behavior directly. Outrage makes the sender's signal more visible. As a side-effect, the target is harmed.

behavior of other signalers. In equilibrium, being the first to display is always beneficial, as one is able to attract many followers; in contrast, being the last to display is assumed to be net costly for individuals of low quality. As a result, a partial pooling equilibrium (Bergstrom and Lachmann, 1998) is obtained, in which individuals of lower quality opt out of the costly display entirely (Dessalles, 2014-06; Gintis et al., 2001).

This issue is exacerbated when the display is binary, as in Gintis et al.'s model (Gintis, Smith, and Bowles, 2001), and ours below. When individuals all invest in the same display, signaling is uniform, and therefore dishonest. Universal investment in a binary display should be doubly impossible. Not only should low quality signalers opt out of investing in a net costly display, but receivers should not pay attention to a entirely uninformative signal.

To explain universal investment in commitment displays, non-senders must face additional costs. We propose an endogenous source for those costs. In the type of group interaction we model, an individual's signal is susceptible to be observed by only a fraction of potential followers. Senders may be motivated to exploit non-senders, if this allows them to advertise their own signal beyond its direct observers. They may denounce or bully individuals who do not comply with a display to draw attention to their own compliance. As a result, non-senders face new costs. Universal investment could emerge out of a single motivation: advertising one's prosocial commitment, by any means necessary.

More specifically, we argue that universal displays can be propped up by moral condemnation. Moral condemnation may take various forms, ranging from negative gossip to a dyadic partner about a third-party's immoral behavior to public expressions of collective outrage. It can entail a degree of reputational and/or material costs for its target. Here, we encapsulate these differing forms of moral condemnation and the associated costs for targets using the term outrage. Outrage can be a credible signal of moral behavior. To infer the moral quality of our partners, we sometimes use their propensity to verbally condemn a third-party's immoral behavior (Jordan et al., 2017). Conversely, to advertise our investment in desirable behavior, we sometimes express outrage against those who unambiguously display undesirable behavior (Jordan and Rand, 2019); or even against those whose morality is merely ambiguous (Jordan and Kteily, 2022).

In the context of commitment displays, outrage can be thought of as a *second-order signal* — a signal about (the absence of) a signal (Fig. 1). We may for instance draw attention to those who secretly eat during a fast, and whose transgression may have otherwise gone unnoticed. In doing so, we not only broadcast our own investment, but we also indirectly increase others' incentive to display, thus laying the groundwork for universal, and even uniform, signaling.

3. Baseline model

3.1. A multi-player model of commitment displays with uncertain observation

To study commitment displays, we adapt the model introduced by Gintis et al. (2001). We consider a large group of individuals, who are characterized by a continuous quality q. We normalize minimum and maximum quality to 0 and 1 respectively: each individual's quality is drawn according to a continuous probability density function, whose support is $[0, 1]$. Individuals only observe their own quality. For mathematical convenience, the group is considered to be infinite in size.

Individuals alternate between two roles, that of signaler and receiver. Play occurs in three stages.

1. *Signaling stage*. Here, signalers decide whether to pay a cost $c_1(q)$ to send a signal, that is participate in a binary display of commitment. (The only other option is not to send.) Sending the signal is cheaper for high quality individuals: c_1 is a strictly decreasing continuous function of individual quality q, which takes positive values. In the present context, individuals of higher quality can be thought of as individuals who are more committed to the group and/or its moral values, and whose commitment translates into an increased ability or willingness to invest in the display—e.g. because the display will cause them to "burn bridges" with other groups, to which they are relatively uncommitted (Brusse, 2020).
2. *Observation stage*. Here, receivers do two things. First, they decide whether to pay a small positive cost $\nu > 0$ to monitor the signal. Second, receivers who paid the cost of monitoring observe the action chosen by each individual signaler in the previous stage, i.e. whether the signaler opted to send or not send. The probability of observation is p_1 $(0 < p_1 < 1)$. Since the population is infinite, they observe the behavior of a fraction p_1 of signalers. As long as sending occurs with positive probability, monitoring receivers each observe at least one sender (not necessarily the same one). Receivers who did not pay the cost of monitoring do not observe behavior in the signaling stage.
3. *Social interaction stage*. Here, receivers decide whether to follow one signaler, that is to affiliate to one individual from the group. Signalers gain positive payoff $s > 0$ for each receiver who decides to follow them. Receivers derive payoff $f(q')$ from following a signaler of quality q', and null payoff from opting not to follow anyone. Following is on average beneficial, and high quality individuals are more desirable social partners: we assume $\mathbf{E}(f) > 0$, and that f is a strictly increasing continuous function of the followee's quality q'. Following low quality individuals may or may not be detrimental (depending on the sign of $f(0)$).

Signalers may decide to send the signal depending on their quality. Receivers may decide to: not monitor and not follow anyone; not monitor and follow any individual, i.e. follow an individual chosen at random; monitor and follow a sender, i.e. follow an individual chosen at random among those signalers they observed sending the signal; or monitor and follow a non-sender. A pure strategy profile specifies: (i) when in the signaler role, whether to send or not send given own quality q, and (ii) when in the receiver role, which one of the above four strategies to play. We do not consider mixed strategies, in which individuals behave probabilistically.

3.2. Honest signaling equilibrium

There are two evolutionarily stable strategy profiles (ESS; Maynard Smith and Price, 1973). First, the strategy profile in which: (i) signalers never send, and (ii) receivers do not monitor, and follow an individual at random. This trivial strategy profile is always a strict Nash equilibrium, and therefore always an ESS. Indeed, deviation to sending the signal is costly for all signalers, whatever their quality. Deviation

to monitoring is costly for receivers, as is deviation to not following—because following an individual at random is beneficial ($\mathbf{E}(f) > 0$). We ignore this ESS from here on.

The second ESS is obtained by considering a family of strategy profiles. For any threshold quality $\theta \in (0, 1)$, we define the honest signaling strategy profile for θ as the strategy profile whereby: (i) signalers send when their quality q verifies $q > \theta$, and do not send when $q < \theta$, and (ii) receivers monitor, and follow a sender. We note this strategy profile HS(θ). We do not consider signaler strategy given $q = \theta$; since quality is continuously distributed, this occurs with null probability. When individuals play according to such a strategy profile, we define $\pi(\theta) \equiv \mathbf{P}(q > \theta) \in (0, 1)$ the probability that a signaler is of relatively high quality $q > \theta$, and sends.

We show that HS(θ) is an ESS if and only if:

$$\nu < \mathbf{E}(f(q) \mid q > \theta) - \mathbf{E}(f) \tag{3.1}$$

$$c_1(\theta) = \frac{s}{\pi(\theta)} \tag{3.2}$$

Below, we outline the main steps allowing us to derive both conditions. We show that (3.2) at best defines a unique value of θ, and therefore a single strategy profile to consider. The full demonstration is detailed in the Supplementary Information.

3.3. Uniform signaling is unstable

Condition (3.1) is obtained by considering the case of receivers. When signalers play according to HS(θ), receivers pay the cost of monitoring, and follow an individual of relatively high quality $q > \theta$. On average, they gain: $\mathbf{E}(f(q) \mid q > \theta) - \nu$. In contrast, a rare mutant who opts not to monitor and follow any individual can expect to gain $\mathbf{E}(f)$. We deduce the proposed condition by comparing both payoffs.

For signaling to be evolutionarily stable, the relative benefit of conditioning affiliation on the signal must outweigh the cost of monitoring. This is a relatively weak condition. Since the cost of monitoring ν is small, condition (3.1) may be satisfied even when discrimination by the signal is weak (low positive threshold θ), and even when partner quality is weakly associated to payoffs (small derivative f').

In equilibrium, the signal is honest. When they observe the signal, receivers can infer the sender is of relatively high quality, above a certain positive threshold $\theta > 0$. In contrast, uniform signaling ($\theta = 0$) is always uninformative, and can therefore never be evolutionarily stable. If all signalers send, receivers learn nothing from the signal, and mutants who do not monitor can invade.

3.4. Existence of a signaling equilibrium

Condition (3.2) is obtained by considering the case of signalers. When receivers play according to HS(θ), signalers compete to attract followers by sending the signal. Each receiver observes a fraction $p_1 \times \pi(\theta)$ of senders, and chooses one to follow.

A signaler of quality q who sends the signal pays cost $c_1(q)$ in the signaling stage. In the social interaction stage, that signaler is individually observed by each receiver with probability p_1. Each time the signaler is observed by a receiver, she is chosen with probability $\frac{1}{p_1 \times \pi(\theta)}$ (since the receiver chooses one individual at random among all observed senders), in which case she gains s. On average, she gains: $p_1 \times \frac{1}{p_1 \times \pi(\theta)} \times s$. The signaler's expected payoff is then: $-c(q) + \frac{s}{\pi(\theta)}$.

Since not sending is free, a rare mutant who deviates from HS(θ) by not sending given relatively high quality (resp. sending given relatively low quality) earns less than the resident when the above expression is positive for $q > \theta$ (resp. negative for $q < \theta$). We deduce that, for HS(θ) to be an ESS, the above expression must be null for $q = \theta$; this yields condition (3.2). In equilibrium, the signal is net beneficial for high quality Signalers, and net costly for low quality Signalers.

Condition (3.2) is an equation in θ, with at best one solution. When θ varies from 0 to 1, $c_1(\theta)$ strictly decreases from $c_1(0)$, and $\frac{s}{\pi(\theta)}$ strictly increases to infinity from $\frac{s}{1} = s$ (because the distribution's support is the entire interval $[0, 1]$). Following the intermediate value theorem we obtain a unique solution $\theta \in (0, 1)$ if and only if:

$$c_1(0) > s \tag{3.3}$$

A signaling equilibrium, defined for the unique value of $\theta \in (0, 1)$ which solves Eq. (3.2), can only exist when the cost of sending for individuals of minimal quality ($q = 0$) is prohibitively high. Conversely, condition (3.3) guarantees the existence of a signaling ESS, so long as monitoring is sufficiently cheap, as per condition (3.1) (we detail the demonstration in the Supplementary Information).

4. The signal runaway game

4.1. Adding outrage to the baseline model

The signal runaway game occurs when we introduce outrage into the previous model. We view outrage as a *second-order signal*. Outrage refers to the commitment display (the first-order signal), by referring to a target's lack of signaling. Its function is to draw attention to the fact that the outraged individual did send the signal. Outraged senders increase everyone's incentive to send, and may destabilize the honest signaling equilibrium studied above. We modify the game in the following manner.

1. *Signaling stage*. Signalers decide whether to invest in costly signaling, as before, as well as whether to pay a cost $c_2 > 0$ to express outrage.
2. *Observation stage*. By expressing outrage, individuals draw attention to their signaling behavior. Signalers who paid the cost of second-order signaling c_2 are observed with increased probability $p_2 > p_1$ ($p_2 < 1$). In our model, onlookers can only observe whether an individual sent the signal—with probability p_1 or p_2, depending on whether the individual paid to express outrage.

Outraged signalers observe the signal, and select a target. Each individual's signaling behavior is thus observed by receivers who pay the cost of monitoring, as well as signalers who pay the cost of outrage. We assume outraged signalers select a target among all individuals they observe not sending the signal. Since the population is infinite, outraged signalers find a target in all situations but uniform signaling.

3. *Social interaction stage*. Targets of outrage are harmed. Signalers lose $h > 0$ for each individual who expresses outrage against them. As before, receivers can follow signalers.

A pure strategy for the signaler now specifies whether or not to send, and, if opting to send, whether to express outrage or not, depending on own quality q. For every q, there are three possibilities: send and express outrage, send and do not express outrage, and do neither. Receiver strategies are unchanged.

4.2. A note on our assumptions

Note that we do not consider the hypocritical strategy, whereby a signaler of quality q does not send the signal, yet pays the cost of second-order signaling. Due to the simplified manner in which we model observation, this strategy is dominated. Receivers can only condition on a individual's observed signal, and not on whether the individual expresses outrage. Hypocritical signalers pay c_2 to draw attention to their lack of signaling, which is never beneficial in our model because it does not allow them to attract more followers.

Our model is intended to show the consequences of outrage functioning as a second-order signal, that is the consequences of onlookers using outrage to infer compliance (Jordan et al., 2017), for exogenous reasons. Nevertheless, we come back to the issue of hypocrisy in Section 4.7, by extending our model.

The cost of outrage is fixed, and thought of as small. It is intended to reflect the risk of retaliation by the target (and, technically, the cost of monitoring, since outraged senders need to find a target). Note that targets are always non-senders. Unjustified punishment, which can

damage one's reputation (Barclay, 2006), does not arise in our model. Instead, outraged individuals target non-senders, and earn a form of hard-coded reputational benefit, by increasing their chances of being followed when receivers value the signal itself.

Lastly, note that the cost of being outraged h is exogenous. We view h as encapsulating a variety of reputational and/or material costs which are suffered in contexts exterior to the model, ranging from the cost of being the subject of another indiviudual's negative gossip to the cost of constituting a legitimate target for collective punishment. In our simplified model, there can be no endogenous costs: targets of outrage can neither lose future followers nor attract more outrage in the future because all followees and targets of outrage are selected simultaneously, in the observation stage (besides, targets of outrage are non-senders who already do not attract any followers). The simulation presented in Section 5 implements richer dynamics, allowing for such costs. It clarifies that even when allowing for such endogeneous costs, the exogenous cost of being outraged h must be positive for uniform signaling to occur (see Fig. 3).

4.3. Honest signaling with outrage equilibrium

For any threshold quality $\theta \in (0, 1)$, we define the honest signaling with outrage strategy profile HSO(θ) as the strategy profile whereby: (i) signalers send and express outrage when their quality verifies $q > \theta$, and neither send nor express outrage when $q < \theta$, and (ii) receivers monitor, and follow a sender.

We show that HSO(θ) is an ESS if and only if:

$$v < \mathbf{E}(f(q) \mid q > \theta) - \mathbf{E}(f) \tag{4.1}$$

$$c_1(\theta) + c_2 = \frac{s}{\pi(\theta)} + \frac{\pi(\theta)h}{1 - \pi(\theta)} \tag{4.2}$$

$$c_2 < \frac{p_2 - p_1}{p_2} \times \frac{s}{\pi(\theta)} \tag{4.3}$$

The proof is analogous to the one before. Receiver strategy and trade-offs are unchanged, yielding condition (4.1), which is identical to (3.1). Each receiver observes a fraction $p_2 \times \pi(\theta)$ of senders, and chooses one to follow. An analogous calculation to the one before shows that a signaler of quality q who sends and expresses outrage gains on average payoff: $-c_1(q) - c_2 + p_2 \times \frac{s}{p_2 \pi(\theta)} = -c_1(q) - c_2 + \frac{s}{\pi(\theta)}$.

Non-senders now face the cost of being potential targets of others' outrage. Each outraged signaler observes a fraction $p_1 \times (1 - \pi(\theta))$ of non-senders, and selects one as target. Non-senders face an outraged signaler with probability $\pi(\theta)$, and are observed by that individual with probability p_1. They can now expect a negative payoff, equal to: $p_1 \times \frac{\pi(\theta)(-h)}{p_1(1-\pi(\theta))} = -\frac{\pi(\theta)h}{1-\pi(\theta)}$. We obtain condition (4.2) by comparing to the payoff above. When θ verifies this condition, signalers of quality $q = \theta$ are indifferent between sending both the signal and the second-order signal, and sending neither. Deviation to sending neither signal given $q > \theta$ is then detrimental, as is deviation to sending both signals given $q < \theta$.

Finally, condition (4.3) is obtained by considering rare mutants who deviate to sending but not expressing outrage given $q > \theta$. Such an individual saves on the cost of second-order signaling c_2, but is observed with probability $p_1 < p_2$, earning only $p_1 \times \frac{s}{p_2 \pi(\theta)}$ on average in terms of followers. Comparing to the payoff of an outraged sender, we obtain the proposed condition.

4.4. Sufficient condition for the evolution of outrage

Since $\frac{1}{\pi(\theta)} \geq 1$, we deduce a sufficient condition for (4.3), valid whatever the value of $\theta \in (0, 1)$:

$$c_2 < \frac{p_2 - p_1}{p_2} s \tag{4.4}$$

We show that outrage can be expected to invade any honest signaling equilibrium under the same condition (see Supplementary Information). Outrage evolves when the cost of second-order signaling is small relative to the benefit of making one's signal more visible to followers. Under condition (4.4), we do not need to consider the send and do not express outrage strategy, which is dominated in any honest signaling equilibrium.

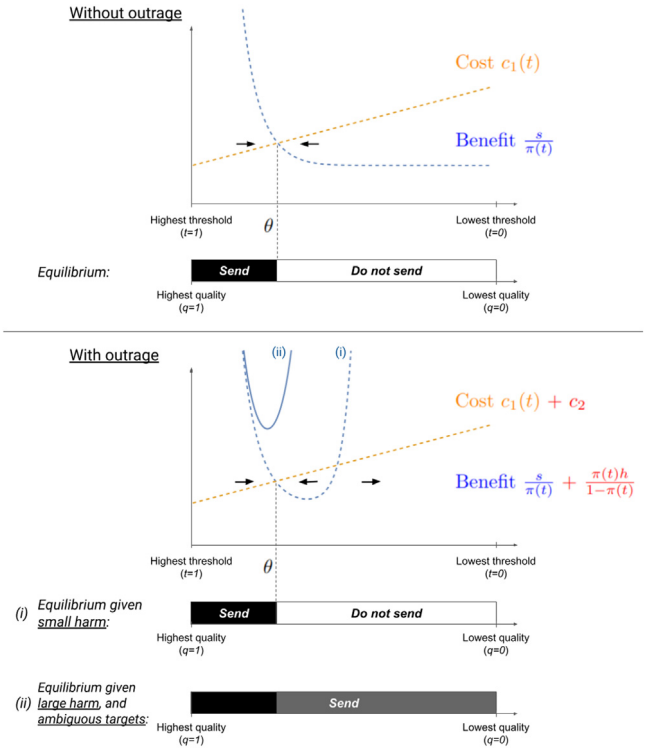


Fig. 2. Effect of outrage on the signaling equilibrium. **Top:** In the absence of outrage, senders compete to attract followers. An honest signaling equilibrium is established at the threshold quality θ which equalizes cost (orange) and benefit of competing against a fraction $\pi(\theta)$ of senders to attract followers (blue). Note that we represent cost and benefit as a function of the potential threshold t, for t varying between 1 and 0 (inverted x-axis). To the left of the graph, when t is large, few high quality individuals ($q > t$) send. Going towards the right of the graph, as t decreases, more and more lower quality individuals join in sending the signal. **Bottom**: Outrage increases the incentive to signal; senders compete to attract followers *and* evade others' outrage. Note that when t tends towards 0, the benefit of evading others' outrage tends towards infinity since individuals then risk becoming the groups' moral punching bag; this is why the blue curve now takes on a U-shape. **(i)** When harm h is low, we obtain another honest signaling equilibrium. The threshold quality θ is obtained at the first intersection of the orange and blue curves. To the left, when t is just above θ, too few high quality individuals send, and individuals whose quality is just below t benefit from joining in. To the right, when t is just below θ, there are too many senders, and those of quality just above t benefit from opting out. In contrast, the other intersection point is repellent. **(ii)** When harm is high, there is no honest signaling equilibrium. When in addition outrage can be directed at ambiguous targets, we obtain uniform signaling. For the purpose of illustration, we assume a linear cost function $c_1(q) = c_1(0) + q(c_1(1) - c_1(0))$, and that quality is normally distributed around $\bar{q} = 0.25$, with standard deviation 0.1. Other parameters: $c_1(0) = 3$, $c_1(1) = 1$, $s = 1$, $c_2 = 0.1$. In condition (i), we take $h = 0.01$; in condition (ii), we take $h = 0.1$.

4.5. Outrage can destabilize the honest signaling equilibrium

Outrage perturbs the signaling equilibrium. Senders now compete to attract followers *and* evade others' outrage. Technically outrage could lead to less signaling—when the cost of expressing outrage is larger than the expected cost of being targeted by the outrage of others ($c_2 > \frac{\pi(\theta)h}{1-\pi(\theta)}$). Since c_2 is considered to be small, outrage will most often push more individuals to send the signal.

There are two possible outcomes, represented in Fig. 2. First, when harm h is low, outrage introduces a small perturbation, and we retain a separating equilibrium. Second, when the consequences of being the subject of others' outrage are dire, outrage introduces a larger

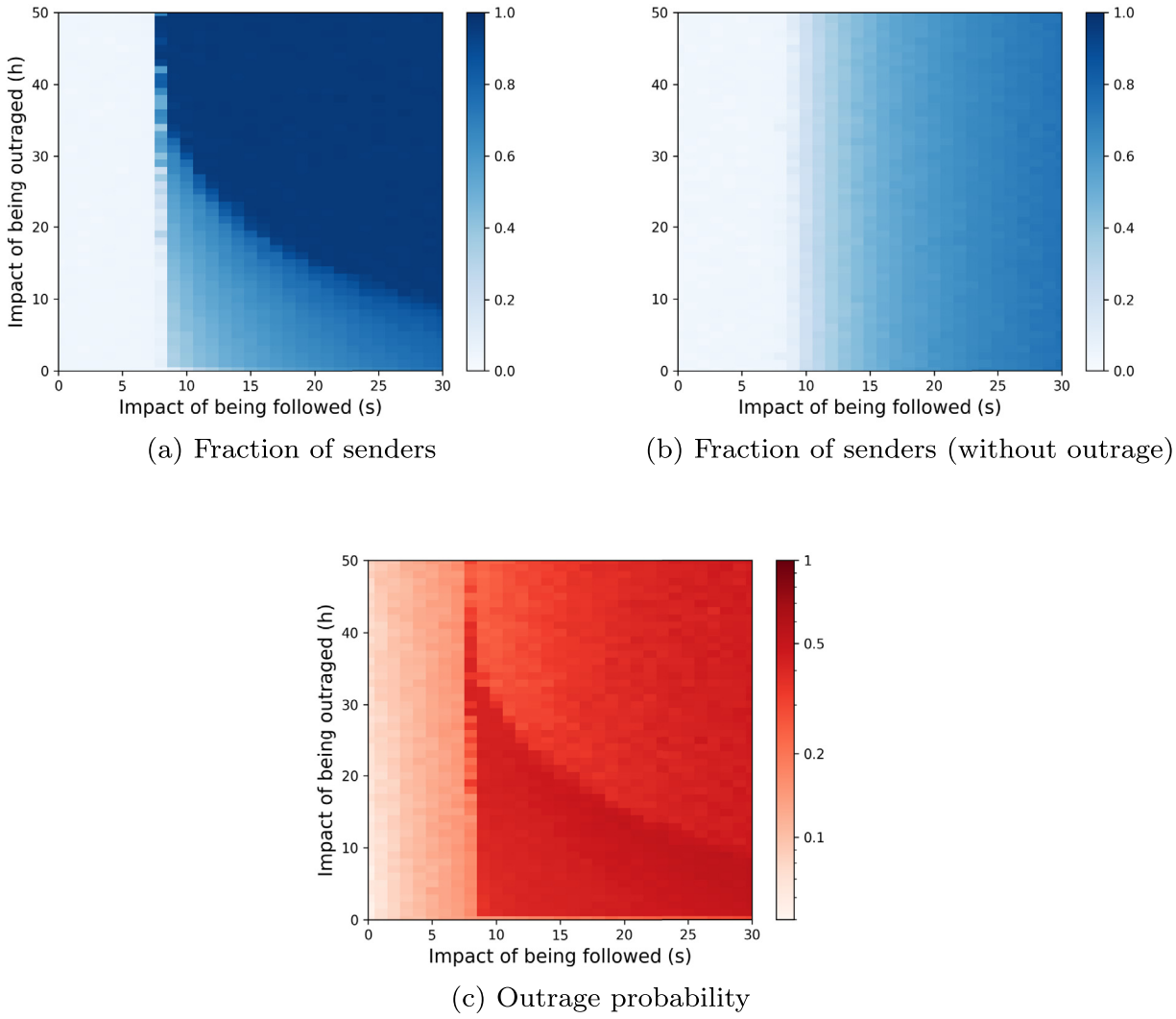


Fig. 3. Results for one level of signaling, after a large number of rounds of simulation. **Top:** fraction of agents who invest in the binary display, as a function of the benefit of being followed s and the cost of being outraged h, when agents can also invest in outrage, and (b) when they cannot. When agents can invest in outrage, signaling (blue regions) is obtained when the benefit of being followed is sufficiently large; near-uniform signaling (dark blue region) is obtained when the cost of being outraged is high. In the absence of outrage, at most 75% of individuals send. **Bottom:** outrage probability is maximal when signaling is non-uniform (light blue zone in (a)).
These simulations are computed with default values including: $h = 30$, $s = 10$, $p_1 = 0.1$, $c_1 = 30$, $c_2 = 5$; the population is composed of 200 agents; they can be affiliated with 2 other individuals and can receive up to 5 affiliation links (see Supplementary Information for more). Code and dynamic illustrations are available on this website.

perturbation—and may completely destabilize the honest signaling equilibrium. We show that, whatever the value of $\theta \in (0,1)$, HSO(θ) is not an ESS if:

$$c_1(0) + c_2 < s + 2\sqrt{hs} \tag{4.5}$$

The above condition is obtained by considering condition (4.2). Multiplying by $\pi(\theta)(1-\pi(\theta))$, we obtain equivalently:

$$(c_1(\theta) + c_2 + h)\pi(\theta)^2 - (c_1(\theta) + c_2 + s)\pi(\theta) + s = 0$$

We recognize a second-order equation in $\pi(\theta)$, whose discriminant is:

$$\Delta = (c_1(\theta) + c_2 + s)^2 - 4(c_1(\theta) + c_2 + h)s = (c_1(\theta) + c_2 - s)^2 - 4sh$$

HSO(θ) cannot be an ESS when there is no solution $\theta \in (0,1)$ to condition (4.2). A sufficient condition for that to occur is $\Delta < 0$. Since $c_1(\theta)$ increases when θ decreases, and since we necessarily have $c_1(0) + c_2 > c_1(0) > s$ (otherwise there is no signaling equilibrium to start from following condition (3.3)), we deduce that the squared term is positive when θ is sufficiently small. We can then take the squared root, and deduce a sufficient condition by replacing θ with 0; we obtain the proposed condition (4.5).

4.6. *Uniform signaling can be stable when outrage harms ambiguous targets*

Our main result is therefore negative: if outrage is sufficiently cheap to express, as per condition (4.4), and being the target of others' outrage is sufficiently costly, as per condition (4.5), then outrage invades, and fully destabilizes any honest signaling equilibrium. Under such conditions, there can be no signaling ESS. Uniform signaling remains impossible here, because the function of outrage is merely to attract more followers, and receivers stop monitoring the signal when it is uniform.

Uniform signaling can however be made possible by extending the target selection mechanism. When all individuals signal, there are no non-senders to target. In our model, for technical reasons, this does not prevent signalers from investing in second-order signaling (because the model occurs in separate stages for simplicity, and we need outraged signalers' visibility to increase before the observation stage). We may instead assume that when individuals do not find non-senders, they use more ambiguous targets instead, in order to express outrage. Although outrage at ambiguous targets is less justified, and therefore riskier (Barclay, 2006), in some contexts it is used to attract reputational benefits (Jordan and Kteily, 2022).

We modify our model, by having outraged senders select as target: (1) a non-sender whom they observe, or, if they do not observe any non-sender (because signaling is uniform) (2) a signaler whose behavior they do not observe. Second-order signaling now serves two functions. Senders who are not observed not only miss out on potential followers but also risk becoming targets of others' outrage.

We define the uniform signaling with outrage strategy profile USO as the strategy profile whereby: (i) senders signal and express outrage, whatever their quality, and (ii) receivers do not monitor, and follow an individual at random.

When individuals play according to USO, each signaler chooses an ambiguous target from the $1-p_2$ percent of individuals that they do not observe. A signaler of quality q pays both costs of signaling, and is a potential target of outrage for another individual with probability $(1-p_2)$. That signaler earns average payoff: $-c_1(q)-c_2-(1-p_2)\times\frac{h}{1-p_2} = -c_1(q)-c_2-h$.

Since the population is infinite, deviation to not sending is immediately detrimental: any signaler who attempts to save on the cost of sending risks becoming the group's moral punching bag, by constituting a preferential, unambiguous target for others' outrage. In addition, a rare mutant who deviates to not expressing outrage saves on cost c_2 but is unobserved, and therefore targeted, with increased probability $(1-p_1) > (1-p_2)$. On average, that mutant earns payoff: $-c_1(q)-(1-p_1)\times\frac{h}{1-p_2}$. Comparing with the payoff of a resident, we deduce that USO is an ESS if and only if:

$$c_2 < \frac{(p_2-p_1)h}{1-p_2} \tag{4.6}$$

4.7. Outrage is honest when it targets hypocrites first

We can further extend the target selection mechanism to cater for hypocrites. Hypocrites are oft reviled, and judged more severely than individuals who admit to engaging in immoral behavior (Jordan et al., 2017). In the context of our model, hypocrites could constitute preferential targets for outrage.

Let us assume that hypocrites are preferential targets of outrage, that is that outraged individuals select as target: (0) an observed hypocrite, i.e. an individual whom they observe expressing outrage but not sending the signal, and, if no hypocrites are observed, (1) an observed non-sender, as before (and possibly, (2) an ambiguous target if they do not observe any non-sender).

In the model up until now, receivers cannot condition on others' outrage behavior, precluding any social benefits for hypocrites. Instead, let us assume the most favorable case for hypocrisy, that is that receivers follow at random among all individuals they observe either sending the signal, or expressing outrage—so long as they do not also observe a hypocrite not sending the signal. We assume that the probability that outrage is observed is $p' = \frac{p_2-p_1}{1-p_1}$ (such that $p_2 = p_1 + p' - p_1p'$ is the probability that an outraged sender is observed sending either signal).

Under such conditions, HSO(θ) is immune to hypocrisy for every $\theta \in (0,1)$. Since the population is infinite, deviation to not sending and expressing outrage is immediately detrimental, because this entails becoming a preferential target for a positive fraction of the population, and losing infinite payoff.

In addition, we can derive a sufficient condition given any positive fraction π_H of hypocrites. Let us assume that, for a certain threshold $\theta \in (0,1)$, signalers whose quality exceeds θ send and express outrage; and signalers whose quality is under θ never send, and express outrage with probability $\frac{\pi_H}{1-\pi(\theta)}$. We assume that receivers play as described above. This situation is analogous to the HSO(θ) strategy profile, with a total fraction $\pi_H > 0$ of hypocrites.

In such a situation, receivers each observe a total of $p_2\pi(\theta) + p'(1-p_1)\pi_H$ potential followees, and chose one to follow; while outraged individuals each observe a fraction $p'p_1\pi_H$ of hypocrites, and chose one to target. On average, hypocrites earn: $-c_2 + p'(1-p_1)\frac{s}{p_2\pi(\theta)+p'(1-p_1)\pi_H} - p'p_1\frac{h}{p'p_1\pi_H}$. Since non-hypocritical non-senders earn null payoff (hypocrites concentrate outrage), a sufficient condition is obtained when, for hypocrites, the cost of facing other's outrage exceeds the benefit of attracting followers, that is (using $p'(1-p_1) = (p_2-p_1)$) when:

$$\frac{h}{\pi_H} > \frac{s}{\pi_H + \frac{p_2}{p_2-p_1}\pi(\theta)}$$

The expression on the right is always smaller than $\frac{s}{\pi_H+0}$. We deduce that, when senders express outrage and outrage is directed at hypocrites first, a sufficient condition for outrage to be honest is that the cost of being outraged exceed the benefit of attracting a follower, i.e. that:

$$h > s \tag{4.7}$$

5. Simulation

5.1. Outrage enables uniform signaling

We implement our model into an agent-based simulation. Contrary to the model, we consider a finite population, and implement local interactions. Individuals can choose a limited number of other agents to follow. The number of followers that an individual can have is also limited, in order to avoid winner-take-all effects.

Agents observe senders and non-senders directly, with probability p_1. In addition, they observe non-senders indirectly, through dyadic interactions with partners, who may express outrage at a non-sender they previously observed. Agents preferentially follow individuals whom they observe sending the signal (directly or indirectly), or whom they observe expressing outrage (during a dyadic encounter).

Agents interact based on two flexible behavioral traits: their investment in a binary display (one level of signaling), and their probability of expressing outrage at non-senders. In the initial round of simulation, these traits are set at 0. With a small probability, agents may try out another value of the trait.

Fig. 3 shows attained fraction of senders in the case of a binary display, depending on the benefit of being followed s and the cost of being outraged h. It also illustrates the crucial role of outrage in enabling uniform signaling.

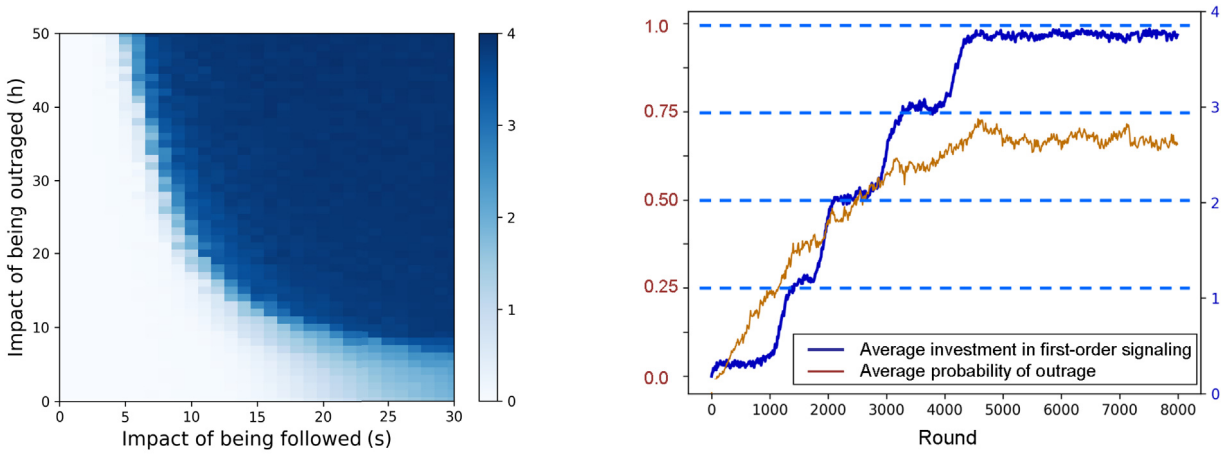
5.2. Runaway costs

When signaling becomes uniform, onlookers can no longer determine who are the top-quality individuals. To attract followers, these individuals may find it in their interest to create and adopt a new discrete signal level, requiring an additional investment of $\Delta c_1(q)$. Again, we assume Δc_1 is a decreasing function of individual quality q. Overperformers have every incentive to advertise their increased investment — e.g. by finding new targets of outrage. We assume they may now pay Δc_2 to express outrage at individuals who are observed sending at the lower level, and guarantee visibility $p_3 > p_2$; targets lose h. Similarly to before, individuals are pushed to increase their investment in the signal (they are prevented from decreasing their investment to 0 for the same reasons as before). We expect full escalation to the new signal level when:

$$\Delta c_1(0) + \Delta c_2 < s + 2\sqrt{hs} \tag{5.1}$$

Outrage could thus lead a population to adopt a costlier display. We relaunch our simulation with several evenly spaced levels of signaling (proportional costs). Agents may now express outrage at non-senders and lower-level senders (whom they still observe directly and indirectly). They preferentially follow: (i) first, an individual observed sending at level $n+1$, (ii) second, an individual observed expressing outrage against a n-level sender, and (iii) third, a n-level sender.

When h and s are sufficiently large, outrage enables a step-by-step runaway process: individuals gradually learn to invest in the highest level of signaling (see Fig. 4). This is in accordance with Eq. (5.1);



(a) Average level of signaling (b) Step-by-step runaway

Fig. 4. Results for four non-null, evenly spaced levels of signaling. **(a)** Average attained level of investment. Agents learn to invest in the highest signal level as soon as h and s are significant. **(b)** Step-by-step runaway, computed with a small value of $s = 2$ and a large value of $h = 150$, to show a clean ratchet effect. A dynamic illustration can be seen on the website.

when levels are evenly spaced, the marginal cost of signaling one level above is constant from one level to the next, and signal escalation may continue indefinitely. In reality, we expect marginal costs to increase at each step to infinity, as individuals are forced to miss out on increasingly important opportunities. The process will necessarily come to a halt. Eventually, high quality individuals will not benefit from creating a costlier display (and advertising it at the expense of others), and low quality individuals will prefer not to increase their investment, even if this means appearing relatively uncommitted.

6. Discussion

This paper offers a proof of concept for the existence of uniform, high-cost displays of commitment which serve to attract followers. The model is general, and may apply to other situations in which signalers compete for followers, and signaling seems exaggerated. Tentatively, our model could apply to high engagement on online social networks, and widespread prestige-motivated help in other species (Zahavi, 1995).

Our model is agnostic about any function the emerging behavior may serve at the level of the collective (e.g. encouraging group cohesion or cooperation; Atran and Henrich, 2010; Whitehouse and Lanman, 2014; Durkheim, 2008; Xygalatas et al., 2013; Irons, 2001; Cimino, 2011; Bulbulia and Sosis, 2011; Gambetta, 2009). Uniform signals are explained at the individual level. Outrage benefits senders, by making their signal easier to spot. We show that, under certain conditions, outrage is sufficient to generate uniform signaling, and escalating costs.

We consider signals which take discrete values. Our model applies for binary displays of commitment, and for displays which categorize individuals (e.g. into participants of a high-ordeal ritual, of a low-ordeal ritual, and non-participants; Xygalatas et al., 2013). Of course this is a simplified vision of reality (Barker et al., 2019). Rituals do not occur in isolation, and receivers may make richer inferences by considering investment in related activities, or other qualities affecting a signaler's ability to invest in a display (e.g. status, Dumas et al., 2021). For instance, by broadening the temporal scope, we can look at long-term attendance in a frequent ritual, which is a continuous metric. Individuals who attend ritual activities more frequently are on average more generous towards other group members (Ruffle and Sosis, 2006; Soler, 2012), and are perceived as such (Power, 2017; Purzycki and Arakchaa, 2013).

Nevertheless, our focus on discrete rather than continuous signals should be seen as a feature of the model, and not a bug. Though continuously-valued signals are more informative, and appear more reasonable in a variety of situations, outrage requires clear-cut comparisons. In some cases, committed individuals could design discrete displays precisely for the purpose of expressing outrage.

We made the simplifying assumption that outrage is honest, in our model and simulation. Outrage is generally believed to be honest when hypocrites suffer sufficient retaliatory costs; yet retaliation against hypocrites is subject to much variation (Sommers and Jordan, 2022). In an extension of our model, we show that honesty can arise when hypocrites are preferential targets for others' outrage. Further research should investigate more systematically the conditions under which outrage is more likely to be honest, and/or treated as such by onlookers, ensuring that it can function as a second-order signal.

If we broaden the picture, second-order signaling can be seen as a specific case of signal amplification. The idea of amplifiers has been introduced to designate signals whose correlation with quality is indirect (Hasson, 1991). For instance, contour lines that accentuate margins on bird feathers or bars across feathers may have evolved as secondary features that make the primary signal, in this case healthy undamaged feathers, more conspicuous. Amplifiers may explain the sophistication of some mating signals. Contrary to signals, some amplifiers may not need to be costly to be reliable, as it is not in the interest of low-quality individuals to draw attention to their poor signaling (Gualla et al., 2008). In our model, outrage serves a function which is analogous to that of an amplifier: it increases the probability that the sender's signal will be detected. Outrage is a rather specific type of amplifier however, as it imposes costs on its target—through which uniform investment and runaway costs may emerge.

Our model may help explain mandatory displays of commitment, such as rites of passage (see also: Cimino, 2011; Densley, 2012; Gambetta, 2009; Iannaccone, 1992). Outrage can create a positive feedback loop, and sustain uniform, and therefore uninformative, displays. The resulting behavior is a specific type of norm. In general, norms can emerge from a variety of positive feedback loops, such as those created by social punishment or benchmark effects (Young, 2015). In our case, uniform displays arise endogenously, from the motivation to advertise one's prosocial commitment to group members, via first- and second-order signaling (we do not need to assume non-senders are punished).

Our model may also help explain exaggerated displays of commitment, e.g. during wartime (see also: Whitehouse, 2018; Sosis et al., 2007). Times of crisis tend to favor expression of commitment over others (Hahl et al., 2018), and may provide the initial push enabling

signal runaway. In such cases, the system is expected to stop at extreme levels of signaling and outrage, pushing individuals to ever greater lengths to avoid appearing uncommitted. A similar logic may be at play with witch hunts or other collective crazes which follow a seemingly self-fulfilling pattern (Lotto, 1994).

The present model is kept minimal. It needs to be completed to explain why many uniform signals remain stable without reaching extreme values, or why, and when, they may deescalate. Depending on the context, individuals may look for commitment to other groups or values. Signals and non-signals can change meaning (e.g. pacifism instead of cowardice, or closed-mindedness instead of dedication to the group). We hope that our model can serve as a basis for investigation into these rich phenomena.

CRediT authorship contribution statement

Julien Lie-Panis: Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing. **Jean-Louis Dessalles:** Conceptualization, Software, Validation, Resources, Writing – review & editing, Supervision.

Declaration of competing interest

None.

Acknowledgments

We thank A. Sijilmassi for feedback on a early version of the manuscript, as well as two anonymous reviewers for their numerous and constructive comments. This research was supported by funding from the EURIP Graduate School for Interdisciplinary Research, and from the Agence Nationale pour la Recherche (ANR-17-EURE-0017, ANR-10-IDEX-0001-02).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jtbi.2023.111586.

References

Atran, S., Henrich, J., 2010. The Evolution of Religion: How Cognitive By-Products, Adaptive Learning Heuristics, Ritual Displays, and Group Competition Generate Deep Commitments to Prosocial Religions. Biol. Theory 5 (1), 18–30. http://dx.doi.org/10.1162/BIOT_a_00018.

Barclay, P., 2006. Reputational benefits for altruistic punishment. Evol. Hum. Behav. 27 (5), 325–344. http://dx.doi.org/10.1016/j.evolhumbehav.2006.01.003, URL https://www.sciencedirect.com/science/article/pii/S1090513806000055.

Barker, J.L., Power, E.A., Heap, S., Puurtinen, M., Sosis, R., 2019. Content, cost, and context: A framework for understanding human signaling systems. Evolut. Anthropol. Issues News Rev. 28 (2), 86–99. http://dx.doi.org/10.1002/evan.21768.

Bergstrom, C.T., Lachmann, M., 1998. Signalling among relatives. II. Beyond the Tower of Babel. (ISSN: 0040-5809) URL https://digital.lib.washington.edu:443/researchworks/handle/1773/2012.

Brusse, C., 2020. Signaling theories of religion: Models and explanation. Religion Brain Behav. 10 (3), 272–291. http://dx.doi.org/10.1080/2153599X.2019.1678514.

Bulbulia, J., Sosis, R., 2011. Signalling theory and the evolution of religious cooperation. Religion 41 (3), 363–388. http://dx.doi.org/10.1080/0048721X.2011.604508.

Cimino, A., 2011. The Evolution of Hazing: Motivational Mechanisms and the Abuse of Newcomers. J. Cognit. Culture 11 (3–4), 241–267. http://dx.doi.org/10.1163/156853711X591242.

Densley, J.A., 2012. Street Gang Recruitment: Signaling, Screening, and Selection. Soc. Probl. 59 (3), 301–321. http://dx.doi.org/10.1525/sp.2012.59.3.301.

Dessalles, J.-L., 2014-06. Optimal investment in social signals. Evolution 68 (6), 1640–1650. http://dx.doi.org/10.1111/evo.12378, URL http://doi.wiley.com/10.1111/evo.12378.

Dumas, M., Barker, J.L., Power, E.A., 2021. When does reputation Lie? Dynamic feedbacks between costly signals, social capital and social prominence. Philos. Trans. R. Soc. B 376 (1838), 20200298. http://dx.doi.org/10.1098/rstb.2020.0298, URL https://royalsocietypublishing.org/doi/full/10.1098/rstb.2020.0298.

Durkheim, E., 2008. The Elementary Forms of Religious Life, Abridged edition Oxford University Press,

Enquist, M., 1985-11-01. Communication during aggressive interactions with particular reference to variation in choice of behaviour. Anim. Behav. 33 (4), 1152–1161. http://dx.doi.org/10.1016/S0003-3472(85)80175-5, URL https://www.sciencedirect.com/science/article/pii/S0003347285801755.

Fisher, R.A., 1915. The evolution of sexual preference. Eugenics Rev. 7 (3), 184.

Gambetta, D., 2009. Codes of the Underworld: how Criminals Communicate. Princeton University Press.

Gintis, H., Smith, E.A., Bowles, S., 2001. Costly Signaling and Cooperation. J. Theoret. Biol. 213 (1), 103–119. http://dx.doi.org/10.1006/jtbi.2001.2406.

Godfray, H.C.J., 1991-07. Signalling of need by offspring to their parents. Nature 352 (6333), 328–330. http://dx.doi.org/10.1038/352328a0, URL https://www.nature.com/articles/352328a0.

Grafen, A., 1990. Biological signals as handicaps. J. Theor. Biol. 144 (4), 517–546. http://dx.doi.org/10.1016/S0022-5193(05)80088-8.

Gualla, F., Cermelli, P., Castellano, S., 2008. Is there a role for amplifiers in sexual selection? J. Theoret. Biol. 252 (2), 255–271.

Hahl, O., Kim, M., Zuckerman Sivan, E.W., 2018. The Authentic Appeal of the Lying Demagogue: Proclaiming the Deeper Truth about Political Illegitimacy. Am. Sociol. Rev. 83 (1), 1–33. http://dx.doi.org/10.1177/0003122417749632.

Hasson, O., 1991. Sexual displays as amplifiers: practical examples with an emphasis on feather decorations. Behav. Ecol. 2, 189–197, URL https://academic.oup.com/beheco/article{-}pdf/2/3/189/446074/2-3-189.pdf.

Iannaccone, L.R., 1992. Sacrifice and Stigma: Reducing Free-riding in Cults, Communes, and Other Collectives. J. Polit. Econ. 100 (2), 271–291. http://dx.doi.org/10.1086/261818.

Irons, W., 2001. Religion as a hard-to-fake sign of commitment. In: Nesse, R.M. (Ed.), Evolution and the Capacity for Commitment.

Jordan, J.J., Kteily, N.S., 2022. People punish moral transgressions for reputational gain, even when they personally question whether punishment is merited. Available at PsyArXiv.

Jordan, J.J., Rand, D.G., 2019. Signaling when no one is watching: A reputation heuristics account of outrage and punishment in one-shot anonymous interactions. J. Personal. Soc. Psychol. http://dx.doi.org/10.1037/pspi0000186.

Jordan, J.J., Sommers, R., Bloom, P., Rand, D.G., 2017. Why Do We Hate Hypocrites? Evidence for a Theory of False Signaling. Psychol. Sci. 28 (3), 356–368. http://dx.doi.org/10.1177/0956797616685771.

Leimar, O., 1997-08. Reciprocity and communication of partner quality. Proc. R. Soc. B 264 (1385), 1209–1215. http://dx.doi.org/10.1098/rspb.1997.0167, URL http://www.royalsocietypublishing.org/doi/10.1098/rspb.1997.0167.

Lotto, D., 1994. On Witches and Witch Hunts: Ritual and Satanic Cult Abuse. J. Psych. New York 21 (4), 373–396.

Maynard Smith, J., Price, G.R., 1973. The Logic of Animal Conflict. Nature 246 (5427), 15–18. http://dx.doi.org/10.1038/246015a0.

Power, E.A., 2017. Discerning devotion: Testing the signaling theory of religion. Evol. Hum. Behav. 38 (1), 82–91. http://dx.doi.org/10.1016/j.evolhumbehav.2016.07.003.

Purzycki, B.G., Arakchaa, T., 2013. Ritual Behavior and Trust in the Tyva Republic. Curr. Anthropol. 54 (3), 381–388. http://dx.doi.org/10.1086/670526.

Ruffle, B.J., Sosis, R., 2006. Cooperation and the in-group-out-group bias: A field test on Israeli Kibbutz members and city residents. J. Econ. Behav. Organ. 60 (2), 147–163. http://dx.doi.org/10.1016/j.jebo.2004.07.007.

Smith, J.M., Harper, D., 2003-11-06. Animal Signals. OUP Oxford, arXiv:SUA51MeG1lcC.

Soler, M., 2012. Costly signaling, ritual and cooperation: Evidence from Candomblé, an Afro-Brazilian religion. Evol. Hum. Behav. 33 (4), 346–356. http://dx.doi.org/10.1016/j.evolhumbehav.2011.11.004.

Sommers, R., Jordan, J., 2022. When does moral engagement risk triggering a hypocrisy penalty?. http://dx.doi.org/10.31234/osf.io/w23ec.

Sosis, R., 2003. Why aren't we all hutterites?: Costly signaling theory and religious behavior. Hum. Nat. 14 (2), 91–127. http://dx.doi.org/10.1007/s12110-003-1000-6.

Sosis, R., Kress, H.C., Boster, J.S., 2007. Scars for war: Evaluating alternative signaling explanations for cross-cultural variance in ritual costs. Evol. Hum. Behav. 28 (4), 234–247. http://dx.doi.org/10.1016/j.evolhumbehav.2007.02.007.

Spence, M., 1974-03-01. Competitive and optimal responses to signals: An analysis of efficiency and distribution. J. Econom. Theory 7 (3), 296–332. http://dx.doi.org/10.1016/0022-0531(74)90098-2, URL http://www.sciencedirect.com/science/article/pii/0022053174900982.

Veblen, T., 1973. The Theory of the Leisure Class: with an Introd. By John Kenneth Galbraith. Houghton Mifflin.

Whitehouse, H., 2018. Dying for the group: Towards a general theory of extreme self-sacrifice. Behav. Brain Sci. 41, http://dx.doi.org/10.1017/S0140525X18000249.

Whitehouse, H., Lanman, J.A., 2014. The Ties That Bind Us: Ritual, Fusion, and Identification. Curr. Anthropol. 55 (6), 674–695. http://dx.doi.org/10.1086/678698.

Xygalatas, D., Mitkidis, P., Fischer, R., Reddish, P., Skewes, J., Geertz, A.W., Roepstorff, A., Bulbulia, J., 2013. Extreme Rituals Promote Prosociality. Psychol. Sci. 24 (8), 1602–1605. http://dx.doi.org/10.1177/0956797612472910.

Young, H.P., 2015. The Evolution of Social Norms. Annu. Rev. Econ. 7 (1), 359–387. http://dx.doi.org/10.1146/annurev-economics-080614-115322.

Zahavi, A., 1975-09. Mate selection—a selection for a handicap. J. Theoret. Biol. 53 (1), 205–214. http://dx.doi.org/10.1016/0022-5193(75)90111-3, URL http://linkinghub.elsevier.com/retrieve/pii/0022519375901113.

Zahavi, Z., 1995. Altruism as a Handicap: The Limitations of Kin Selection and Reciprocity. J. Avian Biol. 26 (1), 1. http://dx.doi.org/10.2307/3677205.

Supplementary Information for

Runaway signals:

Exaggerated displays of commitment may result from second-order signaling

Contents

S1 Baseline model **1**
- S1.1 A multi-player model of costly signaling with uncertain observation 1
- S1.2 Honest signaling equilibrium 2
- S1.3 Interpretation 5

S2 Runaway signal game **6**
- S2.1 Adding outrage as a second-order signal 6
- S2.2 Effect of outrage on the previous signaling equilibrium 6
- S2.3 Uniform signaling can be stable when outrage harms ambiguous targets 8

S3 Simulation **10**
- S3.1 Presentation of the simulation 10
- S3.2 Differences between model and simulation 13
- S3.3 Parameters 13
- S3.4 One signal level 14

S1 Baseline model

S1.1 A multi-player model of costly signaling with uncertain observation

We adapt Gintis, Smith and Bowles' (2001) multi-player model, by making two small changes: importantly, we assume uncertain observation (with a certain probability $p_1 < 1$); more incidentally, we consider a continuous distribution of quality.

We consider a large group of individuals, who are characterized by a continuous quality q. We normalize minimum and maximum quality to 0 and 1 respectively: each individual's quality is drawn according to a continuous probability density function, whose support is $[0, 1]$. Individuals only observe their own quality. For mathematical convenience, the group is considered to be infinite in size.

Individuals alternate between two roles, that of signaler and receiver. Play occurs in three stages.

1. *Signaling stage.* Here, signalers decide whether to pay a cost $c_1(q)$ to send a signal, that is participate in a binary display of commitment. (The only other option is not to send.) Sending the signal is cheaper for high quality individuals: c_1 is a strictly decreasing continuous function of individual quality q, which takes positive values.

2. *Observation stage.* Here, receivers do two things. First, they decide whether to pay a small positive cost $\nu > 0$ to monitor the signal. Second, receivers who paid the cost of monitoring observe the action chosen by each individual signaler in the previous stage, i.e. whether the signaler opted to send or not send. The probability of observation is p_1 ($0 < p_1 < 1$). Since the population is infinite, they observe the behavior of a fraction p_1 of signalers. As long as sending occurs with positive probability, monitoring receivers each observe at least one sender (not necessarily the same one). Receivers who did not pay the cost of monitoring do not observe behavior in the signaling stage.

3. *Social interaction stage.* Here, receivers decide whether to follow one signaler, that is to affiliate to one individual from the group. Signalers gain positive payoff $s > 0$ for each receiver who decides to follow them. Receivers derive payoff $f(q')$ from following a signaler of quality q', and null payoff from opting not to follow anyone. Following is on average beneficial, and high quality individuals are more desirable social partners: we assume $\mathbf{E}(f) > 0$, and that f is a strictly increasing continuous function of the followee's quality q'. Following low quality individuals may or may not be detrimental (depending on the signal of $f(0)$).

Signalers may decide to send the signal depending on their quality. Receivers may decide to: not monitor and not follow anyone; not monitor and follow any individual, i.e. follow an individual chosen at random; monitor and follow a sender, i.e. follow an individual chosen at random among those signalers they observed sending the signal; or monitor and follow a non-sender. A pure strategy profile specifies: (i) when in the signaler role, whether to send or not send given own quality q, and (ii) when in the receiver role, which one of the above four strategies to play. We do not consider mixed strategies, in which individuals behave probabilistically.

S1.2 Honest signaling equilibrium

S1.2.1 Honest signaling strategy profile

There are two evolutionarily stable strategy profiles (ESS; Maynard Smith & Price, 1973). First, the strategy profile in which: (i) signalers never send, and (ii) receivers do not monitor, and follow an individual at random. This trivial strategy profile is always a strict Nash equilibrium, and therefore always an ESS. Indeed, deviation to sending the signal is costly for all signalers, whatever their quality. Deviation to monitoring is costly for receivers, as is deviation to not following—because following an individual at random is beneficial ($\mathbf{E}(f) > 0$). We ignore this ESS from here on.

The second ESS is obtained by considering a family of strategy profiles. For

any threshold quality $\theta \in (0,1)$, we define the honest signaling strategy profile for θ as the strategy profile whereby: (i) signalers send when their quality q verifies $q > \theta$, and do not send when $q < \theta$, and (ii) receivers monitor, and follow a sender. We note this strategy profile $\text{HS}(\theta)$. We do not consider signaler strategy given $q = \theta$; since quality is continuously distributed, this occurs with null probability. When individuals play according to such a strategy profile, we define $\pi(\theta) \equiv \mathbf{P}(q > \theta) \in (0,1)$ the probability that a signaler is of relatively high quality $q > \theta$, and sends.

Any pure strategy equilibrium where signaling occurs with positive probability must follow this form. Indeed, note first that if receivers do not monitor the signal, signalers strictly lose from signaling, whatever their quality: signaling can only occur when senders positively affect their chances of being accepted, i.e. when receivers play according to (ii). Note second that θ must belong to $(0,1)$: if it is equal to 1, then signaling occurs with null probability; and if it is equal to 0, receivers strictly benefit from deviation to not monitoring.

The below demonstration further shows that signalers must play according to a threshold reaction norm of this form. We show that there can be only one honest signaling equilibrium, corresponding to a specific value of θ, and second, that this equilibrium exists under a wide range of parameter values.

S1.2.2 Characteristics of the honest signaling equilibrium

Proposition 1 *HS(θ) is an ESS if and only if:*

$$\pi(\theta) = \frac{s}{c_1(\theta)} \tag{S1.1}$$

$$\nu < \mathbf{E}(f(q) \mid q > \theta) - \mathbf{E}(a) \tag{S1.2}$$

Proof: let us assume that individuals play according to the strategy profile $\text{HS}(\theta)$, for a given value of $\theta \in (0,1)$. We first show that $\text{HS}(\theta)$ defines a strict Nash equilibrium if and only if both of the above conditions are verified.

$\text{HS}(\theta)$ is strict Nash if and only signalers of relatively high quality $q_H > \theta$, signalers of relatively low quality $q_L < \theta$, and receivers all stand to lose from deviation. We obtain equation [S1.1] by considering the case of signalers first. A signaler of quality q can pay $c_1(q)$ to send, in which case she will face a fraction p_1 of well-disposed receivers in the future, who chose one individual to follow among the fraction $p_1 \times \pi(\theta)$ of the population that they observe sending the signal, their chosen followee earning s. Dividing the fraction of well-disposed receivers by the fraction of signals they chose from, we deduce that a sender on average recruits fraction $\frac{1}{\pi(\theta)}$ of receivers, and obtains an expected payoff of $-c_1(q) + \frac{s}{\pi(\theta)}$.

Signalers who do not send earn null payoff. By comparing the above expression to 0, we deduce that signalers of relatively high quality $q_H > \theta$ stand to lose from deviation iff $c_1(q_H) > \frac{s}{\pi(\theta)}$, and that signalers of relatively low quality $q_L < \theta$ stand to lose from deviation iff $c_1(q_L) < \frac{s}{\pi(\theta)}$. Since c_1 is a strictly decreasing function of quality, these two conditions are verified for all $q_H > \theta > q_L$ if and only if $c_1(\theta) = \frac{s}{\pi(\theta)}$; re-arranging, we obtain equation [S1.1]. (Note that signalers may send or not send indifferently when their quality q is precisely equal to the threshold θ; since this occurs with null probability, we neglect this possibility).

We obtain equation [S1.2] by considering next the case of receivers. A receiver pays ν to monitor the signal, and, since the population is infinite, is certain to observe at least one signal, and ally with a signaler of relatively high quality; earning $\mathbf{E}(f(q) \mid q > \theta) - \nu$ on average. If she deviates to accepting at random, she gains instead $\mathbf{E}(f) > 0$; if she deviates to rejecting, she gains null payoff. By comparing these payoffs, we deduce that receivers can expect to lose from deviation if and only if condition [S1.2] is verified.

We have proven that HS(θ) is strict Nash if and only if conditions [S1.1-S1.2] are verified. Hence, under these conditions, the strategy profile is an ESS. Conversely, we show that when these conditions are not verified, HS(θ) is not an ESS: if θ is different to the critical quality determined by condition [S1.1], the previous reasoning shows that the strategy profile cannot be Nash, and therefore cannot be an ESS; and if the second condition [S1.2] is unverified, it can be invaded by a strategy profile in which receivers do not monitor and accept at random. This proves the proposed equivalence.

S1.2.3 Existence of an honest signaling equilibrium

When satisfied, condition [S1.1] defines a unique critical quality θ. Condition [S1.2] adds a constraint on θ: the critical quality must be high enough to guarantee that the net gain from allying with a sender instead of an individual at random exceeds the cost of monitoring.

Gintis et al. show that when signaling is overly costly for low quality individuals ($c_1(0) > s$), an honest signaling equilibrium exists for a range of possibly binary distributions of quality. Below, we extend this result to continuous distributions of quality. When $c_1(0) > s$, an honest signaling equilibrium exists for a wide family of continuous distributions, depending on ν (those for which [S1.2] will be satisfied).

Proposition 2 *When the signal is overly costly for the lowest quality signalers, there exists a range of possible values for the cost of monitoring* $(0, \hat{\nu})$ *for which an honest signaling equilibrium can be defined. In particular, there exists an honest signaling equilibrium where the cost of monitoring is arbitrarily small if and only if:*

$$c_1(0) > s \tag{S1.3}$$

Proof: when t varies in $[0, 1]$, $\pi(t)$ strictly decreases from 1, and $\frac{s}{c_1(t)}$ strictly increases from $\frac{s}{c_1(0)}$. Following the intermediate value theorem, a non-trivial critical quality $\theta \in (0, 1)$ which satisfies condition [S1.1] can be found if and only if condition [S1.3] is verified (Figure S1 gives a graphic argument). In addition, condition [S1.2] is verified if and only if the cost of monitoring is smaller than:

$$\hat{\nu} = \mathbf{E}(f(q)|q > \theta) - \mathbf{E}(f).$$

$\hat{\nu}$ is positive since θ is greater than the minimum quality 0. Condition [S1.2] is verified whenever the cost ν of monitoring is smaller than $\hat{\nu}$.

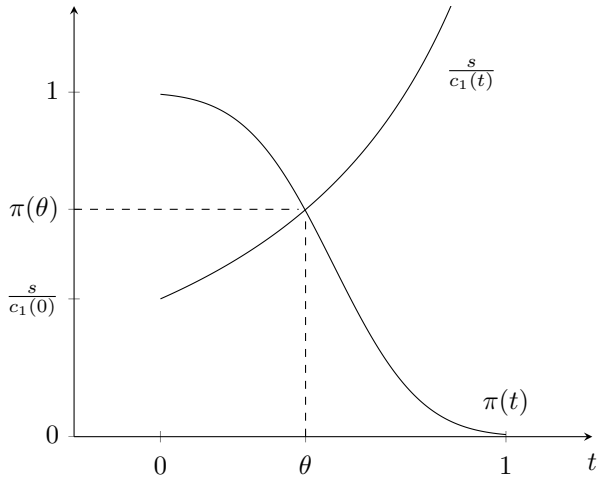


Fig. S1: Graphic determination of the critical threshold θ

S1.3 Interpretation

S1.3.1 To evolve, a signal cannot be overly widespread

Following equation [S1.2], signaling can only be evolutionary stable when the relative benefit of conditioning alliance on the signal outweighs the cost of monitoring. In equilibrium, the signal is informative: when they observe the signal, receivers can infer the sender is of relatively high quality $q > \theta > 0$. More widespread signals (lower minimum bar θ) are less informative to receivers, and less likely to evolve (depending on the cost of monitoring). In particular, a universal signal ($\theta = 0$) is always uninformative, and can never be evolutionarily stable (even when monitoring is free).

S1.3.2 In equilibrium, desirable individuals signal and obtain a net benefit

Following equation [S1.1], the equilibrium value of the threshold quality θ is the value which balances cost $c_1(\theta)$ and benefit $\frac{s}{\pi(\theta)}$ of signaling. In equilibrium, desirable individuals of quality $q > \theta$ signal, and obtain a net benefit. When θ tends towards maximum quality 1, the benefit of signaling tends towards infinity: we can always expect signaling to emerge in the presence of a large motivated audience, since the first individuals to send will gain a large following.

When in contrast θ tends towards 0, the benefit of signaling falls to s. For signaling to remain informative, joining in with everyone else must be prohibitively costly for minimum quality individuals, i.e. we must have $c_1(0) > s$. Proposition 2 shows there is a form of equivalence; signals which are prohibitively costly for minimum quality individuals can evolve as long as monitoring is sufficiently cheap.

S2 Runaway signal game

S2.1 Adding outrage as a second-order signal

The signal runaway game occurs when we introduce outrage into the previous model. We view outrage as a *second-order signal*. Outrage refers to the commitment display (the first-order signal), by referring to a target's lack of signaling. Its function is to draw attention to the fact that the outraged individual did send the signal. Outraged senders increase everyone's incentive to send, and may destabilize the honest signaling equilibrium studied above. We modify the game in the following manner.

1. *Signaling stage.* Signalers decide whether to invest in costly signaling, as before, as well as whether to pay a cost $c_2 > 0$ to express outrage.

2. *Observation stage.* By expressing outrage, individuals draw attention to their signaling behavior. Signalers who paid the cost of second-order signaling c_2 are observed with increased probability $p_2 > p_1$ ($p_2 < 1$). Outrage is never observed in our model. Onlookers can only observe whether an individual sent the signal—with probability p_1 or p_2, depending on whether the individual paid to express outrage.

Outraged signalers observe the signal, and select a target. Each individual's signaling behavior is thus observed by receivers who pay the cost of monitoring, as well as signalers who pay the cost of outrage. We assume outraged signalers select a target among all individuals they observe not sending the signal. Since the population is infinite, outraged signalers find a target in all situations but uniform signaling.

3. *Social interaction stage.* Targets of outrage are harmed. Signalers lose $h > 0$ for each individual who expresses outrage against them. As before, receivers can follow signalers.

A pure strategy for the signaler now specifies whether or not to send, and, if opting to send, whether to express outrage or not, depending on own quality q. For every q, there are three possibilities: send and express outrage, send and do not express outrage, and do neither. Receiver strategies are unchanged.

Note that we do not consider the hypocritical strategy, whereby a signaler of quality q does not send the signal, yet pays the cost of second-order signaling. Due to the simplified manner in which we model observation, this strategy is dominated. Hypocritical signalers pay c_2 to draw attention to their lack of signaling, which is never beneficial in our model because receivers do not observe outrage.

S2.2 Effect of outrage on the previous signaling equilibrium

S2.2.1 Honest signaling with outrage equilibrium

For any threshold quality $\theta \in (0, 1)$, we define the honest signaling with outrage strategy profile $\text{HSO}(\theta)$ as the strategy profile whereby: (i) signalers send and

express outrage when their quality verifies $q > \theta$, and neither send nor express outrage when $q < \theta$, and (ii) receivers monitor, and follow a sender.

Proposition 3 *HSO(θ) is an ESS if and only if:*

$$c_1(\theta) + c_2 = \frac{s}{\pi(\theta)} + \frac{\pi(\theta)h}{1-\pi(\theta)} \tag{S2.1}$$

$$\nu < \mathbf{E}(f(q) \mid q > \theta) - \mathbf{E}(f) \tag{S2.2}$$

$$c_2 < \frac{p_2 - p_1}{p_2} \times \frac{s}{\pi(\theta)} \tag{S2.3}$$

Proof: analogous to the proof of Proposition 1. We assume that individuals play according to HSO(θ), for a given value of $\theta \in (0,1)$. We show first that HSO(θ) defines a strict Nash equilibrium if and only if all three of the above conditions are verified.

Receiver strategy and trade-offs are unchanged, yielding condition (S2.2), which is identical to (S1.2). Each receiver observes a fraction $p_2 \times \pi(\theta)$ of senders, and chooses one to follow. An analogous calculation to the one before shows that a signaler of quality q who sends and expresses outrage gains on average payoff: $-c_1(q) - c_2 + p_2 \times \frac{s}{p_2\pi(\theta)} = -c_1(q) - c_2 + \frac{s}{\pi(\theta)}$.

Non-senders now face the cost of being potential targets of others' outrage. Each outraged signaler observes a fraction $p_1 \times (1 - \pi(\theta))$ of non-senders, and selects one as target. Non-senders face an outraged signaler with probability $\pi(\theta)$, and are observed by that individual with probability p_1. They can now expect a negative payoff, equal to: $p_1 \times \frac{\pi(\theta)(-h)}{p_1(1-\pi(\theta))} = -\frac{\pi(\theta)h}{1-\pi(\theta)}$. We obtain condition (S2.1) by comparing to the payoff above. When θ verifies this condition, signalers of quality $q = \theta$ are indifferent between sending both the signal and the second-order signal, and sending neither. Deviation to sending neither signal given $q > \theta$ is then detrimental, as is deviation to sending both signals given $q < \theta$.

Finally, condition (S2.3) is obtained by considering rare mutants who deviate to sending but not expressing outrage given $q > \theta$. Such an individual saves on the cost of second-order signaling c_2, but is observed with probability $p_1 < p_2$, earning only $p_1 \times \frac{s}{p_2\pi(\theta)}$ on average in terms of followers. Comparing to the payoff of an outraged sender, we obtain the proposed condition.

To conclude, we have proven that HSO(θ) is strict Nash if and only if conditions [S2.1-S2.3] are verified. Under these conditions, the strategy profile is an ESS. Conversely, we can show that both of the first two conditions are necessary, using an analogous argument to the one in Proposition 1. In addition, the last condition is necessary because otherwise rare mutants who deviate to sending but not expressing outrage given $q > \theta$ could invade. This proves the proposed equivalence.

S2.2.2 Sufficient condition for the evolution of outrage when signaling is honest

Proposition 4 *When receivers follow conditionally on the signal, senders all invest in outrage if:*

$$c_2 < \frac{p_2 - p_1}{p_2} s \tag{S2.4}$$

Proof: since $\frac{1}{\pi(\theta)} \geq 1$, the above constitutes a sufficient condition for (S2.3), that is a sufficient conditions for senders to lose from deviation to not expressing outrage when individuals play according to HSO(θ), for a certain $\theta \in (0,1)$.

The only other family of pure strategies in which receivers follow conditionally on the signal are strategies that are akin to the baseline honest signaling equilibrium, whereby senders opt not to express outrage. In such a situation, there exists $\theta \in (0,1)$ such that senders are observed with probability p_1, and gain $\frac{s}{p_1 \pi(\theta)}$ when observed. Deviation to expressing outrage costs c_2 and increases one's visibility, leading to relative benefit $(p_2 - p_1)\frac{s}{p_1 \pi(\theta)} > (p_2 - p_1)\frac{s}{p_2}$. When the above condition holds, that deviation is net beneficial, and outrage invades.

S2.2.3 Condition under which $HSO(\theta)$ cannot be an ESS

Condition [S2.1] captures the effect of outrage on the equilibrium value of θ. When $c_2 < \frac{\pi(\theta)h}{1-\pi(\theta)}$, we obtain a lower threshold than in the baseline case. Outrage then increases the incentive to signal, pushing more individuals to send both signals. Under certain conditions, the minimum bar θ will be pushed all the way to 0. When this occurs, honest signaling can no longer be stable. The below proposition gives a sufficient condition for this to happen.

Proposition 5 *For every positive threshold θ, HSO(θ) is not an ESS if:*

$$c_1(0) + c_2 < s + 2\sqrt{hs} \tag{S2.5}$$

Proof: For HSO(θ) to be an equilibrium, $\pi_S = \pi(\theta)$ must verify equation [S2.1]. Multiplying by $\pi_S(1-\pi_S)$ (π_S is always positive and smaller than 1 at such an equilibrium), we obtain equivalently:

$$(c_1(\theta) + c_2 + h)\pi_S^2 - (c_1(\theta) + c_2 + s)\pi_S + s = 0$$

We recognize a second-order equation in π_S, whose discriminant is equal to:

$$\Delta = (c_1(\theta) + c_2 + s)^2 - 4(c_1(\theta) + c_2 + h)s = (c_1(\theta) + c_2 - s)^2 - 4sh$$

Outrage will push θ all the way to 0 when the above equation has no solution in the interval $(0,1)$. A sufficient condition for that to occur is $\Delta < 0$. Since $c_1(\theta)$ increases when θ decreases, and since we necessarily have $c_1(0) + c_2 > c_1(0) > s$ (otherwise there is no signaling equilibrium to start from following Proposition 2), we deduce that the squared term is positive when θ is sufficiently small. We can then take the squared root and obtain a sufficient condition by replacing θ with 0; we obtain the proposed condition.

S2.3 Uniform signaling can be stable when outrage harms ambiguous targets

S2.3.1 Extension to ambiguous secondary targets of outrage

Our main result is therefore negative: if outrage is sufficiently cheap to express, as per condition (S2.4), and being the target of others' outrage is sufficiently costly, as per condition (S2.5), then outrage invades, and fully destabilizes any

honest signaling equilibrium. Under such conditions, there can be no signaling ESS. Uniform signaling remains impossible here, because the function of outrage is merely to attract more followers, and receivers stop monitoring the signal when it is uniform.

Uniform signaling can however be made possible by extending the target selection mechanism. When all individuals signal, there are no non-senders to target. In our model, for technical reasons, this does not prevent signalers from investing in second-order signaling (because the model occurs in separate stages for simplicity, and we need outraged signalers' visibility to increase before the observation stage). We may instead assume that when individuals do not find non-senders, they use more ambiguous targets instead, in order to express outrage.

We modify our model, by having outraged senders select as target: (1) a non-sender whom they observe, or, if they do not observe any non-sender (because signaling is uniform) (2) a signaler whose behavior they do not observe. Second-order signaling now serves two functions. When senders aren't observed, they miss out on possible followers *and* risks being the target of others' outrage.

S2.3.2 Uniform signaling with outrage equilibrium

Let us consider the universal signaling with outrage (USO) strategy profile, whereby: (i) signalers send and express outrage whatever their quality, and (ii) receivers do not monitor the signal, and accept a signaler at random.

Proposition 6 *USO is an ESS if and only if:*

$$c_2 < (p_2 - p_1) \times \frac{h}{1 - p_2} \tag{S2.6}$$

Proof: let us assume individuals play according to the USO strategy profile. Since receivers do not monitor the signal, senders do not recruit more followers than non-senders. All signalers send and express outrage, by targeting one of the $1 - p_2$ individuals they each do not observe sending. With probability $1 - p_2$, a signaler will constitute a potential (ambiguous) target for another signaler; dividing, we deduce that each individual loses h, on average.

No individual benefits from deviation to not sending. Any individual who does so risks become a priority target for other individuals with probability p_1, and faces an infinite loss. If an individual opts not to express outrage, she saves on cost c_2, but increases her chance of constituting a target for others from $1 - p_2$ to $1 - p_1$, losing $\frac{1-p_1}{1-p_2}h$ on average. By comparing with h, we deduce that USO is strict Nash, and therefore ESS, if (S2.6) holds. Conversely, if this condition is unverified, senders do not lose from deviation to not expressing outrage; mutants who do not express outrage can then invade. This proves the proposed equivalency.

S2.3.3 Sufficient condition for outrage

Under the conditions derived in this section, outrage may transform the honest signaling equilibrium into a stable equilibrium where all individuals signal, and the signal is completely uninformative. When condition (S2.5) is verified, outrage should push all individuals to signal, destabilizing the honest signaling

strategy profile. As long as it is sufficiently cheap, as per condition (S2.6), we may end up with generalized signaling.

More precisely, we derive a sufficient condition for outrage to exist in all the potential situations under consideration. To simplify, we assume $\nu = 0$ in the below proposition; such that we should either be in a case of the form HSO(θ), when $\theta > 0$, and otherwise be in the case of USO.

Proposition 7 *When monitoring is free ($\nu = 0$), in any ESS where signaling occurs with positive probability, senders express outrage if:*

$$c_2 < (p_2 - p_1) \times \min\{\frac{s}{p_2}, \frac{h}{1-p_2}\} \tag{S2.7}$$

Proof: let us assume we are in an ESS where signaling occurs with positive probability, and where senders express outrage. Since the cost of sending both signals $c_1(q)+c_2$ is a decreasing function of individual quality q, signaler behavior can be described according to a threshold $\theta \in [0, 1)$ above which they send both signals.

If $\theta > 0$, we must be in the case of honest signaling with outrage. Since $\nu = 0$, receivers strictly benefit from using the signal. Let us consider a signaler of quality $q \geq \theta$, who sends both signals, and earns on average $p_2 \times \frac{s}{p_2 \pi(\theta)} - c_1(q) - c_2$. Were such an individual to deviate to not expressing outrage, she would save on the cost of outrage c_2, but decrease her chances of being observed from p_2 to p_1. On average deviation to not expressing outrage for a sender leads to payoff differential: $c_2 - (p_2 - p_1)\frac{s}{p_2 \pi(\theta)} \leq c_2 - (p_2 - p_1)\frac{s}{p_2}$. Since we are in an ESS, and since $\theta < 1$, we deduce that we must have: $c_2 < (p_2 - p_1)\frac{s}{p_2}$.

If $\theta = 0$, we must be in the case of the USO ESS, and therefore have $c_2 < (p_2 - p_1)\frac{h}{1-p_2}$, following Proposition 6. This proves the implication.

Finally, let us assume instead that players are playing according to a strategy profile in which signaling occurs with positive probability, and senders do not express outrage. We prove the strategy profile cannot be ESS when the above condition holds. To do this, note first that we must be in (a situation akin to) the baseline honest signaling equilibrium. The same steps as in the proof of Proposition 4 show that deviation to expressing outrage is net beneficial under the above condition. The strategy profile under consideration can therefore not be an ESS. This proves the proposed equivalency.

S3 Simulation

S3.1 Presentation of the simulation

The multi-agent simulation, written in Python, is based on the *Evolife*[1] platform. Agents differ by their quality. Agent qualities are uniformly distributed between 0 and 100. They may signal at a certain level at a cost that smoothly decreases with their quality. Agents learn two features through a simple local search: their investment in signaling and their probability of expressing outrage (investment in signal monitoring is an optional learned feature). A typical example of run can be see on the website.

[1] All programs are open source and are available at this Website: https://evolife.telecom-paris.fr/outrage. The program described here can be found in the *Evolife* package at `Evolife/Apps/Patriot/Patriot.py`

All interactions in the simulation are meant to be local. The population is structured in groups. Individuals meet each other in a randomized order within groups. During their first encounter (Algorithm 1), they observe each other's signal with a certain probability which depends on a global parameter called $InitialVisibility$ (parameter p_1 in the model and on a feature, $MonitoringProbability$ (ν in the model) (set to 1 by default, but that can be learned by individuals as an option).

Algorithm 1 Observe

Input: $self, Partner$
if $random() \leq self.MonitoringProbability$
and $random() \leq InitialVisibility$ **then**
  if $self.signal < Partner.signal$ **then**
    ▷▷ *self remembered as potential outrage target*
    add $(self, self.signal)$ **to** $Partner$**'s outrage memory**
  end if
  ▷▷ *self remembered as potential affiliation target*
  add $(self, self.signal)$ **to** $Partner$**'s affiliation memory**
end if

During a second randomized encounter (Algorithm 2), individuals may express outrage toward third parties. The point of outrage is to indicate that one's own signal is superior to the target's signal (this translates in the apparent signal $Target.signal + 1$ in the algorithm). Each individual learns a feature named $OutrageProbability$ and decides to be outraged accordingly.

Algorithm 2 Outrage

Input: $self$**,** $Partner$
if $random() \leq self.OutrageProbability$ **then**
  ▷▷ *self communicates outrage target*
  $Target \leftarrow$ **worst individual in self's outrage memory**
  if $Target.signal < Partner.signal$ **then**
    add $(Target, Target.signal)$ **to** $Partner$**'s outrage memory**
  end if
  add $(self, Target.signal + 1)$ **to** $Partner$**'s affiliation memory**
end if

In a third randomized encounter, individuals attempt to establish friendship based on the observed signals (Algorithm 3).

After these three rounds, payoffs are computed (Algorithm 4): individuals get rewarded for having attracted affiliates (they receive $FollowerImpact$, corresponding to parameter s in the model) and for being affiliated with high quality individuals (they receive $FollowingImpact \times Partner.Quality$ for each partner; function $a(q')$ in the model). Individuals get punished if they were the target of outrage (parameter h in the model). Agents' memory is reset after the assessment phase. However, they store payoffs and learn periodically from them. Agents have a limited lifespan and get fully reinitialized when being reborn with the same quality.

Algorithm 3 Interact

Input: $self$, $Partner$
if $Partner$ **in** $self$**'s affiliation memory then**
    $PartnerSignal \leftarrow Partner$**'s memorized signal**
else
    $PartnerSignal \leftarrow 0$
end if
if **self's affiliation set is not full**
or $PartnerSignal \geq self$**'s current worst friend's signal then**
    ⋈ *Partner becomes self's friend*
    $self.affiliate(Partner, PartnerSignal)$
end if

Algorithm 4 Assessment

Input: $self$
for F **in** $self$**'s friends do**
    ⋈ *payoff for having attracted a follower (s)*
    $F.Points \mathrel{+}\leftarrow FollowerImpact$
    ⋈ *payoff for being affiliated with F (depends on F's quality)*
    $self.Points \mathrel{+}\leftarrow FollowingImpact \times F.Quality$
end for
$self.Points \mathrel{-}\leftarrow$ **cost of signaling**
$self.Points \mathrel{-}\leftarrow OutrageCost \times self.OutrageProbability$
$self.Points \mathrel{-}\leftarrow MonitoringCost$
if $self.Outrage$ **then**
    $Target \leftarrow self$**'s outrage memory worst individual**
    ⋈ *outrage target is harmed*
    $Target.Points \mathrel{-}\leftarrow OutragePenalty$
end if
$self.resetMemory()$

S3.2 Differences between model and simulation

In the model, we consider an infinite population, such that one individual's strategy does not affect overall probabilities. In addition, receivers may monitor, observe and choose senders in a perfectly balanced way. In contrast, the simulation program is meant to implement a more realistic setting in which all interactions remain local. The population is periodically split into random groups within which interactions occur. Agents meet systematically, though in a randomized order. An agent may or may not see the partner's signal, and may or may not express outrage at some previously seen individual, in order to prove its own signaling to the partner. Due to locality and chance, there is a variance in the number of affiliates each visible sender may attract. To prevent a winner-take-all effect, we limit the number of affiliations per individual and the number of affiliates each sender may recruit.

Another divergence with the model lies in the payoff function $f(q)$ that depends on the quality of the individual with whom one gets affiliated. The role of this function in the model is to motivate agents to search for high-quality individuals to affiliate with. In the simulation, we made a simplifying assumption and hard-wire the preference for intense signals.

Another difference comes from the fact that agents do not always adopt the ideal strategy corresponding to their quality. They need time to learn their various options (sending the signal, expressing outrage) and they constantly explore alternatives with a certain probability. Despite behavioral variance due to chance and to this "learning noise", the simulation is robust, i.e. it produces similar outcomes for a wide range of parameter values.

Variance can be seen as an advantageous feature of the simulation. When all individuals end up sending the same signal, there are no obvious outrage targets. Hence the possibility introduced in the model of expressing outrage at ambiguous individuals, i.e. individuals that either do not send or were not observed while sending. By contrast, in the simulation, the constant existence of exploring individuals maintains potential outrage targets.

S3.3 Parameters

The simulation program relies on a variety of parameters. The most relevant ones are listed in table 1. Individuals get 'Follower Impact' (s in the model) for each agent that affiliates with them. The 'Signaling cost coefficient' provides the scale of signal cost: it corresponds to the the cost c_1 paid by a medium-quality Sender. 'Signaling cost decrease' controls the variation of signaling cost depending on quality ($c_1(q)$ in the model) (0: no variation; 1: linear decrease; higher values: steeper, non-linear convex decrease). 'Outrage penalty' (h in the model) is endured by individuals each time they are someone's outrage target. The parameter 'Outrage cost' implements a gradual version of model's fixed cost c_2: outraged individuals pay a cost that is proportional to this parameter and to their (learned) propensity to express outrage. 'Initial visibility' is the probability of individuals' signals to be seen during the observation round (p_1 in the model). Finally, two parameters control the learning speed. For each learned feature, the value explored next may totally change according to 'Jump probability' or locally change according to 'Additive exploration'.

Parameters' values are systematically explored in the simulations of the next

Description	*Default value*
Population size	200
Number of groups	5
Maximum number of followers (affiliates)	5
Number of affiliations (followees)	2
Impact of being followed (s)	10
Impact of outrage (h)	30
Signaling cost coefficient	100
Signaling cost decrease	5
Outrage cost (c_2)	5
Initial visibility (p_1)	0.1
Jump probability	0.05
Additive exploration	20%

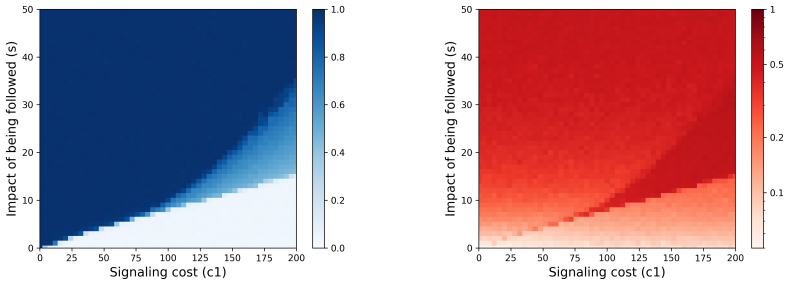
Table 1: List of most relevant parameters.

section, while non-varying parameters are set to the default values of table 1.

S3.4 One signal level

The emergence of uniform signaling due to outrage is a robust phenomenon that occurs for a wide range of parameters (see figures in the main article and dynamic examples on the website.

Figure S2a shows that three regions can be distinguished, based on costs and payoffs: a no-signal zone, a uniform signaling zone (dark blue) and an intermediary zone (light blue) corresponding to a separating equilibrium with a smaller fraction of senders. Uniform signaling (dark blue region) is obtained for low values of c_1 and high values of s. Figure S2b reveals that outrage is maximally probable in the intermediary zone, where it is used by agents as a way to increase the probability of being perceived as sender.

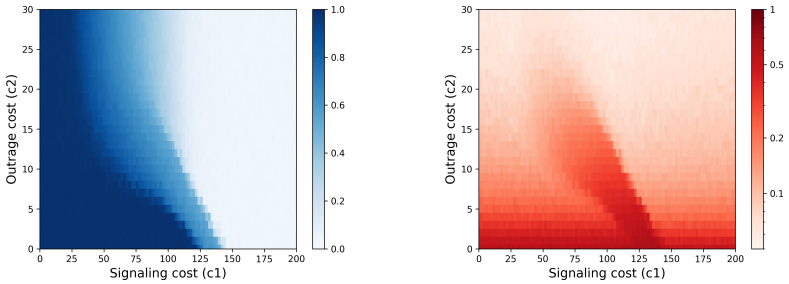


(a) Fraction of senders (b) Probability of outrage

Fig. S2: First- and second-order signal after many rounds, depending on signaling cost c_1 and follower impact s. (a) Fraction of senders; (b) average probability of outrage.

Figure S3 shows investment in both first- and second-order signaling, de-

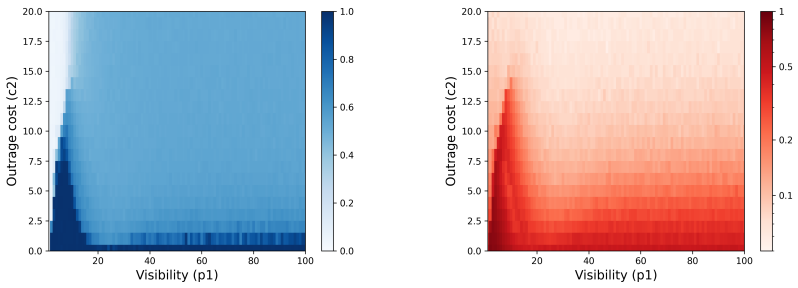
pending on signaling costs. The figure reveals the role of outrage as signal enhancer: in Fig. S3a, uniform signaling (dark blue region) expands toward costly signals at the bottom where outrage is cheap; in Fig. S3b we can see that outrage is intense in the separating equilibrium zone that corresponds to the light blue zone in Fig. S3a.



(a) Fraction of senders (b) Fraction of senders (no outrage)

Fig. S3: First- and second-order signal after many rounds, depending on the signaling cost coefficient c_1 and the cost of expressing outrage c_2. (a) Fraction of senders; (b) average probability of outrage.

Figure S4a shows that uniform signaling (dark blue) emerges when visibility (p_1) is low and outrage cost (c_2) is not too high. For other values of visibility, the separating equilibrium (light blue) is observed except when outrage is free. Figure S4b clearly shows that outrage promotes uniform signaling: outrage probability is significant in the zone that corresponds to uniform signaling and where outrage is cheap.



(a) Fraction of senders (b) Average probability of outrage

Fig. S4: Fraction of senders and average probability of outrage, as a function of visibility p_1 and outrage cost c_2.

Figure S5 shows that the emergence of a signaling situation depends on two learning parameters. The first one controls the agents' maximal additive exploration during the learning of features (here signal and outrage probability). The jump probability coefficient controls the probability of "jumping" to any value from time to time. A moderate value of either parameter is necessary for learning to function properly. Too large values generate mere noise.

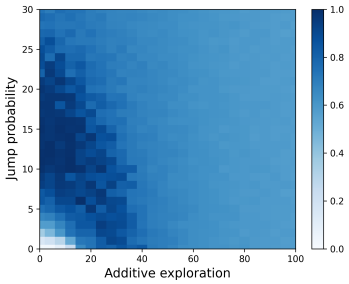
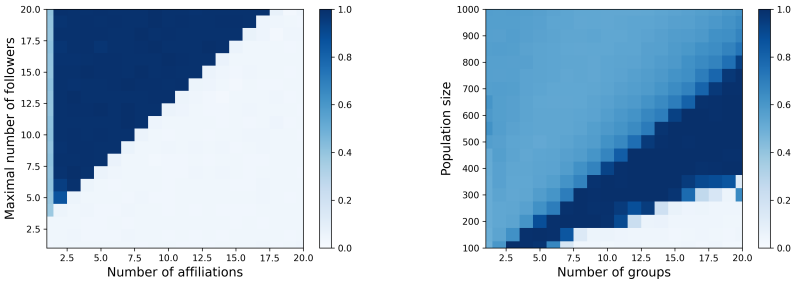


Fig. S5: Fraction of senders (blue shades) depending on additive exploration and jump probability.

Figure S6a shows the necessity of an imbalance in the number of affiliations and the number of affiliates per individual. For uniform signal to emerge in the simulation, the benefit of attracting k affiliates beyond the expected value (i.e. beyond the number of affiliations) must exceed the cost of enduring outrage (here $k \times 10 \geq 30$). Note that when individuals can have only one affiliation, the top half of them become senders and attract all available votes (hence the light-blue vertical line in figure S6a).

The population in the simulation is finite. It is structured in randomly drawn groups in which interactions occur (groups are periodically redrawn). Figure S6b shows the proportion of signalers as a function of the number of groups and the size of the population. We can observe that groups should be neither too small nor too large for signaling to emerge. In a very small group, all individuals attract the maximum number of affiliations anyway and sending the signal is useless (white region). In a large group, enough individuals are visible to each agent (up to the numnber of affiliates it can accept) and outrage becomes useless (light blue region).



(a) Affiliation imbalance

(b) Group size

Fig. S6: Fraction of senders as a function of (a) the number of given and received affiliation links and (b) the number of groups vs. the population size.

In addition, Figure S7 shows how attained investment in signaling varies with the 'Signaling cost coefficient' (variation of $c_1(q)$ in the model). It reveals that cost inequality between the low-quality (or least motivated) individuals and the high-quality (or highly motivated) individuals promotes runaway toward high-

cost signal levels.

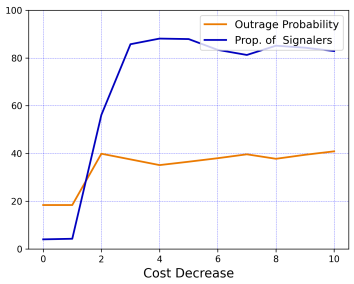


Fig. S7: Average level of signaling as a function of the 'Signaling cost decrease' coefficient.

References

Gintis, H., Smith, E. A., & Bowles, S. (2001). Costly Signaling and Cooperation. Journal of Theoretical Biology, 213(1), 103–119. https://doi.org/10.1006/jtbi.2001.2406

Maynard Smith, J., & Price, G. R. (1973). The Logic of Animal Conflict. Nature, 246(5427), 15–18. https://doi.org/10.1038/246015a0

Chapter 5

A model of endogeneous institution formation through limited reputational incentives

Objectives and summary

Across societies, humans rely on institutions to stabilize cooperation (Henrich & Muthukrishna, 2021; Powers et al., 2016). Yet institutions are not a magic bullet: even the best institutional structure on paper cannot conjure cooperation out of thin air (Bersch, 2019; McCloskey, 2016). Rather, as shown by a large body of evidence from psychology (Muthukrishna et al., 2017; Spadaro et al., 2023), economics (Beekman et al., 2014; Nannicini et al., 2013) or political science (Gutiérrez et al., 2011; Putnam et al., 1994), good functioning institutions depend on people's dedication to the common good in the first place. People must devote time and resources to design institutional rules and reward institutional agents; these agents, in turn, must resist corruption and avoid abusing their power.

In other words, institutions are *second-order cooperative interactions*—cooperative interactions aimed at promoting cooperation—which emerge from the very communities they are supposed to regulate (Ostrom, 1990). Any satisfying model of institutions should then explain both how institutions generate new incentives for collective action and how endogenous social mechanisms within the community allow the formation of institutions in the first place. Existing models, however, have proposed other mechanisms, coming from the institutions themselves, such as second-order punishment (Sigmund et al., 2010) or a compensation mechanism for individuals who take on the costs of punishment (Wang et al., 2018). Through these mechanisms, institutions are stabilized without individuals having to pay any costs—these models side-step the problem of second-order cooperation, sometimes explicitly (Currie et al., 2021). While of course these models deepen our understanding of institutions, ultimately, they are unsatisfying. If institutional rules can be designed in such an optimal manner on paper, why aren't good functioning institutions ubiquitous?

In the manuscript below, pre-printed in 2023, we build a model of endogenous institution formation, which, we argue, better captures the actual processes through which institutions solve cooperative dilemmas in human societies (for a verbal argument, see Lie-Panis and André, 2023). Our premise is that cooperative dilemmas vary in difficulty. Some are easy: existing reputational incentives are sufficient to stabilize cooperation because cooperation is cheap, behaviors are observable, or interactions occur within small groups of kith and kin. Other cooperative dilemmas are hard: they cannot be solved by reputation alone.

In our model, individuals engage in a hard form of first-order cooperation, and in an easy form of second-order cooperation—whereby they contribute to an institution that collects all individual contributions, and transforms them into incentives for first-order cooperation. Individuals who contribute to the institution are more likely to be trusted as partners for first-order cooperative interactions, thus enjoying reputational benefits. We show that reputation can indirectly stabilize first-order cooperation, by stabilizing an institution that generates enough new incentive for this hard cooperative dilemma.

Besides providing an endogenous account for institution formation, our model speaks to the cultural evolution of institutions. It suggests that institutions can be understood as technologies that humans have invented and gradually refined to build the most mutually beneficial social organization that can be sustained by reputation alone. Just as a pulley system helps lift heavy loads with minimal effort, institutions maximize the potential of limited reputational incentives, helping humans achieve extended levels of cooperation.

Our model also generates distinctive, testable predictions for the design features, social mechanisms, and cross-cultural variations of institutions. Our discussion reviews evidence for these predictions, from across the social sciences. This includes evidence that better functioning institutions emerge when social capital is high (Putnam et al., 1994), when people are intrinsically motivated to cooperate (Gächter & Schulz, 2016), and when their level of income increases (Paldam & Gundlach, 2008). This also includes evidence that institutional agents' dedication to the common good depends on reputational incentives, including in small-scale communities (Garfield et al., 2020; Ostrom, 1990).

The manuscript, printed below, is followed by a supplementary information, in which we detail the mathematical model and its results. Then, we follow by printing an accepted commentary on a BBS Target article (Glowacki, 2022).

The Mathematica file used to plot the figures can be accessed at https://osf.io/b8fy3/.

A model of endogenous institution formation through limited reputational incentives

Julien Lie-Panis*[a,b], Léo Fitouchi[a], Nicolas Baumard[a], and Jean-Baptiste André[a]

[a] *Institut Jean Nicod, Département d'études cognitives, Ecole normale supérieure, Université PSL, EHESS, CNRS, 75005 Paris, France*
[b] *LTCI, Télécom Paris, Institut Polytechnique de Paris, 91120 Palaiseau, France*

July 11, 2023

Abstract

Institutions explain humans' exceptional levels of cooperation. Yet institutions are at the mercy of the very problem they are designed to solve. They are themselves cooperative enterprises, so to say that institutions stabilize cooperation just begs the question: what stabilizes institutions? Here, we use a mathematical model to show that reputation can sustain institutions without such a second-order problem. Our premise is that cooperative dilemmas vary in difficulty. Some are easy: they can be solved by reputation alone because cooperation is cheap, behaviors are observable, or interactions occur within small groups of kith and kin. Others are hard: they cannot be solved by reputation alone. Humans need not tackle hard cooperation problems head on. Instead, they can design an institution, which (a) is based on an easy cooperation dilemma, and (b) generates enough new incentives to solve the initial hard cooperation problem. Our model leads us to view institutions as technologies that humans have invented and gradually refined to build the most mutually beneficial social organizations that can be sustained by reputation alone. Just as a pulley system helps lift heavy loads with minimal effort, institutions maximize the potential of limited reputational incentives, helping humans achieve extended levels of cooperation.

*Corresponding author. Email: jliep@protonmail.com

Large-scale cooperation is central to the success of the human species (Henrich & Muthukrishna, 2021). Yet its origins remain poorly understood. Canonical explanations, such as kin altruism (Hamilton, 1963; Ohtsuki et al., 2006), reciprocity (Axelrod & Hamilton, 1981; Barclay, 2020; Trivers, 1971), and reputation (Barclay et al., 2021; Giardini & Vilone, 2016; Lie-Panis & André, 2022; Nowak & Sigmund, 1998; Panchanathan & Boyd, 2003; Quillien, 2020), seem insufficient to explain the scale and intensity of human cooperation. In large human societies, more often than not, partners are unrelated, interactions are one-shot, and reputational information is narrowly disseminated (Lehmann et al., 2022; Powers et al., 2021).

To get around this difficulty, humans have designed institutions, such as clans (Schulz et al., 2019), age sets (Lienard, 2016), merchant guilds (Greif et al., 1994), assemblies (Hadfield & Weingast, 2013), governments (Fukuyama, 2011), and justice systems (Fitouchi & Singh, 2023; Milgrom et al., 1990; Sznycer & Patrick, 2020). These institutions make rules of good behavior explicit, specify role-specific obligations, and organize the monitoring and punishment of free-riders (Currie et al., 2021; Gavrilets & Currie, 2022). Essentially, they solve the free-rider problem by instituting new incentives for cooperation (North, 1990; Powers et al., 2016).

Institutions, however, are themselves cooperative enterprises, and as such they face a second-order free-rider problem (Yamagishi, 1986). People must devote time and resources to create new rules and pay institutional operatives. These operatives, in turn, must resist corruption; they must, for instance, rebuff bribes (Muthukrishna et al., 2017) and avoid abuses of power (Acemoglu & Robinson, 2013). In other words, institutions are a form of second-order cooperation, by which people cooperate to increase cooperation (Dixit, 2018; Ostrom, 1990; Persson et al., 2013). Saying that institutions stabilize cooperation only pushes the problem one step further: What stabilizes second-order cooperation? As long noted, stabilizing second-order cooperation seems to require third-order cooperation, which in turn seems to require fourth-order cooperation, and so on (Boyd, 2017; Boyd & Richerson, 1992).

Here, we show that reputation can stabilize institutions without such an infinite recursion. Our premise is that cooperative dilemmas vary in difficulty. Some cooperative dilemmas are hard to solve; because the temptation to cheat is high, because cheaters are unlikely to be observed, or because the dilemma involves many unrelated individuals. Other cooperative dilemmas are easy; because cooperation is cheap, behaviors are observable, and interactions occur within small groups of kith and kin.

Humans need not tackle hard cooperation problems head on. Instead, they can design another cooperative interaction that is easier to solve (e.g., because behaviors are more observable), and that generates new incentives for cooperation in the hard dilemma (e.g., by organizing the monitoring of free-riding). Institutions, we argue, consist of these easier, second-order interactions. When (a) second-order cooperation is cheap enough to be incentivized by reputation alone, and (b) the institution generates enough new incentives to solve the hard cooperation problem, cooperation is stabilized. Reputation then solves the hard

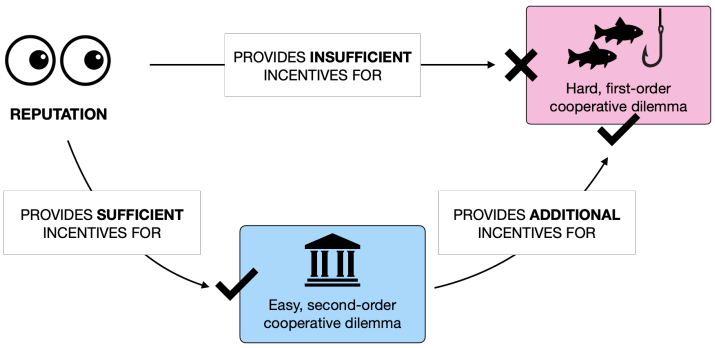


Figure 1: **Institutions allow reputation to solve hard cooperation problems indirectly**. Reputation can solve hard cooperation problems indirectly, by incentivizing an easier form of second-order cooperation, which in turn increases the incentive to cooperate at the first order. By engineering an institution based on such a form of second-order cooperation, humans engineer a technological solution to a hard cooperation problem, using only the limited reputational incentives at their disposal.

cooperation problem indirectly, by incentivizing an institution that generates new incentives for cooperation (see Figure 1).

Take a historical example. In rural Japan, villagers needed to cooperate to preserve communal forests from overuse (McKean, 1992; Ostrom, 1990, pp. 65-69). This cooperation problem was hard: it was strongly in each villager's interest to overuse the communal forest, and it was difficult to check that no one was doing so. To solve this hard problem, villages hired specialized monitors called detectives, thus generating new incentives for cooperation. This institution was itself a cooperative enterprise: for the whole thing to work, detectives had to cooperate themselves, instead of soliciting bribes, or exacting unfair penalties. Thankfully, this was a highly prestigious position. Detectives faced an easy cooperation problem: if they abused their power, they were likely to be spotted, and, thus, to lose their hard-earned reputation. Essentially, by hiring detectives, the villagers had found a way to solve their hard problem indirectly, using only the limited reputational incentives at their disposal.

Here, we formalize this idea using a mathematical model. Our model focuses on individuals who can cooperate in two different ways: sometimes they can pay to help an individual partner (first-order cooperation), and sometimes they can pay to contribute to an institution (second-order cooperation). In both cases, the only benefit they gain from cooperation is reputational. Each time individuals are observed cooperating, whether at the first- or second-order, they enhance their reputation, and become more likely to be trusted by partners in the future.

The institution collects individual contributions, and transforms them into incentives for first-order cooperation. We show that the institution extends the domain of reputation-based cooperation, to include hard cooperative dilemmas that could not be solved directly. What's more, we show that the amount of ad-

ditional cooperation generated by the institution varies with its *efficiency*—the amount of incentives the institution produces for every dollar it receives. This underscores the idea that institutions should be viewed as a social technology. Just as a pulley system helps lift heavy loads with minimal effort, institutions maximize the potential of reputational incentives, helping humans address hard cooperation problems that reputation couldn't solve directly. Institutions are social engineering tools that humans have invented and gradually refined to build the most mutually beneficial social organizations that can be sustained by reputation alone.

Model

Life of an individual actor

We consider a repeated game with two types of individuals, actors and choosers. Actors are long-run players: they play all infinite rounds of the repeated game. Choosers are short-run players: they play only one round. For mathematical convenience, actors and choosers are members of two separate populations of infinite size.

Our model focuses on actors (see Figure 2). Actors can cooperate in two ways: sometimes they can pay to help an individual chooser (first-order cooperation), and sometimes they can pay to contribute to an institution (second-order cooperation). More precisely, in each round, actors either play one trust stage game, with probability q, or they play the institution stage game, with probability $1 - q$ (from here on: one trust game, or the institution game). A trust game is played with one chooser. The institution game is played with the $1 - q$ percent of the actor population which draws that stage game in that round. Both stage games are described below.

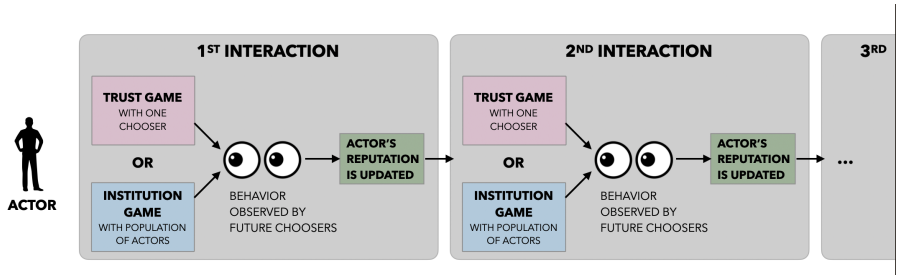


Figure 2: **Life of an individual actor**. Throughout her[1] life, an actor engages in infinitely many interactions. These interactions either involve a chooser and follow the logic of a trust game, or involve the population of actors and follow the logic of the institution game (both games are described below). After each interaction, the actor's behavior may be observed by future choosers. Her reputation is updated accordingly.

Every actor begins with an empty reputation. At the end of each round, an actor's behavior is observed by all choosers with a certain probability. The

value of this observation probability depends both on the type of game the actor played—either a trust game or the institution game—and on the incentives produced by the institution (as will be further detailed below). With the complementary probability, the actor's behavior remains invisible to choosers. We assume that choosers only have access to information from the previous round, and not from those before (memory 1). For instance, consider an actor who plays the institution game in round 1, decides to contribute, and happens to be observed by choosers. Entering round 2, her reputation is 'contributed'. If this actor then plays a trust game without being observed, choosers receive no information in that round, and the actor's reputation entering round 3 reverts to being empty.

We vary actors' ability to invest in their future reputation by varying their time preferences. Each actor is characterized by a private discount factor δ ($0 < \delta < 1$). Payoffs throughout an actor's life are calculated following a discounted utility model, whereby the present value of a payoff unit that will be received in t rounds is δ^t. When δ is high, the actor is patient. Individual values of δ are drawn at birth, depending on the population distribution of discount factors. We consider a normal distribution of mode μ and standard deviation σ, truncated over the interval $[0, 1]$ ($0 < \mu < 1$, $0 < \sigma < 1$). When μ is high, most individual actors are patient. We refer to μ as the *patience of the population*.

Trust Game (first-order cooperation)

A trust game is a two-step process. In the first step, a chooser decides whether or not to trust an actor, depending on her reputation. Trust costs $k > 0$ to the chooser, and brings reward $r > 0$ to the actor. If trusted by the chooser, the actor decides whether or not to reciprocate, in the second step. Reciprocation costs $c_1 > 0$ to the actor, and brings benefit $b > 0$ to the chooser.

We assume $b > k$. Choosers obtain a net benefit when they partner with a trustworthy actor. When trusted by their partner, actors are observed with baseline probability p_1 by future choosers ($0 < p_1 \leq 1$; actors who are not trusted do not exhibit any behavior). The probability of observation in the trust game may be increased through the action of the institution (see below).

Institution Game (second-order cooperation)

The institution game consists in a collective action involving all actors who draw that stage game. In any given round, it involves infinitely many individuals: $1-q$ percent of the infinite population of actors. Each of them decides whether or not to pay c_2 in order to contribute to the institution. Their behavior is observed by choosers with fixed probability p_2 ($0 < p_2 \leq 1$).

The institution collects actors' contributions. In a given round, we note f_2 the fraction of contributors; that is, the number of actors who decide to contribute to the institution divided by the total number of actors who face the institution game. In that round, the total amount of contributions received

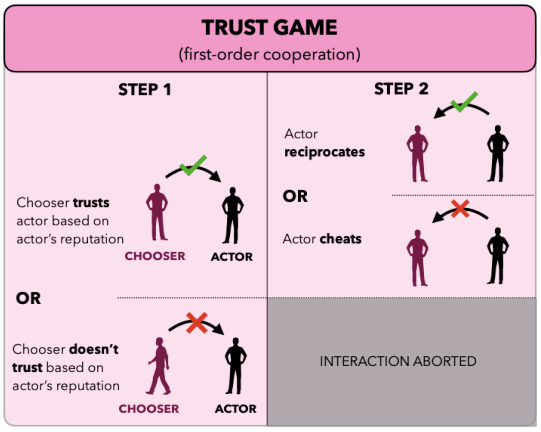


Figure 3: **Trust Game**. In a trust game, one actor interacts with one chooser. The chooser acts first: on the basis of the actor's reputation, the chooser may either trust the actor or put an early end to the interaction. If trusted, the actor may, second, either reciprocate the chooser's trust, or cheat.

by the institution is proportional to: $(1-q)f_2c_2$ (since the actor population is infinite, the total amount of contributions is infinite as well).

Mechanism of the institution

The institution transforms these contributions into incentives for first-order cooperation. One portion is allocated to rewarding cooperators, another portion is used for punishing cheaters, and the remaining portion is dedicated to monitoring. These incentives are uniformly applied to every trust game played that round; that is, the trust games played by the q percent of the actor population that interact with a chooser that round. Every actor who reciprocated a partner's trust earns reward $\beta \geq 0$, every actor who cheated is punished by $\gamma \geq 0$, and the probability of observation in every trust game is increased by $\pi_1 \geq 0$. In other words, the total amount of incentives generated by the institution is proportional to: $q(\beta + \gamma + c_1\pi_1)$ (again, this quantity is infinite). Note that we apply a factor of conversion c_1 to convert the probability increase π_1 into a dollar amount.

We define the *efficiency of the institution* ρ as the ratio between output and input; that is, the ratio between the incentives the institution generates and the contributions it receives. Mathematically:

$$\rho = \frac{\text{Incentives generated by the institution}}{\text{Contributions received by the institution}} = \frac{q(\beta + \gamma + c_1\pi_1)}{(1-q)f_2c_2} \qquad (1)$$

With this general model, we can consider different types of institutions by choosing different parameter values. For instance, a purely punishing institution is obtained by taking $\beta = \pi_1 = 0$. In that case, every dollar collected by the

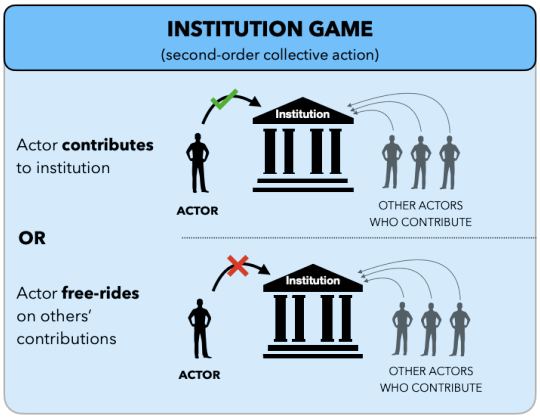


Figure 4: **Institution Game**. All actors who face the institution game in a given round take part in a collective action. They can each either contribute to the institution or free-ride on others' contributions.

institution is converted into a penalty for defectors in the trust game, who lose $\gamma = \rho f_2 c_2(1-q)/q$. A purely monitoring institution is obtained by taking $\beta = \gamma = 0$; in that case, observability in the trust game increases by $\pi_1 = \rho f_2(c_2/c_1)(1-q)/q$. Finally, a (purely) rewarding institution is obtained by taking $\gamma = \pi_1 = 0$.

Taking into account the effect of the institution, we calculate the net cost of cooperation by subtracting the total payoff of cooperators from the total payoff of defectors, and obtain: $(r-\gamma)-(r-c_1+\beta) = c_1-\beta-\gamma$. The total observability of cooperation is equal to $p_1+\pi_1$. Here, we assume that, even after accounting for the institution, first-order cooperation remains costlier and less observable than second-order cooperation, that is: $c_2 \leq c_1-\beta-\gamma$ and $p_1+\pi_1 \leq p_2$.

Results

Equilibrium analysis

We analyze our model by characterizing all possible endpoints of an evolutionary process. To do so, we use the concept of subgame perfection. A Nash equilibrium is subgame perfect when it is stable given a small likelihood of perturbing mistakes (Selten, 1983).

Baseline: cooperation in the absence of an institution

To establish a baseline, we turn off the institution, by assuming that choosers do not observe second-order cooperation ($p_2 = 0$). In such a situation, the institution is moot. Actors never contribute to the collective action, since doing so is costly and cannot lead to reputational benefits.

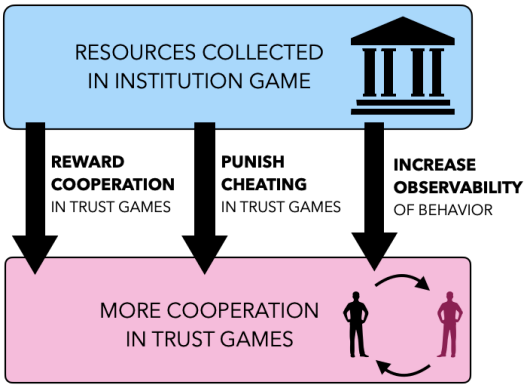


Figure 5: **Mechanism of the institution**. The institution transforms contributions made in the institution game into incentives for cooperation in trust games.

We show that there then exists a unique subgame perfect equilibrium in which cooperation occurs, which we call the baseline equilibrium. In this equilibrium, reputation incentivizes first-order cooperation only. We fully characterize the baseline equilibrium in the Methods section at the end of this document.

The baseline equilibrium is characterized by two values: the probability that choosers trust actors whose reputation is empty, and a threshold discount factor $\hat{\delta}^b$, which separates trustworthy actors from untrustworthy actors. Sufficiently future-oriented actors ($\delta \geq \hat{\delta}^b$) always reciprocate their partners' trust, and present-oriented actors ($\delta < \hat{\delta}^b$) always cheat.

In the most favorable case, the threshold discount factor is equal to: $\hat{\delta}^b = c_1/(p_1 qr)$ (in other cases, $\hat{\delta}^b > c_1/(p_1 qr)$). We refer to this minimum value as the *intrinsic difficulty of cooperation*; that is, the difficulty of cooperation in the absence of an institution. We note it δ^b (without a hat). Re-arranging, $\delta \geq \delta^b$ is equivalent to $(p_1 q) \times (\delta \times r) \geq c_1$. Actors cooperate when they can afford to pay c_1 in order to obtain r in the future with probability $p_1 q$—the probability of being observed in the current trust game *and* facing another chooser in the next interaction. When cooperation is costlier or less observable, its difficulty δ^b increases, and fewer actors are able to cooperate.

Institution equilibrium

When choosers do observe second-order cooperation ($p_2 > 0$), another subgame perfect equilibrium becomes possible. We call this equilibrium the institution equilibrium. In this equilibrium, reputation incentives both first- and second-order cooperation. As with the baseline equilibrium, we fully characterize the institution equilibrium in the Methods section at the end of this document.

The institution equilibrium is characterized by three values: the probability that choosers trust actors whose reputation is empty, and two threshold discount factors $\hat{\delta}_1$ and $\hat{\delta}_2$. These discount factors respectively separate trustworthy

actors from cheaters, and contributors from free-riders. An actor whose discount factor is δ reciprocates her partners' trust if $\delta \geq \hat{\delta}_1$ (otherwise, she cheats on them), and contributes to the institution if $\delta \geq \hat{\delta}_2$ (otherwise, she free-rides).

In the most favorable case, the threshold discount factors are equal to: $\hat{\delta}_1 = [c_1 - (\beta + \gamma)]/[(p_1 + \pi_1)q(r - \gamma)]$ and $\hat{\delta}_2 = c_2/[p_2 q(r - \gamma)]$. We note these values δ_1 and δ_2 respectively.

All types of institution lower the difficulty of cooperation: we verify that $\delta_1 < \delta^b$ whatever the balance operated between rewards, punishment and monitoring (i.e. the value given to the parameters β, γ and π_1). In addition, under our assumptions, second-order cooperation is always more difficult than first-order cooperation ($\delta_2 < \delta_1$).

Numerical resolution

To illustrate our results, we fix the institution type. We consider a monitoring-punishing institution, which allocates incentives equally between increasing the observability of cooperation and punishing defectors ($\beta = 0$, $\gamma = 1/2(\rho f_2 c_2)(1 - q)/q$, $\pi_1 = 1/2(\rho f_2 c_2/c_1)(1 - q)/q$). In the Supplementary Information, we consider other types of institution, and obtain similar results.

We consider three cases: (a) the baseline equilibrium obtained when choosers do not observe second-order cooperation ($p_2 = 0$), (b) the institution equilibrium obtained when the institution is inefficient ($\rho = 1/3$), and (c) the institution equilibrium obtained when the institution is efficient ($\rho = 3$). Figure 6 shows the rate of cooperation in each of these three cases, as a function of the patience of the population μ on the x-axis, and the intrinsic difficulty of cooperation δ^b on the y-axis.

Efficient institutions extend the domain of cooperation

In the absence of an institution, hard cooperation problems cannot be solved by reputation. On panel (a) of Figure 6, null cooperation rates are obtained as soon as the difficulty of cooperation δ^b exceeds 1.

Efficient institutions extend the domain of reputation-based cooperation, to include hard problems. On panel (c) of Figure 6, positive cooperation rates are obtained even when the difficulty of cooperation exceeds 1—in fact, even for $\delta^b = 3$. Efficient institutions allow reputation to stabilize hard cooperation problems, by amplifying its limited effects. In contrast, an inefficient institution does not make much of a dent, as visible on panel (b) of Figure 6.

Institutions are stable when the population is patient

This beneficial effect of efficient institutions is confined to large values of μ. All other things being equal, the institution equilibrium is more likely when the population is patient. Since institutions are a form of cooperation, they require that individuals pay immediate costs to invest in their long-term reputation.

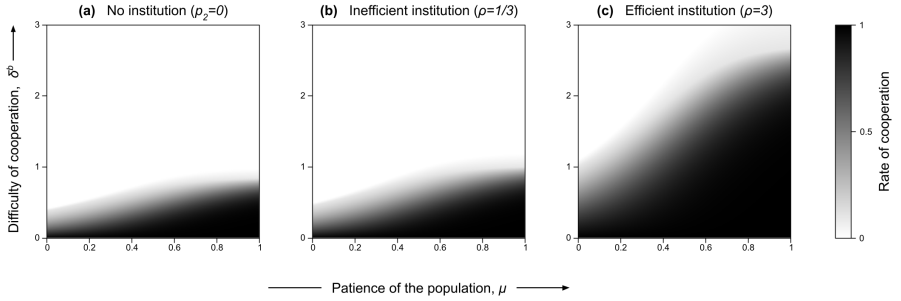


Figure 6: **Rate of cooperation**. The rate of cooperation is defined as the probability of cooperation given that a random actor and chooser interact in the first round; that is, the probability that the chooser trusts an actor whose reputation is empty and that the actor then reciprocates that trust. It is computed as a function of the patience of the population μ (x-axis), and the intrinsic difficulty of cooperation δ^b (y-axis), in three cases: **(a)** the baseline equilibrium obtained when $p_2 = 0$ (or no institution), **(b)** the institution equilibrium obtained when $\rho = 1/3$ (inefficient institution), and **(c)** the institution equilibrium obtained when $\rho = 3$ (efficient institution). The shade of gray indicates the rate of cooperation at a given point; black: maximum rate of cooperation of 1. We consider a monitoring-punishing institution ($\beta = 0$, $\gamma = 1/2(\rho f_2 c_2)(1-q)/q$, $\pi_1 = 1/2(\rho f_2 c_2/c_1)(1-q)/q$). To vary $\delta^b = c_1/(p_1 q r)$ between 0 and 3, we fix $q = 0.5$, $r = 2$, $p_1 = 1/3$, and vary c_1 between 0 and 1. We assume $c_2 = c_1/3$ and $p_2 = 1$ to ensure that second-order cooperation is always costlier and less visible than first-order cooperation. Other parameters are fixed: $\sigma = 0.25$, $k = 0.1$, $b = 1$.

Institutions are wasteful when cooperation is easy and the population is very patient

When δ^b is small in addition to μ being large, institutions become unnecessary. Large rates of cooperation can already be achieved in the non-institution equilibrium in that region. Since institutions require that individuals pay costs, they will then become wasteful.

To make this more apparent, we subtract the rate of cooperation obtained in the baseline equilibrium with $p_2 = 0$ to the rate of cooperation obtained in the institution equilibrium with $\rho = 3$, and plot the difference, in panel (a) of Figure 7. We do the same operation for the expected payoff, and plot results in panel (b). When δ^b is small and μ is large, the institution leads only to a marginal increase in the rate of cooperation. As a result, individuals are worse off.

Discussion

How do institutions for collective action develop? Unlike previous evolutionary models (Gavrilets & Duwal Shrestha, 2021; Powers & Lehmann, 2014; Sasaki et al., 2015; Schoenmakers et al., 2014; Sigmund et al., 2010; Wang et al., 2018), our model considers institutions for collective action as themselves emerging from people's cooperative behaviors (see also: Powers & Lehmann, 2013). This is consistent with a large body of evidence from psychology (Muthukrishna

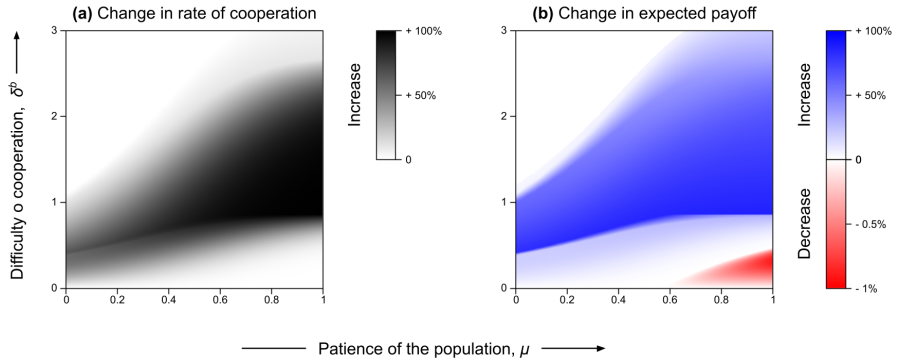


Figure 7: **Comparison between an efficient institution and no institution**. We subtract the value of **(a)** the rate of cooperation and **(b)** the expected payoff in the baseline equilibrium with $p_2 = 0$, to those same values in the institution equilibrium with $\rho = 3$. The expected payoff is defined as the normalized payoff of an individual drawn at random in both the actor and chooser populations, once many rounds of the repeated game have been played. For the precise computation, see the Supplementary Information. The rate of cooperation is defined as in Figure 6; we consider the same monitoring-punishing institution, and the same parameter values. The shade of gray indicates the increase in the rate of cooperation at a given point; black: maximum increase of 100%. Shades of blue indicate an increase in the expected payoff, and shades of red indicate a decrease. Blue: maximum increase of 100%. Red: decrease of 1%. To explain these small decreases, note that with our chosen parameters, an actor who contributes to the institution pays on average $(1-q)c_2 = c_1/6$ throughout her life. In the parameter region in which the institution appears unnecessary, c_1 is small, and $c_1/6$ is very small.

et al., 2017; Spadaro et al., 2023), economics (Beekman et al., 2014; Rose-Ackerman & Palifka, 2016), and political science (Bersch, 2019; Putnam et al., 1994). Institutions are not a magic bullet. They are made of people with their own interests, which often conflict with the common good (McCloskey, 2016; Montinola & Jackman, 2002). If these people are not motivated to pay personal costs for the benefit of their community, institutions simply fail to promote cooperation (Acemoglu & Robinson, 2013). To quote McCloskey:

> You can set up British-style courts of law, and even provide the barristers with wigs, but if the judges are venal and the barrister have no professional pride and if the public disdains them both, then the introduction of such a nice-sounding institution will fail to improve the rule of law. (McCloskey, 2016, chapter 15)

In the following, we derive distinctive predictions from our model, and show that they are supported by evidence from across the social sciences.

Institutions require social capital and intrinsic honesty

In the model, the institution relies on individual contributions. The more individuals are willing to pay to contribute to the institution, the more incentives it

can produce. Unsurprisingly, individuals who tend to bear costs to help dyadic partners (first-order cooperation) also tend to bear the costs of second-order cooperation (in our model, these are sufficiently patient individuals).

A first prediction of our model, thus, is that the effect of institutions on cooperation depends on individuals' disposition to cooperate in the first place. In a famous study, Putnam et al. (1994) showed that the best predictor of institutional performance across Italian regions was people's propensity to engage in grassroots cooperative interactions such as sports clubs, literary guilds, or choral societies. Putnam explained this association in terms of social capital; the social networks and norms of reciprocity that emerge from a long history of grassroots cooperation. The importance of social capital for institutional functioning replicates in other geographic areas and historical periods (Andrews & Brewer, 2014; Coffé & Geys, 2005; Cusack, 1999; Gutiérrez et al., 2011; Knack, 2002; Nannicini et al., 2013; Pierce et al., 2016). More recently, across 23 societies, institutional quality has been associated with people's intrinsic honesty—that is, people's propensity to cooperate even when they are not incentivized by institutions to do so (Gächter & Schulz, 2016).

Institutional honesty depends on reputational incentives

If institutional quality depends on agents' intrinsic honesty, what compels agents to be honest in the first place? In line with previous models (Jordan & Rand, 2017; Pal & Hilbe, 2022; Panchanathan & Boyd, 2004) and experimental evidence (Barclay, 2006; Dhaliwal et al., 2021; Jordan et al., 2016), our model shows that reputation can incentivize second-order cooperation. Second-order cooperators enhance their reputation, and thereby increase their chances of being rewarded by a partner's trust.

In the real world, individuals who take on an institutional role are indeed motivated by reputation and social rewards. In her famous review, Ostrom underlines how, in communities that create long-lasting institutions for common-pool resources, monitors are incentivized through reputation: "The individual who finds a rule-infractor gains status and prestige for being a good protector of the commons" (Ostrom, 1990, p.96). Similar dynamics can be found in nonindustrial societies. Among the Enga of Papua New Guinea, for example, mediators who resolve conflicts in customary courts gain a good reputation (Wiessner, 2020). Among the Amazonian Tsimane, similarly, men who mediate more conflicts are more frequently cited as cooperation partners (Glowacki & von Rueden, 2015). More largely, across nonindustrial societies, informal leaders tend to resolve conflicts on the one hand, and enjoy high status on the other (Garfield et al., 2020).

Reputation-based institutions develop in patient populations

In our model, both first- and second-order cooperation involve a present-future trade off: cooperative individuals pay to acquire a good reputation today, and

increase their chances of being trusted tomorrow (Fitouchi et al., 2022; Lie-Panis & André, 2022). As a result, more patient individuals are more likely to engage in either form of cooperation, and more patient populations are more likely to sustain an institution.

Time preferences allow us to put two stylized facts in perspective. First, they allow us to revisit the importance of social capital for institutional functioning (Putnam et al., 1994). As Putnam explains, a long history of cooperation makes social capital. It also makes the future loom large. In communities with strong social networks and norms of reciprocity, individuals can expect more from their cooperative future. With respect to their reputation, they can be characterized as patient.

Time preferences also explain why material circumstances matter. In more affluent environments, individuals' most pressing needs are met, allowing them to explore other opportunities, like investing in their reputation or social network (Boon-Falleur et al., 2022; Mell et al., 2021). Thus, all other things being equal, individuals in more affluent environments should be more patient, and more able to trust that others will also invest in their cooperative reputation. Supporting this, experimental evidence shows that political leaders are more corrupt when their voters are poor (Denly & Gautam, n.d.), and that poorer individuals more often have to pay bribes to government officials (Justesen & Bjørnskov, 2014). At the macroscopic level, a country's level of corruption is negatively associated with its wealth (Montinola & Jackman, 2002; Serra, 2006). It should be noted, however, that the relationship is bidirectional (Apergis et al., 2010; Dimant & Tosato, 2018). While economic hardship paves the way for enduring corruption (Paldam & Gundlach, 2008), corrupt institutions can also lead to economic hardship (Acemoglu & Robinson, 2013).

Social engineering and the cultural evolution of institutions

Lastly, our models speaks to the cultural evolution of institutions. A crucial parameter in our model is the institution's efficiency—the amount of incentives it produces for every dollar it receives. In the same population, more efficient institutions generate more incentives, and allow individuals to solve harder cooperation problems.

Our model leads us to view institutions as social engineering tools that humans have invented and gradually refined to build the most mutually beneficial social organizations that can be sustained by reputation alone. As we've seen, monitors are held accountable by their communities, and face reputational incentives (Ostrom, 1990); in contrast, sanctions are less legitimate, and less effective at increasing cooperation, when monitors are selected without the accord of the community (Baldassarri & Grossman, 2011). In addition, rather than assign monitoring and punishment tasks to all, people prefer to rely on specialized monitors (Traulsen et al., 2012): by doing so, they ensure that these individuals face strong reputational incentives, and an easier cooperative dilemma (Lie-Panis & André, 2023). Finally, more complex institutional arrangements are nested (Ostrom, 1990): by grouping individuals into lower level units, nested

enterprises ensure that reputation can continue to act as a strong incentive even as the number of total individuals increases (Lehmann et al., 2022).

Methods

To analyze our model, we assume that choosers can trust probabilistically an actor whose reputation is empty, and that choosers behave in a deterministic manner when faced with an actor whose reputation is non-empty, e.g., trust actors whose reputation is 'reciprocated' or 'contributed' and do not trust actors whose reputation is 'cheated' or 'free-rode'.

Throughout this section, we note θ the probability that choosers trust an actor whose reputation is empty; we establish the equilibrium value of θ in the baseline and the institution equilibria below. We allow θ to vary between 0 and 1 in order to capture all situations in which cooperation is possible. As shown below, in some cases, the equilibrium value of θ belongs to $(0, 1)$—considering only pure chooser strategies would lead to miss certain cases, and the plots shown in Figures 6 and 7 would have holes.

In the Supplementary Information, we show that our model admits three subgame perfect equilibria, when considering this chooser strategy space. In one equilibrium, which we call the trivial equilibrium, choosers never trust actors, whatever their reputation; and actors never reciprocate and never contribute, whatever their reputation and whatever their discount factor. The trivial equilibrium is always subgame perfect because trust, first- and second-order cooperation are costly. We describe the two remaining subgame perfect equilibria below.

Baseline equilibrium

The baseline equilibrium occurs when choosers trust actors whose reputation is 'reciprocated', and do not trust actors whose reputation is 'cheated', 'contributed' or 'free-rode'. Recall that we note θ the probability that choosers trust actors whose reputation is empty.

When choosers play according to this strategy, we show that in a subgame perfect equilibrium, actors reciprocate their partner's trust if and only if their discount factor δ is greater than $\hat{\delta}^b(\theta)$, whatever their current reputation. $\hat{\delta}^b(\theta)$ is given by the following equation:

$$\hat{\delta}^b(\theta) = \frac{c_1}{p_1 q(r - \theta c_1)} \tag{2}$$

Actors never contribute to the institution, whatever their reputation or discount factor. When θ varies between 0 and 1, $\hat{\delta}^b(\theta)$ strictly increases from $\delta^b = c_1/(p_1 qr)$ to $c_1/(p_1 qr(r - c_1))$. As we have defined it, the difficulty of cooperation δ^b provides a lower bound on the threshold discount factor for actors.

We show that in a subgame perfect equilibrium, θ must equal $\theta^{*,b}$, as given by the following equation:

$$\theta^{*,b} = \begin{cases} 0 & \text{if } \mathbf{P}(\delta \geq \hat{\delta}^b(0)) \leq \frac{k}{b} \\ 1 & \text{if } \mathbf{P}(\delta \geq \hat{\delta}^b(1)) \geq \frac{k}{b} \\ t & \text{such that } \mathbf{P}(\delta \geq \hat{\delta}^b(t)) = \frac{k}{b} \end{cases} \tag{3}$$

In other words, the equilibrium value of θ is 0 when the probability that an actor reciprocates is smaller than the relative cost of trust k/b *even* in the best case scenario for actors; that is, even when the threshold discount factor is at its lowest possible value δ^b. Conversely, the equilibrium value of θ is 1 when the probability that an actor reciprocates is larger than the relative cost of trust k/b *even* in the worst case scenario for actors. In all other cases, we find a unique value $0 < \theta < 1$. We allow choosers to mix given empty reputation in order to include these cases.

Taking $\theta = \theta^{*,b}$ as defined by the above equation, we show that the baseline equilibrium is in fact a subgame perfect equilibrium if and only if:

$$\hat{\delta}_1^b(\theta^{*,b}) < 1 \tag{4}$$

Note that p_2 does not appear in any of the formulas here: taking $p_2 = 0$ does not affect the form of the baseline equilibrium, nor does it affect its domain of existence. It does however negatively affect the expected payoff of actors, which is used in Figure 7.

Institution equilibrium

The institution equilibrium occurs when choosers trust actors whose reputation is 'reciprocated' and actors whose reputation is 'contributed', and do not trust actors whose reputation is 'cheated' or 'free-rode'. Again, we note θ the probability that choosers trust actors whose reputation is empty.

When choosers play according to this strategy, we show that in a subgame perfect equilibrium, actors reciprocate their partner's trust if and only if their discount factor δ is greater than $\hat{\delta}_1(\theta)$, whatever their current reputation. $\hat{\delta}_1(\theta)$ is given by the following equation:

$$\hat{\delta}_1(\theta) = \frac{c_1 - \beta - \gamma}{(p_1 + \pi_1)q[r - \gamma - \theta(c_1 - \gamma - \beta)]} \tag{5}$$

In addition, we show that in a subgame perfect equilibrium, actors contribute to the institution if and only if if their discount factor δ is greater than $\hat{\delta}_2(\theta)$, whatever their current reputation. $\hat{\delta}_2(\theta)$ is given by the following equation:

$$\hat{\delta}_2(\theta) = \frac{c_2}{p_2[q(r - \gamma) - (p_1 + \pi_1)\theta c_2]} \tag{6}$$

When θ varies between 0 and 1, $\hat{\delta}_1(\theta)$ strictly increases from δ_1 and $\hat{\delta}_2(\theta)$ strictly increases from δ_2; as we have defined them, δ_1 and δ_2 provide lower bounds on the relevant threshold discount factors.

Similarly to before, we show that in a subgame perfect equilibrium, θ must be equal to θ^*, where:

$$\theta^* = \begin{cases} 0 & \text{if } \mathbf{P}(\delta \geq \hat{\delta}_1(0)) \leq \frac{k}{b} \\ 1 & \text{if } \mathbf{P}(\delta \geq \hat{\delta}_1(1)) \geq \frac{k}{b} \\ t & \text{such that } \mathbf{P}(\delta \geq \hat{\delta}_1(t)) = \frac{k}{b} \end{cases} \tag{7}$$

Finally, taking $\theta = \theta^*$ as defined by the above equation, we show that the institution equilibrium is in fact a subgame perfect equilibrium if and only if:

$$\hat{\delta}_1(\theta^*) < 1 \tag{8}$$

$$\mathbf{P}(\delta \geq \hat{\delta}_1(\theta^*) \mid \delta \geq \hat{\delta}_2(\theta^*)) \geq \frac{k}{b} \tag{9}$$

Acknowledgements

We would like to thank Mélusine Boon-Falleur for her most excellent feedback. This research was funded by Agence Nationale pour la Recherche (ANR-17-EURE-0017, ANR-10-IDEX-0001-02).

Notes

1. We use the pronouns she/her to refer to actors throughout this document.

References

Acemoglu, D., & Robinson, J. A. (2013). *Why nations fail: The origins of power, prosperity, and poverty*. Profile Books
OCLC: 792662070.

Andrews, R., & Brewer, G. A. (2014). Social Capital and Public Service Performance: Does Managerial Strategy Matter? *Public Performance & Management Review*, *38*(2), 187–213. Retrieved May 22, 2023, from https://www.jstor.org/stable/24735250

Apergis, N., Dincer, O. C., & Payne, J. E. (2010). The relationship between corruption and income inequality in U.S. states: Evidence from a panel cointegration and error correction model. *Public Choice*, *145*(1), 125–135. https://doi.org/10.1007/s11127-009-9557-1

Axelrod, R., & Hamilton, W. D. (1981). The Evolution of Cooperation, 11.

Baldassarri, D., & Grossman, G. (2011). Centralized sanctioning and legitimate authority promote cooperation in humans. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(27), 11023–11027. https://doi.org/10.1073/pnas.1105456108

Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution and Human Behavior*, *27*(5), 325–344. https://doi.org/10.1016/j.evolhumbehav.2006.01.003

Barclay, P. (2020). Reciprocity creates a stake in one's partner, or why you should cooperate even when anonymous. *Proceedings of the Royal Society B*. https://doi.org/10.1098/rspb.2020.0819

Barclay, P., Bliege Bird, R., Roberts, G., & Számadó, S. (2021). Cooperating to show that you care: Costly helping as an honest signal of fitness interdependence. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1838), 20200292. https://doi.org/10.1098/rstb.2020.0292

Beekman, G., Bulte, E., & Nillesen, E. (2014). Corruption, investments and contributions to public goods: Experimental evidence from rural Liberia. *Journal of Public Economics*, *115*, 37–47. https://doi.org/10.1016/j.jpubeco.2014.04.004

Bersch, K. (2019, January 31). *When Democracies Deliver: Governance Reform in Latin America* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781108559638

Boon-Falleur, M., Baumard, N., & André, J.-B. (2022, June 24). Optimal resource allocation and its consequences on behavioral strategies, personality traits and preferences. https://doi.org/10.31234/osf.io/2r3ef

Boyd, R. (2017, October 23). *A Different Kind of Animal: How Culture Transformed Our Species*. Princeton University Press. https://doi.org/10.1515/9781400888528

Boyd, R., & Richerson, P. J. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology*, *13*(3), 171–195. https://doi.org/10.1016/0162-3095(92)90032-Y

Coffé, H., & Geys, B. (2005). Institutional Performance and Social Capital: An Application to the Local Government Level. *Journal of Urban Affairs*, *27*(5), 485–501. https://doi.org/10.1111/j.0735-2166.2005.00249.x

Currie, T. E., Campenni, M., Flitton, A., Njagi, T., Ontiri, E., Perret, C., & Walker, L. (2021). The cultural evolution and ecology of institutions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1828), 20200047. https://doi.org/10.1098/rstb.2020.0047

Cusack, T. R. (1999). Social capital, institutional structures, and democratic performance: A comparative study of german local governments. *European Journal of Political Research*, *35*(1), 1–34. https://doi.org/10.1111/1475-6765.00440

Denly, M., & Gautam, A. (n.d.). Poverty, Party Alignment, and Reducing Corruption through Modernization: Evidence from Guatemala, 216.

Dhaliwal, N., Patil, I., & Cushman, F. (2021). Reputational and cooperative benefits of third-party compensation. *Organizational Behavior and Human Decision Processes*, *164*, 27–51. https://doi.org/10.1016/j.obhdp.2021.01.003

Dimant, E., & Tosato, G. (2018). Causes and Effects of Corruption: What Has Past Decade's Empirical Research Taught Us? A Survey. *Journal of Economic Surveys*, *32*(2), 335–356. https://doi.org/10.1111/joes.12198

Dixit, A. (2018). Anti-corruption Institutions: Some History and Theory. In K. Basu & T. Cordella (Eds.), *Institutions, Governance and the Control*

of Corruption (pp. 15–49). Springer International Publishing. https://doi.org/10.1007/978-3-319-65684-7_2

Fitouchi, L., André, J.-B., & Baumard, N. (2022). Moral disciplining: The cognitive and evolutionary foundations of puritanical morality. *Behavioral and Brain Sciences*, 1–71. https://doi.org/10.1017/S0140525X22002047

Fitouchi, L., & Singh, M. (2023). Punitive justice serves to restore reciprocal cooperation in three small-scale societies. *Evolution and Human Behavior*. https://doi.org/10.1016/j.evolhumbehav.2023.03.001

Fukuyama, F. (2011). *The origins of political order: From prehuman times to the French Revolution* (1st ed). Profile books.

Gächter, S., & Schulz, J. F. (2016). Intrinsic honesty and the prevalence of rule violations across societies. *Nature*, *531*(7595), 496–499. https://doi.org/10.1038/nature17160

Garfield, Z. H., Syme, K. L., & Hagen, E. H. (2020). Universal and variable leadership dimensions across human societies. *Evolution and Human Behavior*, *41*(5), 397–414. https://doi.org/10.1016/j.evolhumbehav.2020.07.012

Gavrilets, S., & Currie, T. E. (2022). Mathematical models of the evolution of institutions. https://doi.org/10.31235/osf.io/kuxvd

Gavrilets, S., & Duwal Shrestha, M. (2021). Evolving institutions for collective action by selective imitation and self-interested design. *Evolution and Human Behavior*, *42*(1), 1–11. https://doi.org/10.1016/j.evolhumbehav.2020.05.007

Giardini, F., & Vilone, D. (2016). Evolution of gossip-based indirect reciprocity on a bipartite network. *Scientific Reports*, *6*(1), 37931. https://doi.org/10.1038/srep37931

Glowacki, L., & von Rueden, C. (2015). Leadership solves collective action problems in small-scale societies. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1683), 20150010. https://doi.org/10.1098/rstb.2015.0010

Greif, A., Milgrom, P., & Weingast, B. R. (1994). Coordination, Commitment, and Enforcement: The Case of the Merchant Guild. *Journal of Political Economy*, *102*(4), 745–776.

Gutiérrez, N. L., Hilborn, R., & Defeo, O. (2011). Leadership, social capital and incentives promote successful fisheries. *Nature*, *470*(7334), 386–389. https://doi.org/10.1038/nature09689

Hadfield, G. K., & Weingast, B. R. (2013). Law without the State: Legal Attributes and the Coordination of Decentralized Collective Punishment. *Journal of Law and Courts*, *1*(1), 3–34. https://doi.org/10.1086/668604

Hamilton, W. D. (1963). The Evolution of Altruistic Behavior. *The American Naturalist*, *97*(896), 354–356. https://doi.org/10.1086/497114

Henrich, J., & Muthukrishna, M. (2021). The Origins and Psychology of Human Cooperation. *Annual Review of Psychology*, *72*(1), 207–240. https://doi.org/10.1146/annurev-psych-081920-042106

Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature*, *530*(7591), 473–476. https://doi.org/10.1038/nature16981

Jordan, J. J., & Rand, D. G. (2017). Third-party punishment as a costly signal of high continuation probabilities in repeated games. *Journal of Theoretical Biology*, *421*, 189–202. https://doi.org/10.1016/j.jtbi.2017.04.004

Justesen, M. K., & Bjørnskov, C. (2014). Exploiting the Poor: Bureaucratic Corruption and Poverty in Africa. *World Development*, *58*, 106–115. https://doi.org/10.1016/j.worlddev.2014.01.002

Knack, S. (2002). Social Capital and the Quality of Government: Evidence from the States. *American Journal of Political Science*, *46*(4), 772–785. https://doi.org/10.2307/3088433

Lehmann, L., Powers, S. T., & van Schaik, C. P. (2022). Four levers of reciprocity across human societies: Concepts, analysis and predictions. *Evolutionary Human Sciences*, *4*, e11. https://doi.org/10.1017/ehs.2022.7

Lienard, P. (2016). Age Grouping and Social Complexity. *Current Anthropology*, *57*(S13), S105–S117. https://doi.org/10.1086/685685

Lie-Panis, J., & André, J.-B. (2022). Cooperation as a signal of time preferences. *Proceedings of the Royal Society B: Biological Sciences*, *289*(1973), 20212266. https://doi.org/10.1098/rspb.2021.2266

Lie-Panis, J., & André, J.-B. (2023). Peace is a form of cooperation, and so are the cultural technologies which make peace possible. https://doi.org/10.31234/osf.io/nr6ek

McCloskey, D. N. (2016). *Bourgeois equality: How ideas, not capital or institutions, enriched the world*. The University of Chicago Press.

McKean, M. (1992). Management of Traditional Common Lands in Japan. *undefined*. Retrieved June 22, 2022, from https://www.semanticscholar.org/paper/Management-of-Traditional-Common-Lands-in-Japan-McKean/07dad95f3c6390b00c083de3bf91fc973e66a3d5

Mell, H., Baumard, N., & André, J.-B. (2021). Time is money. Waiting costs explain why selection favors steeper time discounting in deprived environments. *Evolution and Human Behavior*, *42*(4), 379–387. https://doi.org/10.1016/j.evolhumbehav.2021.02.003

Milgrom, P., North, D., & Weingast, B. (1990). The Role of Institutions in the Revival of Trade: The Law Merchant, Private Judges, and the Champagne Fairs. *Economics and Politics*, *2*, 1–23. https://doi.org/10.1111/j.1468-0343.1990.tb00020.x

Montinola, G. R., & Jackman, R. W. (2002). Sources of Corruption: A Cross-Country Study. *British Journal of Political Science*, *32*(1), 147–170. https://doi.org/10.1017/S0007123402000066

Muthukrishna, M., Francois, P., Pourahmadi, S., & Henrich, J. (2017). Corrupting cooperation and how anti-corruption strategies may backfire. *Nature Human Behaviour*, *1*(7), 1–5. https://doi.org/10.1038/s41562-017-0138

Nannicini, T., Stella, A., Tabellini, G., & Troiano, U. (2013). Social Capital and Political Accountability. *American Economic Journal: Economic Policy*, *5*(2), 222–250. https://doi.org/10.1257/pol.5.2.222

North, D. C. (1990). *Institutions, institutional change, and economic performance*. Cambridge University Press.

Nowak, M. A., & Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. *Nature*, *393*(6685), 573–577. https://doi.org/10.1038/31225

Ohtsuki, H., Hauert, C., Lieberman, E., & Nowak, M. A. (2006). A simple rule for the evolution of cooperation on graphs and social networks. *Nature*, *441*(7092), 502–505. https://doi.org/10.1038/nature04605

Ostrom, E. (1990, November 30). *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press.

Pal, S., & Hilbe, C. (2022). Reputation effects drive the joint evolution of cooperation and social rewarding. *Nature Communications*, *13*(1), 5928. https://doi.org/10.1038/s41467-022-33551-y

Paldam, M., & Gundlach, E. (2008). Two Views on Institutions and Development: The Grand Transition vs the Primacy of Institutions. *Kyklos*, *61*(1), 65–100. https://doi.org/10.1111/j.1467-6435.2008.00393.x

Panchanathan, K., & Boyd, R. (2003). A tale of two defectors: The importance of standing for evolution of indirect reciprocity. *Journal of Theoretical Biology*, *224*(1), 115–126. https://doi.org/10.1016/S0022-5193(03)00154-1

Panchanathan, K., & Boyd, R. (2004). Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature*, *432*(7016), 499–502. https://doi.org/10.1038/nature02978

Persson, A., Rothstein, B., & Teorell, J. (2013). Why Anticorruption Reforms Fail—Systemic Corruption as a Collective Action Problem. *Governance*, *26*(3), 449–471. https://doi.org/10.1111/j.1468-0491.2012.01604.x

Pierce, J., Lovrich, N., & Budd, W. (2016). Social capital, institutional performance, and sustainability in Italy's regions: Still evidence of enduring historical effects? *The Social Science Journal*, *53*. https://doi.org/10.1016/j.soscij.2016.06.001

Powers, S. T., & Lehmann, L. (2013). The co-evolution of social institutions, demography, and large-scale human cooperation (M. V. Baalen, Ed.). *Ecology Letters*, *16*(11), 1356–1364. https://doi.org/10.1111/ele.12178

Powers, S. T., & Lehmann, L. (2014). An evolutionary model explaining the Neolithic transition from egalitarianism to leadership and despotism. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1791), 20141349. https://doi.org/10.1098/rspb.2014.1349

Powers, S. T., van Schaik, C. P., & Lehmann, L. (2016). How institutions shaped the last major evolutionary transition to large-scale human societies. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1687), 20150098. https://doi.org/10.1098/rstb.2015.0098

Powers, S. T., van Schaik, C. P., & Lehmann, L. (2021). Cooperation in large-scale human societies—What, if anything, makes it unique, and how did it evolve? *Evolutionary Anthropology: Issues, News, and Reviews*, *30*(4), 280–293. https://doi.org/10.1002/evan.21909
_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/evan.21909

Putnam, R. D., Leonardi, R., & Nanetti, R. (1994). *Making democracy work: Civic traditions in modern Italy* (5. print., 1. Princeton paperback print). Princeton Univ. Press.

Quillien, T. (2020). Evolution of conditional and unconditional commitment. *Journal of Theoretical Biology*, *492*, 110204. https://doi.org/10.1016/j.jtbi.2020.110204

Rose-Ackerman, S., & Palifka, B. J. (2016). *Corruption and Government: Causes, Consequences, and Reform*. Cambridge University Press.

Sasaki, T., Uchida, S., & Chen, X. (2015). Voluntary rewards mediate the evolution of pool punishment for maintaining public goods in large populations. *Scientific Reports*, *5*(1), 8917. https://doi.org/10.1038/srep08917

Schoenmakers, S., Hilbe, C., Blasius, B., & Traulsen, A. (2014). Sanctions as honest signals – The evolution of pool punishment by public sanctioning institutions. *Journal of Theoretical Biology*, *356*, 36–46. https://doi.org/10.1016/j.jtbi.2014.04.019

Schulz, J. F., Bahrami-Rad, D., Beauchamp, J. P., & Henrich, J. (2019). The Church, intensive kinship, and global psychological variation. *Science*, *366*(6466), eaau5141. https://doi.org/10.1126/science.aau5141

Selten, R. (1983). Evolutionary stability in extensive two-person games. *Mathematical Social Sciences*, *5*(3), 269–363. https://doi.org/10.1016/0165-4896(83)90012-4

Serra, D. (2006). Empirical determinants of corruption: A sensitivity analysis. *Public Choice*, *126*(1), 225–256. https://doi.org/10.1007/s11127-006-0286-4

Sigmund, K., De Silva, H., Traulsen, A., & Hauert, C. (2010). Social learning promotes institutions for governing the commons. *Nature*, *466*(7308), 861–863. https://doi.org/10.1038/nature09203

Spadaro, G., Molho, C., Van Prooijen, J.-W., Romano, A., Mosso, C. O., & Van Lange, P. A. M. (2023). Corrupt third parties undermine trust and prosocial behaviour between people. *Nature Human Behaviour*, *7*(1), 46–54. https://doi.org/10.1038/s41562-022-01457-w

Sznycer, D., & Patrick, C. (2020). The origins of criminal law. *Nature Human Behaviour*, *4*(5), 506–516. https://doi.org/10.1038/s41562-020-0827-8

Traulsen, A., Röhl, T., & Milinski, M. (2012). An economic experiment reveals that humans prefer pool punishment to maintain the commons. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1743), 3716–3721. https://doi.org/10.1098/rspb.2012.0937

Trivers, R. L. (1971). The Evolution of Reciprocal Altruism. *The Quarterly Review of Biology*, *46*(1), 35–57. https://doi.org/10.1086/406755

Wang, Q., He, N., & Chen, X. (2018). Replicator dynamics for public goods game with resource allocation in large populations. *Applied Mathematics and Computation*, *328*, 162–170. https://doi.org/10.1016/j.amc.2018.01.045

Wiessner, P. (2020). The role of third parties in norm enforcement in customary courts among the Enga of Papua New Guinea. *Proceedings of the National Academy of Sciences*, *117*(51), 32320–32328. https://doi.org/10.1073/pnas.2014759117

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology*, *51*, 110–116. https://doi.org/10.1037/0022-3514.51.1.110

Supplementary Information for:

A model of endogenous institution formation

Contents

S1 Set up **1**
- S1.1 Disclaimer . . . 1
- S1.2 Stage game . . . 2
- S1.3 Reputation of the actor . . . 3
- S1.4 Life of the actor . . . 3
- S1.5 Effect of the institution on first-order cooperation . . . 3

S2 Technical assumptions and strategy space **4**
- S2.1 History equivalence classes and strategies . . . 4
- S2.2 Three possible subgame perfect equilibria . . . 6

S3 Institution equilibrium **7**
- S3.1 Objective and simplifying notations . . . 8
- S3.2 Characterization of actor equilibrium strategy . . . 8
- S3.3 Conditions under which chooser strategy is optimal . . . 11

S4 Baseline equilibrium **12**

S5 Implementation into Mathematica **13**
- S5.1 Motivation and algorithm . . . 13
- S5.2 Other useful formula . . . 14
- S5.3 Mathematica output . . . 17

A Pdf print of the Mathematica file **22**

S1 Set up

S1.1 Disclaimer

In this document, we introduce a model of reputation-based first- and second-order cooperation. The model is presented as an infinitely repeated game involving just two players: one short-lived chooser, who plays only one round of the repeated game; and one long-lived actor, who plays all rounds. This is for the sake of mathematical simplicity; limiting our model to two players allows us to evade certain technical issues (like having to define the game history for infinitely many players), and to mobilize the framework developed by Mailath and Samuelson (2006).

Note that, to avoid lengthy repetitions of the terms chooser and actor, we have assigned a gender to each player based on the result of a coin toss; throughout this text, we will use masculine pronouns to refer to the chooser (he/him/his), and feminine pronouns (she/her) to refer to the actor.

Though our model features only two players, we have designed it with an infinite population in mind. The actor represents an infinite population of individuals, who can cooperate in two different manners. Each round of the repeated game sees certain actors interacting with one chooser, in a subgame of the stage game we call the trust game. If trusted by their partner, these actors can then pay to help them (first-order cooperation). Each round of the repeated games sees other actors interacting with one another in a collective action, in a subgame of the stage game we call the institution game. These actors can pay to contribute to an institution (second-order cooperation). The institution collects individual contributions from that second group of actors, and transforms them into incentives for first-order cooperation by the first group; that is, the institution uses those contributions to change the rules of the trust game which governs all actor-chooser interactions, pushing more actors to pay to help their partner.

As is more classic, the short-lived chooser represents a succession of potential partners for the actor. In the two-player model detailed below, everything is as if the actor faces a new chooser each time the trust game is draw. In the infinite-player model we have in mind, the chooser population provides a succession of different partners for each actor. The role of the chooser is to motivate cooperation. Each time the actors pays the cost of first- or second-order cooperation, her reputation is enhanced, and her chances of being trusted by the next potential chooser increase—as long as the chooser uses the actor's reputation to determine whether to trust her (cooperation by the actor depends on the chooser's strategy).

We begin, in the below section, by describing all the fundamental assumptions on which our model relies, in relatively little detail. In section S2, we go into further technical detail, and show how our assumptions restrict the set of possible strategies, and possible subgame perfect equilibria. We find three such equilibria, and detail the mathematical steps leading to our result in sections S3 and S4. Finally, we motivate and explain the numerical resolution of our model in section S5.

S1.2 Stage game

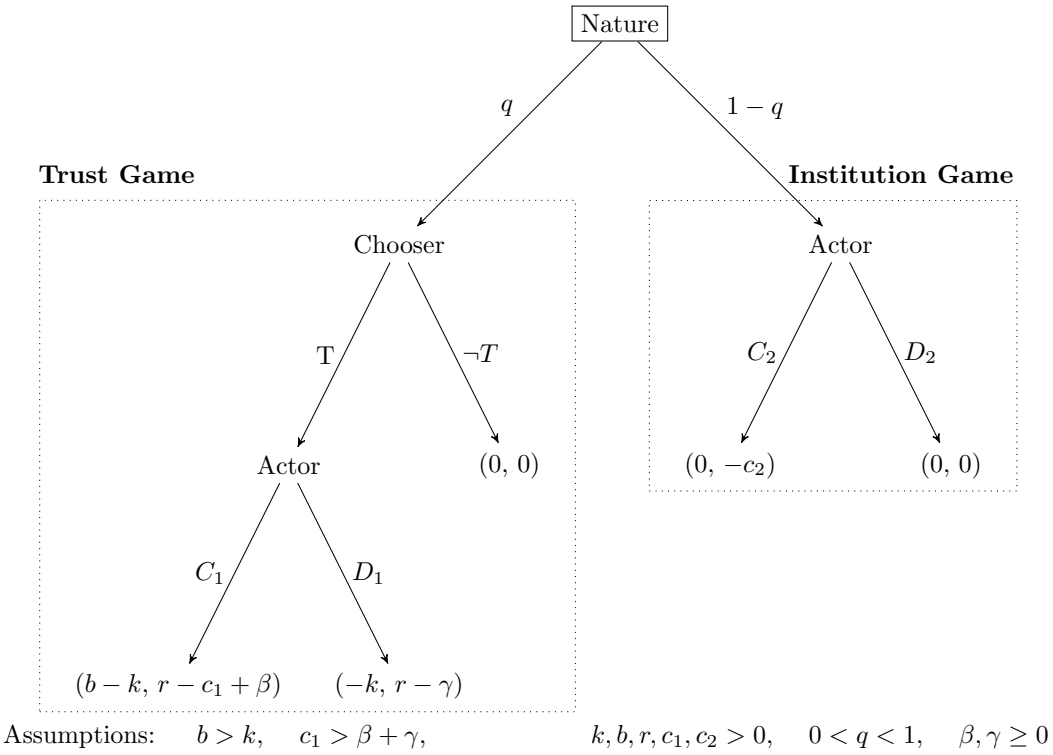


Figure 1: **Stage game**. Nature begins by setting the interaction type: the trust game is drawn with probability q, and the institution game is drawn with probability $1-q$. In the **trust game** (left branch), the chooser and the actor play an asymmetric prisoner's dilemma. We assume $b > k$ and $c_1 > \beta + \gamma$ to keep the structure of an asymmetric prisoner's dilemma. For the chooser, playing T instead of $\neg T$ is net beneficial if the actor subsequently plays C_1, but net costly if the actor subsequently plays D_1. For the actor, playing C_1 instead of D_1 is always net costly, despite the effect of the institution. The institution is materialized here by a reward $\beta \geq 0$ granted in the case that the actor plays C_1, and a penalty $\gamma \geq 0$ inflicted in the case that the actor plays D_1 (we detail the functioning of the institution in section S1.5). In the **institution game** (right branch), the actor plays alone, and decides whether or not to contribute to the institution by paying $c_2 > 0$.

Two players, one chooser and one actor, engage in an infinitely repeated game. Our model, which builds on Lie-Panis and André (2022), uses the framework of Mailath and Samuelson (2006), as well as their notations and definitions. The stage game illustrated in Figure 1 is infinitely repeated, for each of the rounds $t \in \mathbb{N}$.

Each stage proceeds as follows. First, nature draws between two types of interaction: the trust game with probability q, and the institution game with probability $1-q$. Then, the actor and the chooser play according to the rules of the interaction at hand. The trust game and the institution game are thus shorthands. We use these terms to refer to two subgames within the larger extensive form game that constitutes the stage game.

In the trust game, both players play an asymmetric prisoner's dilemma with two steps. In the first step, the chooser decides whether to trust (i.e. play action T) or not trust ($\neg T$) the actor, putting an early end to the

interaction. If he trusts her, the actor then decides whether to reciprocate (C_1) that trust, or cheat (D_1), in the second step. We refer to reciprocation as first-order cooperation, and to cheating as first-order defection (or simply cooperation and defection when there is no ambiguity), which is why we use the labels C_1 and D_1 to designate these two actor actions.

Trust costs $k > 0$ to the chooser, and brings benefit $r > 0$ to the actor. First-order cooperation (reciprocation) costs $c_1 > 0$ to the actor, and brings benefit $b > 0$ to the chooser. We assume $b > k$: the chooser benefits from trusting the actor if she subsequently reciprocates that trust.

In the institution game, the actor plays alone. She can either contribute (C_2) to an institution, whose functioning is described in section S1.5, or free-ride (D_2). We refer to contribution as second-order cooperation, and to free-riding as second-order defection, which is why we use the labels C_2 and D_2 to designate these two other actor actions. Second-order cooperation (contribution) costs $c_2 > 0$ to the actor.

S1.3 Reputation of the actor

The chooser is short-lived, and plays only round of the repeated game. He earns payoffs in that round only. At the end of each round, the chooser is replaced by another individual who takes on the same role in the following round. In a given round, we refer to the actor's current co-player as the current chooser (or simply the chooser, when there is no ambiguity), and to her co-player in the next round as the next chooser.

We restrict the information available to the chooser in the following manner. At the end of each round, we assume that the next chooser privately observes the actor's action in that round with baseline probability p_1 if the actor faced the trust game and was trusted by the current chooser (this probability can be increased through the effect of the institution; see section S1.5), and with fixed probability p_2 if the actor faced the institution game ($0 < p_1 \leq 1$, $0 < p_2 \leq 1$). We assume that the next chooser does not observe the actor's behavior in rounds before the one who just ended, and does not observe the behavior of previous choosers.

What this means is that when the current chooser faces the trust game, and therefore the option to trust or not trust the actor, he can be in one of five situations. If the actor did not play in round $t-1$ (because we are in the initial round $t = 0$, or because the actor previously faced the trust game and was not trusted) or if her action was not observed, the chooser does not have access to any information. Otherwise, the chooser has access to one piece of information, pertaining to the actor's action in round $t-1$.

We refer to this piece of information as the actor's *reputation*, or, interchangeably, as the information available to the (current) chooser. (Note that the actor's reputation is defined with respect to her current partner.) We note $\mathcal{R} \equiv \{\emptyset, \mathcal{C}_1, \mathcal{D}_1, \mathcal{C}_2, \mathcal{D}_2\}$ the set of possible actor reputations, $\emptyset$ referring to the case of an 'empty' reputation (i.e. the case when the chooser in a given round has no information), $\mathcal{C}_1$ referring to the case when the chooser has observed the actor playing C_1 in the previous round, and so on. We note $\mathcal{R}^* \equiv \mathcal{R} \setminus \{\emptyset\}$ the set of non-empty reputations.

S1.4 Life of the actor

The actor is long-lived, and plays all rounds of the infinitely repeated game. She is characterized by a discount factor $\delta \in (0, 1)$, which is drawn at birth. δ is drawn according to the the population distribution of discount factors, which we assume is continuous and of support $\Delta \equiv [0, 1]$. The value of δ is hidden, and remains constant throughout the repeated game.

Throughout the repeated game, the actor discounts payoffs according to her discount factor δ. Her lifetime payoff is equal to the payoff earned in the initial round ($t = 0$), plus δ times the payoff earned in the next round ($t = 1$), plus δ^2 times the payoffs earned in the round after that ($t = 2$), and so on.

S1.5 Effect of the institution on first-order cooperation

An institution collects the expected contribution of the actor (in the institution game), and transforms it into incentives for first-order cooperation (in the trust game). The institution considered here is not a player; for a given set of parameter values, its functioning is fixed. However, it relies on the actor's behavior in the institution game: if the actor never contributes, the institution cannot provide any incentives for first-order cooperation. As detailed below (see section S2), the actor's strategy is allowed to vary with her discount factor and her current reputation. A priori, her behavior is probabilistic: knowing her strategy, her reputation and the population distribution of discount factors (but not the actor's personal discount factor), one can compute the probability that the actor will play C_2 when faced with the institution game.

We note f_2 that probability. In a given round, the institution receives an amount $(1-q) f_2 c_2$ in expectation—this expected contribution being calculated at the beginning of a round (before either game is drawn), knowing the actor's

strategy, her reputation that round, and the population distribution of discount factors.

In any given round, we assume that the institution receives this expected contribution with certainty. Remember that we have an infinite population in mind, as explained in section S1.1. With an infinite population of actors, each round would see a fraction $(1-q)f_2$ of the total population pay c_2 to contribute to the institution. (Note that we will show that the actor's strategy is stationary in every subgame perfect equilibrium, i.e. does not depend on her current reputation. Were we to extend to an infinite-player model, we could define f_2 in such an equilibrium without having to keep track of each individual's reputation.)

We take the amount received by the institution, multiply it by a factor $\rho > 0$, and split the result $\rho(1-q)f_2c_2$ between three types of incentives: a reward for cooperation $\beta \geq 0$, a penalty for defection $\gamma \geq 0$, and an increase $\pi_1 \geq 0$ in the baseline probability of observation in the trust game. In the same round, if the actor faces the trust game and is trusted by the chooser, she earns total payoff $r - c_1 + \beta$ if she plays C_1, $r - \gamma$ if she plays D_1, and is observed with total probability $p_1 + \pi_1$. We assume that:

$$\rho(1-q)f_2c_2 = q(\beta + \gamma + c_1\pi_1) \tag{1}$$

To interpret this equation, remember again that we have an infinite population in mind. With an infinite population of actors, each round would see a fraction q of trust games which can be incentivized by the funds from the multiplied amount $\rho(1-q)f_2c_2$. We assume that incentives produced by the institution apply equally to every trust game, and that their sum is equal to this multiplied amount (a factor of conversion c_1 is applied to the probability π_1).

Note that we use Greek letters to refer to the institution and the incentives it creates throughout the model. ρ is a measure of the institution's efficiency: for every dollar in total contribution, ρ dollars are created to incentivize first-order cooperation. β, γ and π_1 are left unspecified: with this general model, we can consider different types of institutions. For instance, a purely punishing institution is obtained by taking $\beta = \pi_1 = 0$; in that case, the total contribution is entirely allocated to punishing defectors, who are inflicted a penalty of $\gamma = \rho f_2 c_2(1-q)/q$. A purely monitoring institution is obtained by taking $\beta = \gamma = 0$; in that case, the probability of observation in the trust game increases by $\pi_1 = \rho f_2(c_2/c_1)(1-q)/q$.

Accounting for the effect of the institution in a given round, the net cost of cooperation is equal to the total payoff of defectors minus the total payoff of cooperators, that is: $(r-\gamma) - (r - c_1 + \beta) = c_1 - (\beta + \gamma)$. We assume that, even after accounting for the effect of the institution, cooperation remains costly for actors, that is: $c_1 > \beta + \gamma$. In addition, we naturally assume that the likelihood of observation in the trust game remains below 1, i.e. that: $p_1 + \pi_1 \leq 1$.

S2 Technical assumptions and strategy space

S2.1 History equivalence classes and strategies

S2.1.1 Chooser history equivalence classes and strategy space

The chooser only plays in rounds in which the trust game is drawn. Because we strongly restrict the information available to the chooser, chooser histories of the repeated game can be divided in five equivalence classes, depending on the actor's reputation in the eyes of the (current) chooser. We note $\mathcal{H}_{ch} \mid R$ the equivalence class attained when the actor's reputation is $R \in \mathcal{R}$, and note $\mathcal{H}_{ch} \mid \mathcal{R}$ the set comprised of the five equivalence classes for histories of the repeated game; the set of chooser histories $\mathcal{H}_{ch}$ is the union of those equivalence classes.

For simplicity, we equate $\mathcal{R}$ with $\mathcal{H}_{ch} \mid \mathcal{R}$. That is, we define chooser strategy directly as a function of actor reputation, rather than as a function of the history equivalence class. A pure strategy for a chooser specifies whether to trust or not trust the actor depending on her reputation; it is a map:

$$\sigma_{ch} : \mathcal{R} \to \{T, \neg T\}$$

We restrict to the set of chooser strategies $\mathcal{S}_{ch}$ which is pure for non-empty reputations, i.e. the set of strategies following which the chooser plays either T or $\neg T$ with certainty given any information $R \in \mathcal{R}^*$. We note $\sigma^*_{ch} \equiv \sigma_{ch} \mid_{\mathcal{R}^*}$ the restriction of a chooser's strategy to the non-empty information set. There are $2^4 = 16$ possible values for σ^*_{ch}, and an infinite number of possible chooser strategies since we allow choosers to mix between T and $\neg T$ given $\emptyset$. When the chooser plays according to a strategy $\sigma_{ch} \in \mathcal{S}_{ch}$, we note θ the probability that she trusts given $\emptyset$; a chooser's strategy is completely described by σ^*_{ch} and $\theta \in [0,1]$.

We thus allow the chooser to mix given $\emptyset$. We return to this issue in section S3.3.2 in which we calculate the value of θ in equilibrium. Our calculation shows that restricting to pure strategies would lead us not to consider

certain equilibria—under certain parameter conditions, the chooser benefits from deviation to trusting given $\theta = 0$ and from deviation to trusting given $\theta = 1$ (because the value of θ influences actor strategy). In fact, the equilibrium value of θ will be an important value, capturing the baseline level of trust in that equilibrium (see section S5).

S2.1.2 Actor strategy space

The actor does not always have an opportunity to act. In each round, there are three possibilities: either the trust game is drawn and the chooser plays T, in which case the actor has an opportunity to play C_1 or play D_1; or the trust game is drawn and the chooser plays $\neg T$, in which case the actor does not play this round; or the institution game is drawn, in which case the actor has an opportunity to play C_2 or play D_2. We note $\mathcal{T}$, $\neg\mathcal{T}$ and $\mathcal{I}$ the corresponding events, in the above order.

We restrict actor strategy space in accordance to the restriction applied to chooser strategy space, taking into account only what is relevant to the chooser once either event $\mathcal{T}$ or $\mathcal{I}$ has occurred, and the actor decides between playing C_1 and D_1, or C_2 and D_2, respectively. In other words, since choosers play only according to reputation, we do not need to consider the entire set of possible histories for the actor; we only need to consider elements of the set $\mathcal{R} \times \{\mathcal{T}\}$, and elements of the set $\mathcal{R} \times \{\mathcal{I}\}$. (Since the actor does not play after event $\neg\mathcal{T}$, we do not define actor strategy following that event).

A pure strategy for the actor specifies whether to reciprocate or cheat after being trusted in the trust game, and whether to contribute or free-ride in the institution game, depending on the actor's (current) reputation, and her (fixed) discount factor; it is comprised of two maps:

$$\begin{aligned} \sigma_{act} : \ & \mathcal{R} \times \{\mathcal{T}\} \times \Delta \to \{C_1, D_1\} \\ & \mathcal{R} \times \{\mathcal{I}\} \times \Delta \to \{C_2, D_2\} \end{aligned}$$

We restrict to the set $\mathcal{S}_{act}$ of pure strategies for the actor.

S2.1.3 Continuation strategy profile

For every $R \in \mathcal{R}$, the continuation game associated with R is defined as the infinitely repeated game in which the chooser initially has information R, corresponding to the actor's initial reputation. The continuation game associated with R occurs each time the actor attains reputation R at the end of the previous round.

In the continuation game associated with R, the chooser plays directly after. For every strategy profile σ, we note $\sigma \mid_R$ the continuation strategy profile induced by R.

The actor plays after histories of the form $\{R, \mathcal{T}\}$ and $\{R, \mathcal{I}\}$. For every strategy profile σ, and every $(R, \mathcal{X}) \in \mathcal{R} \times \{\mathcal{T}, \mathcal{I}\}$, we note $\sigma \mid_{R,\mathcal{X}}$ the continuation strategy profile induced by $(R, \mathcal{X})$.

S2.1.4 Payoffs

For every σ and R, we note $u(\sigma \mid_R)$ the expected payoff of the chooser in the continuation game. This is the payoff that the chooser can expect to gain in the current round, given that the trust game is drawn, when players play according to σ, and the chooser has information R on the actor.

The actor earns payoffs throughout the game. When the actor's discount factor is δ, we normalize her lifetime payoffs by multiplying payoffs in each round by $(1-\delta)$. For every δ, σ and R, we note $U_\delta(\sigma \mid_R)$ the lifetime expected payoff of the actor starting from the continuation game associated with R. Since the actor begins with empty reputation, $U_\delta(\sigma) \equiv U_\delta(\sigma \mid_\emptyset)$ is the actor's expected payoff over the entire game.

In addition, we define two other classes of continuation payoffs for the actor, relevant to the histories after which she actually plays, in the trust and institution game respectively. For every δ, σ and R, we note $U_\delta(\sigma \mid_{R,\mathcal{T}})$ the lifetime expected payoff of the actor given history $(R, \mathcal{T})$, and $U_\delta(\sigma \mid_{R,\mathcal{I}})$ the lifetime expected payoff of the actor given history $(R, \mathcal{I})$. These correspond to the lifetime's payoff of the actor in the continuation game associated with R, once even $\mathcal{T}$ or $\mathcal{I}$ has occurred (hence not comprising the benefit of being trusted by the chooser in the first case).

S2.1.5 Objective and equilibrium concept

A strategy profile $\sigma = (\sigma_{ch}, \sigma_{act})$ is a Nash equilibrium of the repeated game if both players' strategy is a best response to the other's, i.e. if:

$$\begin{aligned} &\forall \sigma'_{ch} \in \mathcal{S}_{ch}, && u(\sigma) \geq u(\sigma'_{ch}, \sigma_{act}) \\ &\forall \sigma'_{act} \in \mathcal{S}_{act}, \forall \delta \in \Delta, && U_\delta(\sigma) \geq U_\delta(\sigma_{ch}, \sigma'_{act}) \end{aligned}$$

In lieu of considering all possible Nash equilibria, we consider a more restrictive equilibrium concept—namely, subgame perfection. A strategy profile $\sigma = (\sigma_{ch}, \sigma_{act})$ is a subgame perfect equilibrium of the repeated game if:

$$\begin{aligned} &\forall \sigma'_{ch} \in \mathcal{S}_{ch}, \forall R \in \mathcal{R}, && u(\sigma \mid_R) \geq u((\sigma'_{ch}, \sigma_{act}) \mid_R) \\ &\forall \sigma'_{act} \in \mathcal{S}_{act}, \forall R \in \mathcal{R}, \forall \delta \in \Delta, && U_\delta(\sigma \mid_{R,\mathcal{T}}) \geq U_\delta((\sigma_{ch}, \sigma'_{act}) \mid_{R,\mathcal{T}}) \\ & && U_\delta(\sigma \mid_{R,\mathcal{I}}) \geq U_\delta((\sigma_{ch}, \sigma'_{act}) \mid_{R,\mathcal{I}}) \end{aligned}$$

A Nash equilibrium is subgame perfect if, for every possible continuation game, the induced strategy profile is a Nash equilibrium—even when considering unrealized histories; that is, histories which occur with null probability. Here, seeing the restricting assumptions we have made on histories and therefore strategy space, a subgame perfect equilibrium is a strategy profile such that there are no profitable deviations for either player: (i) even when considering reputations that occur with null probability, because the actor never accomplishes a certain actions—e.g., $\mathcal{C}_1$ given that the actor always defects; and (ii) even when considering unrealized combinations of reputation and event $\mathcal{T}$, because the chooser never trusts given certain reputations—e.g. $(\mathcal{D}_1, \mathcal{T})$ given that the chooser does no trust given information $\mathcal{D}_1$.

Objective: Our goal is to characterize the set $\mathcal{S}$ of subgame perfect equilibria of the repeated game, which belong to the set $\mathcal{S}_{ch} \times \mathcal{S}_{act}$ following our restrictive assumptions.

By restricting to subgame perfect equilibria, we restrict to Nash equilibria which are stable to trembles, that is to either player mistakenly playing an unprescribed action with a small, positive probability (Selten, 1983). Arguably, this is the relevant concept when one is interested in endpoints of an evolutionary process—assuming that mistakes or misunderstandings will occur with non-null probability.

Note that restricting to subgame perfect equilibria also leads to getting rid of certain functionally equivalent Nash equilibria. For instance, consider the *uncooperative* strategy profile, defined as the strategy profile whereby: (i) the chooser always plays $\neg T$, whatever the history $R \in \mathcal{R}$; and (ii) the actor always plays D_1 in the trust game, whatever the history $(R, \mathcal{T}) \in \mathcal{R} \times \{\mathcal{T}\}$, and always plays D_2 in the institution game, whatever the history $(R, \mathcal{I}) \in \mathcal{R} \times \{\mathcal{I}\}$. This strategy profile is always a Nash equilibrium because trust is assumed to be costly for the chooser, and first- and second-order cooperation are assumed to be costly for the actor (it is in fact subgame perfect for those reasons). Yet, there are many neutral deviations available. The chooser may for instance deviate to playing T given history $\mathcal{C}_1$. Since $\mathcal{C}_1$ occurs with null probability when the actor plays according to (ii), this unilateral deviation is payoff-neutral. Similarly, $\mathcal{T}$ occurs with null probability when the chooser plays according to (i). The actor may deviate to playing C_1 given history $(\mathcal{D}_1, \mathcal{T})$ without affecting her payoffs (or in fact given any history of the form $(R, \mathcal{T})$ and any discount factor). In both cases, the obtained strategy profile is also a Nash equilibrium, which is functionally equivalent to the one under consideration.

However, only the uncooperative strategy profile is subgame perfect. For instance, the functionally equivalent Nash equilibrium obtained when the actor cooperates given $(\mathcal{D}_1, \mathcal{T})$ instead of defecting is not subgame perfect: given history $(\mathcal{D}_1, \mathcal{T})$, the actor strictly benefits from deviation back to defecting.

S2.2 Three possible subgame perfect equilibria

S2.2.1 General calculation

Let us consider a subgame perfect equilibrium $\sigma = (\sigma_{ch}, \sigma_{act}) \in \mathcal{S}$. Given history $(R, \mathcal{T})$, the actor either cheats and gains $r - \gamma$, or reciprocates and gains only $r - c_1 + \beta < r - \gamma$. When the actor reciprocates, her future reputation is $\mathcal{C}_1$ with probability $p_1 + \pi_1$; when she cheats, her future reputation is $\mathcal{D}_1$ with probability $p_1 + \pi_1$—in either case, her future reputation is $\emptyset$ with probability $1 - (p_1 + \pi_1)$. Following the actor's action, the continuation game associated with $\mathcal{C}_1$, $\mathcal{D}_1$ or $\emptyset$ occurs, depending on the actor's chosen action and the outcome of observation.

Given $(R, \mathcal{T})$, σ_{act} will prescribe playing C_1 if and only if (for simplicity of notations, we assume that when the actor is indifferent between either option, she plays C_1. In section S3.2.3, this is shown to occur with null probability):

$$\begin{aligned} r - c_1 + \beta + \delta[(p_1 + \pi_1)U_\delta(\sigma \mid_{\mathcal{C}_1}) + (1 - p_1 + \pi_1)U_\delta(\sigma \mid_\emptyset)] \geq \\ r - \gamma + \delta[(p_1 + \pi_1)U_\delta(\sigma \mid_{\mathcal{D}_1}) + (1 - p_1 + \pi_1)U_\delta(\sigma \mid_\emptyset)] \end{aligned}$$

Re-arranging, we obtain:

$$c_1 - \gamma - \beta \leq \delta \times (p_1 + \pi_1)(U_\delta(\sigma \mid_{\mathcal{C}_1}) - U_\delta(\sigma \mid_{\mathcal{D}_1})) \tag{2}$$

Similarly, σ_{act} will prescribe playing C_2 given $(R, \mathcal{I})$ if and only if (again, assuming for simplicity of notations that the actor plays C_2 in the equality case, which is shown to occur with null probability in section S3.2.4):

$$c_2 \leq \delta \times p_2(U_\delta(\sigma \mid_{\mathcal{C}_2}) - U_\delta(\sigma \mid_{\mathcal{D}_2})) \tag{3}$$

In a subgame perfect equilibrium, the actor is expected to reciprocate (contribute) when the net cost of first-order cooperation taking into account the effect of the institution $c_1 - \gamma - \beta$ (the cost of second-order cooperation c_2) is smaller than the future lifetime benefit of achieving reputation $\mathcal{C}_1$ instead of $\mathcal{D}_1$ ($\mathcal{C}_2$ instead of $\mathcal{D}_2$) when observed, with probability $p_1 + \pi_1$ (p_2).

We deduce two characteristics that are shared by all subgame perfect equilibria.

S2.2.2 Actor strategy is stationary in a subgame perfect equilibrium

For the first characteristic, note that R is absent from both equations (2) and (3). The actor's strategy is necessarily stationary in a subgame perfect equilibrium. The actor is expected to reciprocate (contribute) depending on: her discount factor, and the lifetime benefit of achieving reputation $\mathcal{C}_1$ instead of $\mathcal{D}_1$ ($\mathcal{C}_2$ instead of $\mathcal{D}_2$)—which solely depend on chooser equilibrium strategy (see below). In contrast, her current reputation R does not come into play.

(Note that through our simplifying assumptions on histories, we have implicitly assumed that the chooser's strategy is stationary: the chooser only sees the history equivalence class, as defined based on the last observation, if any. In consequence, we have shown here that the actor's strategy in a subgame perfect equilibrium will also be stationary—although the previous reasoning shows this need not be the case in a Nash equilibrium).

S2.2.3 Possible equilibrium chooser strategies for non-empty reputations σ^*_{ch}

For the second characteristic, note that by assumption $c_1 - \gamma - \beta > 0$. Playing C_1 instead of C_2 leads to an immediate payoff loss, which can be upset by a future gain depending on chooser strategy, that will be reflected in the difference $U_\delta(\sigma\,|_{\mathcal{C}_1}) - U_\delta(\sigma\,|_{\mathcal{D}_1})$.

We deduce that there are only two possibilities. Either the chooser trusts given $\mathcal{C}_1$ and does not trust given $\mathcal{D}_1$, in which case $U_\delta(\sigma\,|_{\mathcal{C}_1}) - U_\delta(\sigma\,|_{\mathcal{D}_1}) > 0$, and the actor is expected to cooperate given sufficiently high values of δ following equation (2). Or the chooser plays any other combination of actions given $\mathcal{C}_1$ and $\mathcal{D}_1$, in which case there is no incentive to reciprocate, and it must be that the actor always cheats on her partners.

Similarly, because contribution to the institution is costly ($c_2 > 0$) there are again two possibilities: either the chooser trusts given $\mathcal{C}_2$ and does not trust given $\mathcal{D}_2$, and the actor should play C_2 for sufficiently high values of δ; or the chooser plays any other strategy, in which case it must be that the actor always plays D_2.

Putting these two observations together, we deduce that there are only three possibilities for σ^*_{ch} (we show below that these three possibilities correspond to three different subgame perfect equilibria). First, the chooser can play $\neg T$ given any non-empty information. In such a situation, there is no reputational incentives for either costly prosocial action—the actor will never reciprocate nor contribute in equilibrium, and it must be that the chooser also does not trust given empty information. We are necessarily in the uncooperative subgame perfect equilibrium described before—which we refer to as the *uncooperative equilibrium* for simplicity.

Second, the chooser can play T given $\mathcal{C}_1$, and play $\neg T$ given $R \in \{\mathcal{D}_1, \mathcal{C}_2, \mathcal{D}_2\}$. In such a situation, reputation only incentivizes cooperation. We obtain a potential subgame perfect equilibrium, which we refer to as the *baseline equilibrium*. In section S4, we characterize the baseline equilibrium, and derive conditions under which it is indeed a subgame perfect equilibrium.

Third, the chooser can play T given $\mathcal{C}_1$ and $\mathcal{C}_2$, and play $\neg T$ given $\mathcal{D}_1$ and $\mathcal{D}_2$. In such a situation, reputation incentivizes cooperation and second-order cooperation. We obtain a potential subgame perfect equilibrium, which we refer to as the *institution equilibrium*. In section S3 just below, we characterize the institution equilibrium, and derive conditions under which it is indeed subgame perfect.

Note that in both of these latter cases, characterizing the equilibrium includes finding the equilibrium value of θ—that is, the equilibrium probability that the chooser trusts given $\emptyset$. As of yet, we have not proven anything regarding the equilibrium value of θ (we will in section S3.3).

S3 Institution equilibrium

In this section, we study the institution equilibrium more precisely. After introducing useful notations in section S3.1, we characterize actor strategy in section S3.2. In section S3.3 we characterize chooser strategy, by providing an algorithm for determining the equilibrium value of θ (the equilibrium probability that the chooser trusts given no information). We then obtain necessary conditions for an institution equilibrium to occur, which taken together are sufficient.

Note that the formulas obtained in this section are under-determined. They depend on incentives produced by the institution β, γ and π_1 as well as the probability of contribution f_2, which itself is a function of β, γ and π_1, which depend on f_2, and so on. We return to this issue in section S5, where we demonstrate how to use the results

of this section in order to compute the unique institution equilibrium (if it exists), given a set of parameters and a specific type of institution.

S3.1 Objective and simplifying notations

Let $\sigma = (\sigma_{ch}, \sigma_{act})$ be the institution equilibrium, i.e. the subgame perfect equilibrium in which the chooser trusts given $\emptyset$ with probability θ, trusts given $\mathcal{C}_1$ and $\mathcal{C}_2$, and does not trust given $\mathcal{D}_1$ and $\mathcal{D}_2$. We use the generic (yet to be specified) notations σ, θ and σ_{act} throughout this section.

Our objective is two fold. First, we aim to characterize the form that σ must take for it to constitute a subgame perfect equilibrium, by characterizing the equilibrium value of θ on the one hand (and therefore the chooser's equilibrium strategy), and by characterizing the actor's strategy σ_{act} on the other. Second, we aim to derive necessary and sufficient conditions guaranteeing the existence of the institution equilibrium.

Note that reputation affects only one partner choice in our set up (the chooser's decision to play T or $\neg T$)—as a consequence, reputation incentivizes first- and second-order cooperation the same. We note $U_\delta^G \equiv U_\delta(\sigma \mid_{\mathcal{C}_1}) = U_\delta(\sigma \mid_{\mathcal{C}_2})$ the actor's lifetime payoff in a continuation game associated with $\mathcal{C}_1$ or $\mathcal{C}_2$—in which case we say that the actor is in *good standing*. Similarly, we note $U_\delta^B \equiv U_\delta(\sigma \mid_{\mathcal{D}_1}) = U_\delta(\sigma \mid_{\mathcal{D}_2})$, and say that the actor is in *bad standing* when her reputation is $\mathcal{D}_1$ or $\mathcal{D}_2$. Finally, we note $U_\delta^\emptyset = U_\delta(\sigma \mid_\emptyset)$, and say that the actor is in *null standing* when her reputation is $\emptyset$.

We define the *reputational benefit* of good behavior (or simply reputational benefit) $R_\delta(\sigma)$ to be the difference in continuation payoffs given good vs. bad standing, i.e. we define:

$$R_\delta(\sigma) \equiv U_\delta^G - U_\delta^B$$

The actor's payoffs depend on her (stationary) strategy, her discount factor, and her current reputation—brought down to her current standing (good, bad or null) due to how the chooser makes decisions. We have:

$$U_\delta^\emptyset = \theta U_\delta^G + (1-\theta) U_\delta^B \tag{4}$$

Using the above definition, this is equivalent to:

$$U_\delta^\emptyset - U_\delta^B = \theta R_\delta(\sigma) \tag{5}$$

$$U_\delta^\emptyset - U_\delta^G = -(1-\theta) R_\delta(\sigma) \tag{6}$$

S3.2 Characterization of actor equilibrium strategy

S3.2.1 Form of σ_{act}

The actor reciprocates given $\mathcal{T}$ and contributes given $\mathcal{I}$ depending solely on her discount factor δ (or more simply reciprocates and contributes depending on δ). We note $\Delta_{C_1} \subset \Delta$ the subset of discount factors for which σ_{act} prescribes reciprocation, that is the maximum interval such that, $\forall R \in \mathcal{R}, \forall \delta \in \Delta_{C_1}, \sigma_{act}(R, \mathcal{T}, \delta) = C_1$. We note $\Delta_{D_1} \equiv \Delta \setminus \Delta_{C_1}$ the subset of discount factor which σ_{act} prescribes cheating; and define analogously Δ_{C_2} and Δ_{D_2}.

We have partitioned the support $\Delta = [0,1]$ into four, depending on the two actions prescribed by σ_{act} in both games. For instance, if $\delta \in \Delta_{C_1} \cap \Delta_{C_2}$, the actor will play C_1 and C_2 each time she is given the opportunity to, throughout the repeated game. We show below that this partition can be greatly simplified, based on two threshold discount factors.

S3.2.2 Derivation of the reputational benefit $R_\delta(\sigma)$

Let us begin by noting that prescribed behavior in the institution game does not affect the form of the reputational benefit. To illustrate, let us consider values of the discount factor inside $\Delta_{D_1} \cap \Delta_{D_2}$. We denote any (a priori small) value inside this set by δ^-. Given bad standing, the actor either faces the trust game with probability q—in which case the chooser does not trust and the actor's standing becomes null, or faces the institution game with probability $1-q$—in which case the actor does not contribute, and achieves either bad or null standing depending on whether she is observed. In other words, we have:

$$U_{\delta^-}^B = q(0 + \delta_- U^\emptyset) + (1-q)(0 + \delta_-[p_2 U_{\delta^-}^B + (1-p_2) U^\emptyset])$$

Given good standing, we obtain:

$$U^G_{\delta_-} = q(r - \gamma + \delta_-[(p_1 + \pi_1)U^B_{\delta_-} + (1 - p_1 - \pi_1)U^\emptyset]) \\ + (1 - q)(0 + \delta_-[p_2 U^B_{\delta_-} + (1 - p_2)U^\emptyset])$$

We deduce:

$$R_{\delta_-}(\sigma) = q(r - \gamma + \delta_-(p_1 + \pi_1)(U^B_{\delta_-} - U^\emptyset_{\delta_-})$$

$R_\delta(\sigma)$ depends solely on the actor's behavior in the trust game—in this case cheating, given $\delta^- \in \Delta_{D_1}$. We obtain the same expression for $\delta^- \in \Delta_{D_1} \cap \Delta_{C_2}$.

Replacing $U^\emptyset$ using equation (5) we deduce:

$$R_{\delta_-}(\sigma) = q(r - \gamma - \delta_-(p_1 + \pi_1)\theta R_{\delta_-}(\sigma))$$
$$R_{\delta_-}(\sigma) = \frac{q(r - \gamma)}{1 + q\delta_-(p_1 + \pi_1)\theta} \qquad \forall \delta_- \in \Delta_{D_1}$$

Similarly, for $\delta_+ \in \Delta_{C_1}$, we have:

$$R_{\delta_+}(\sigma) = q \times [r - c_1 + \beta + \delta_+(p_1 + \pi_1)(U^G_{\delta_+}(\sigma) - U^\emptyset_{\delta_+}(\sigma)] + q \times [0]$$

From which we deduce, using equation (6):

$$R_{\delta_+}(\sigma) = q(r - c_1 + \beta + \delta_+(p_1 + \pi_1)(1 - \theta)R_{\delta+}(\sigma))$$
$$R_{\delta_+}(\sigma) = \frac{q(r - c_1 + \beta)}{1 - q\delta_+(p_1 + \pi_1)(1 - \theta)} \qquad \forall \delta_+ \in \Delta_{C_1}$$

S3.2.3 Threshold discount factor for first-order cooperation

Using equation (2) and the definition of Δ_{C_1} and Δ_{D_1}, we have (recall that we have assumed that the actor plays C_1 when she is indifferent between both options):

$$\begin{cases} c_1 - \gamma - \beta > \delta_-(p_1 + \pi_1)R_{\delta_-}(\sigma) & \forall \delta_- \in \Delta_{D_1} \\ c_1 - \gamma - \beta \leq \delta_+(p_1 + \pi_1)R_{\delta_+}(\sigma) & \forall \delta_+ \in \Delta_{C_1} \end{cases}$$

Replacing using the above formulas, we deduce:

$$\begin{cases} c_1 - \gamma - \beta > \delta_-(p_1 + \pi_1)\frac{q(r-\gamma)}{1+q\delta_-(p_1+\pi_1)\theta} & \forall \delta_- \in \Delta_{D_1} \\ c_1 - \gamma - \beta \leq \delta_+(p_1 + \pi_1)\frac{q(r-c_1+\beta)}{1-q\delta_+(p_1+\pi_1)(1-\theta)} & \forall \delta_+ \in \Delta_{C_1} \end{cases}$$

From which we obtain:

$$\begin{cases} c_1 - \gamma - \beta > \delta_-(p_1 + \pi_1)q[r - \gamma - \theta(c_1 - \gamma - \beta)] & \forall \delta_- \in \Delta_{D_1} \\ c_1 - \gamma - \beta \leq \delta_+(p_1 + \pi_1)q[r - c_1 + \beta + (1 - \theta)(c_1 - \gamma - \beta)] & \forall \delta_+ \in \Delta_{C_1} \end{cases}$$

Using $r - \gamma - \theta(c_1 - \gamma - \beta) = r - c_1 + \beta + (1 - \theta)(c_1 - \gamma - \beta)$, we deduce finally that:

$$\Delta_{D_1} = [0, \hat{\delta}_1(\theta)[$$
$$\Delta_{C_1} = [\hat{\delta}_1(\theta), 1]$$

Where:

$$\hat{\delta}_1(\theta) = \frac{c_1 - \gamma - \beta}{(p_1 + \pi_1)q[r - \gamma - \theta(c_1 - \gamma - \beta)]} \tag{7}$$

In an institution equilibrium, the actor cooperates if and only if her discount factor exceeds the threshold $\hat{\delta}_1(\theta)$. We refer to $\hat{\delta}_1(\theta)$ as the *difficulty of first-order cooperation in the institution equilibrium for* θ. $\hat{\delta}_1(\theta)$ is defined in

the institution equilibrium for which the probability of trust given empty reputation is θ; its value depends on the equilibrium value that this probability will take.

Condensed notations: noting $\mathbf{r_C} = r - c_1 + \beta$ the stage payoff to a cooperator, $\mathbf{r_D} = r - \gamma$ the stage payoff to a defector, $\mathbf{c_1} = \mathbf{r_D} - \mathbf{r_C} = c_1 - \gamma - \beta$ the cost of cooperation, and $\mathbf{p_1} = p_1 + \pi_1$ the probability of observation in the trust game—all of which taking into account the effect of the institution at equilibrium, we obtain:

$$\hat{\delta}_1(\theta) = \frac{\mathbf{c_1}}{\mathbf{p_1} q[\mathbf{r_D} - \theta \mathbf{c_1}]} \tag{7'}$$

We can also write $R_\delta(\sigma)$ solely as a function of θ using these notations;

$$R_\delta(\theta) = \begin{cases} \frac{q\mathbf{r_D}}{1+\delta \mathbf{p_1} q\theta} & \delta < \hat{\delta}_1(\theta) \\ \frac{q\mathbf{r_C}}{1-\delta \mathbf{p_1} q(1-\theta)} & \delta \geq \hat{\delta}_1(\theta) \end{cases} \tag{8}$$

S3.2.4 Threshold discount factor(s) for second-order cooperation

Using equation (3) and an analogous reasoning, we deduce that the actor will contribute given $\mathcal{I}$ if and only if her discount factor exceeds the threshold $\hat{\delta}_2(\theta)$ satisfying the equation:

$$c_2 = \hat{\delta}_2(\theta) p_2 R_{\hat{\delta}_2(\theta)}(\theta)$$

Similarly to above, we refer to $\hat{\delta}_2(\theta)$ as the *difficulty of second-order cooperation in the institution equilibrium for* θ. There are two cases, depending on whether this threshold is smaller or greater than $\hat{\delta}_1(\theta)$—that is whether second-order cooperation can be said to be 'easier' or 'more difficult' than first-order cooperation.

First case (second-order cooperation is easier): when $\hat{\delta}_2(\theta) < \hat{\delta}_1(\theta)$, the critical reputational benefit is obtained for a discount factor inside Δ_{D_1}. Replacing using the relevant formula, we deduce:

$$\begin{aligned} c_2 &= \hat{\delta}_2(\theta) p_2 \frac{q(r-\gamma)}{1 + q\hat{\delta}_2(\theta)(p_1+\pi_1)\theta} \\ \hat{\delta}_2(\theta) &= \frac{c_2}{q[p_2(r-\gamma) - (p_1+\pi_1)\theta c_2]} \qquad \text{if } \hat{\delta}_2(\theta) < \hat{\delta}_1(\theta) \end{aligned}$$

Second case (second-order cooperation is more difficult): when $\hat{\delta}_2(\theta) \geq \hat{\delta}_1(\theta)$, we obtain:

$$\begin{aligned} c_2 &= \hat{\delta}_2(\theta) p_2 \frac{q(r-c_1+\beta)}{1 - q\hat{\delta}_2(\theta)(p_1+\pi_1)(1-\theta)} \\ \hat{\delta}_2(\theta) &= \frac{c_2}{q[p_2(r-c_1+\beta) + (p_1+\pi_1)(1-\theta)c_2]} \qquad \text{if } \hat{\delta}_2(\theta) \geq \hat{\delta}_1(\theta) \end{aligned}$$

Using the first condition and the condensed notations introduced above, we deduce that if $\hat{\delta}_2(\theta) < \hat{\delta}_1(\theta)$, then we must have:

$$\begin{aligned} \frac{c_2}{q[p_2\mathbf{r_D} - \mathbf{p_1}\theta c_2]} &< \frac{\mathbf{c_1}}{\mathbf{p_1} q[\mathbf{r_D} - \theta \mathbf{c_1}]} \\ c_2 \mathbf{p_1}(\mathbf{r_D} - \theta \mathbf{c_1}) &< \mathbf{c_1}(p_2 \mathbf{r_D} - \mathbf{p_1}\theta c_2) \\ c_2 \mathbf{p_1} \mathbf{r_D} &< \mathbf{c_1} p_2 \mathbf{r_D} \\ \frac{c_2}{p_2} &< \frac{\mathbf{c_1}}{\mathbf{p_1}} \end{aligned}$$

Using the second condition, we deduce that the above is also a sufficient condition. Second-order cooperation is easier than first-order cooperation if and only if its cost divided by the relevant probability of observation p_2 is smaller than the net cost of first-order cooperation divided by the relevant total probability of observation $\mathbf{p_1}$. In other words:

$$\hat{\delta}_2(\theta) < \hat{\delta}_1(\theta) \iff \frac{c_2}{p_2} < \frac{\mathbf{c_1}}{\mathbf{p_1}} = \frac{c_1 - \gamma - \beta}{p_1 + \pi_1} \tag{9}$$

Bringing together the above formulas, we deduce, in condensed form:

$$\hat{\delta}_2(\theta) = \begin{cases} \frac{c_2}{q[p_2\mathbf{r_D} - \mathbf{p_1}\theta c_2]} & \text{if } \frac{c_2}{p_2} < \frac{\mathbf{c_1}}{\mathbf{p_1}} \\ \frac{c_2}{q[p_2\mathbf{r_C} + \mathbf{p_1}(1-\theta)c_2]} & \text{if } \frac{c_2}{p_2} \geq \frac{\mathbf{c_1}}{\mathbf{p_1}} \end{cases} \tag{10}$$

S3.3 Conditions under which chooser strategy is optimal

In the equilibrium under consideration, the chooser trusts given $\mathcal{C}_1$ or $\mathcal{C}_2$, does not trust given $\mathcal{D}_1$ or $\mathcal{D}_2$, and trusts with probability θ given $\emptyset$ which has yet to be specified. The previous section shows that the actor's strategy can then be fully described by two thresholds, $\hat{\delta}_1(\theta)$ and $\hat{\delta}_2(\theta)$.

In this section, we first give a general condition under which the chooser should trust in equilibrium. We deduce a condition giving the equilibrium value of the probability of trust given empty reputation θ^*, and then obtain four necessary conditions given that $\theta = \theta^*$. Put together, these five conditions are sufficient to obtain an institution equilibrium.

S3.3.1 General condition for inferring trust

Let us assume that the chooser faces the trust game, and has information $R \in \mathcal{R}$. If the chooser does not trust, she gains nothing; in contrast, if she trusts, she pays k and receives b in exchange with probability $\mathbf{P}(C_1 \mid R)$, the probability that the actor reciprocates given reputation R. In a subgame perfect equilibrium, σ_{ch} will prescribe trusting if and only if (again, we assume that choosers trust in the equality case for simplicity of notations):

$$\mathbf{P}(C_1 \mid R) \geq \frac{k}{b} \tag{11}$$

S3.3.2 Given $\emptyset$; determination of θ

Following equation (7), $\hat{\delta}_1(\theta)$ is strictly increasing in θ, and varies between $\hat{\delta}_1(0)$ and $\hat{\delta}_1(1)$ when θ varies between 0 and 1. When θ is high, the actor has less to gain by achieving good standing, since she is already likely to be accepted given null reputation; and the minimum bar to exceed $\hat{\delta}_1(\theta)$ increases.

Given null standing (e.g. in the first round), the actor reciprocates with probability $\mathbf{P}(C_1 \mid \emptyset) = \mathbf{P}(\delta \geq \hat{\delta}_1(\theta))$. Since $\hat{\delta}_1(\theta)$ is strictly increasing in θ, $\mathbf{P}(\delta \geq \hat{\delta}_1(\theta))$ is strictly decreasing in θ, and defines a bijection between $[0,1]$ and $[\mathbf{P}(\delta \geq \hat{\delta}_1(1)), \mathbf{P}(\delta \geq \hat{\delta}_1(0))]$.

We deduce that there are three cases. First, if $\mathbf{P}(\delta \geq \hat{\delta}_1(0)) \leq \frac{k}{b}$, then $\mathbf{P}(\delta \geq \hat{\delta}_1(\theta)) < \frac{k}{b}$ must also be true for all $\theta > 0$—the only possibility is that $\theta = 0$ in equilibrium. Second, if $\mathbf{P}(\delta \geq \hat{\delta}_1(1)) \geq \frac{k}{b}$, then $\mathbf{P}(\delta \geq \hat{\delta}_1(\theta)) \geq \frac{k}{b}$ must also be true for all $\theta < 1$—and θ must be equal to 1. Third, if neither of these conditions are verified, then there must exist a unique $\theta \in (0,1)$ such that $\mathbf{P}(\delta \geq \hat{\delta}_1(\theta) = \frac{k}{b}$.

In other words, the equilibrium value of θ is given by:

$$\theta^* = \begin{cases} 0 & \text{if } \mathbf{P}(\delta \geq \hat{\delta}_1(0)) \leq \frac{k}{b} \\ 1 & \text{if } \mathbf{P}(\delta \geq \hat{\delta}_1(1)) \geq \frac{k}{b} \\ t & \text{such that } \mathbf{P}(\delta \geq \hat{\delta}_1(t)) = \frac{k}{b} \end{cases} \tag{12}$$

The above yields the value of θ in a potential equilibrium. Below, we assume $\theta = \theta^*$, and deduce conditions under which we have a subgame perfect equilibrium.

S3.3.3 Given $\mathcal{C}_1$ and $\mathcal{D}_1$

The actor's discount factor $\hat{\delta}$ is distributed following a continuous distribution, whose support is $[0,1]$. If $\hat{\delta}_1(\theta^*) \geq 1$, $\mathcal{C}_1$ is the null event, and $\mathbf{P}(C_1 \mid \mathcal{C}_1)$ cannot be defined. In such a case, the chooser benefits from deviating by playing $\neg T$ given $\mathcal{C}_1$, since the actor then defects with probability 1. We deduce that a necessary condition to obtain the institution equilibrium is:

$$\hat{\delta}_1(\theta^*) < 1 \tag{13}$$

Otherwise, both $\mathcal{C}_1$ and $\mathcal{D}_1$ are defined, and, since actor equilibrium strategy is stationary, we immediately have:

$$\mathbf{P}(C_1 \mid \mathcal{C}_1) = 1 > \frac{k}{b}$$
$$\mathbf{P}(C_1 \mid \mathcal{D}_1) = 0 < \frac{k}{b}$$

Conversely, if $\hat{\delta}_1(\theta^*) < 1$, the chooser benefits (strictly) from trusting given $\mathcal{C}_1$ and not trusting given $\mathcal{D}_1$.

S3.3.4 Given $\mathcal{C}_2$ and $\mathcal{D}_2$

We deduce similarly that a necessary condition is:

$$\hat{\delta}_2(\theta^*) < 1 \tag{14}$$

Under this condition, we can define $\mathbf{P}(C_1 \mid \mathcal{C}_2) = \mathbf{P}(\delta \geq \hat{\delta}_1(\theta^*) \mid \delta \geq \hat{\delta}_2(\theta^*))$ and $\mathbf{P}(C_1 \mid \mathcal{D}_2) = \mathbf{P}(\delta \geq \hat{\delta}_1(\theta^*) \mid \delta < \hat{\delta}_2(\theta^*))$. When $\theta = \theta^*$, the chooser benefits from trusting given $\mathcal{C}_2$ and not trusting given $\mathcal{D}_2$ if and only if:

$$\mathbf{P}(\delta \geq \hat{\delta}_1(\theta^*) \mid \delta \geq \hat{\delta}_2(\theta^*)) \geq \frac{k}{b} \tag{15}$$

$$\mathbf{P}(\delta \geq \hat{\delta}_1(\theta^*) \mid \delta < \hat{\delta}_2(\theta^*)) < \frac{k}{b} \tag{16}$$

Note that one of these equations is always trivially verified. For instance, when second-order cooperation is 'easier' than first-order cooperation, i.e. when $\hat{\delta}_2(\theta^*) < \hat{\delta}_1(\theta^*)$, the actor always defects given past non-contribution, and (16) is trivially verified. In contrast, when second-order cooperation is 'harder', (15) is trivially verified.

S4 Baseline equilibrium

The characteristics of the baseline equilibrium are deduced from the above section—all is needed is to restrict the definition of good standing to achieving reputation $\mathcal{C}_1$, and to 'turn off' the institution in the formulas above by replacing each Greek letter with 0 (i.e. take $\beta = \gamma = \pi_1 = 0$). In particular, we can use condition (8) to deduce the reputational benefit of playing C_1 instead of D_1 in the baseline equilibrium.

We deduce that in the baseline equilibrium in which the chooser trusts with probability θ given $\emptyset$, the actor reciprocates if and only if her discount factor δ is greater than $\hat{\delta}^b(\theta)$ (the actor never contributes), where:

$$\hat{\delta}^b(\theta) = \frac{c_1}{p_1 q(r - \theta c_1)} \tag{17}$$

The equilibrium value of θ is given by:

$$\theta^{*,b} = \begin{cases} 0 & \text{if } \mathbf{P}(\delta \geq \hat{\delta}^b(0)) \leq \frac{k}{b} \\ 1 & \text{if } \mathbf{P}(\delta \geq \hat{\delta}^b(1)) \geq \frac{k}{b} \\ t & \text{such that } \mathbf{P}(\delta \geq \hat{\delta}^b(t)) = \frac{k}{b} \end{cases} \tag{18}$$

We obtain the baseline equilibrium if and only if:

$$\hat{\delta}^b(\theta^{*,b}) < 1 \tag{19}$$

This time, we refer specifically to $\delta^b \equiv \hat{\delta}^b(0)$ as the *intrinsic difficulty of cooperation*, or simply, the difficulty of cooperation. $\delta^b = c1/(p_1 qr)$ is a function of our parameters, and characterizes the repeated game as a whole; in contrast to before, it is not defined in relation to the value of θ in a specific case.

In fact, we can show that the baseline equilibrium exists if and only if:

$$\delta^b < 1 \tag{20}$$

To prove this, consider first the case when $\delta^b \geq 1$. In that case, we have $\hat{\delta}^b(\theta) \geq \delta^b \geq 1$ for all θ: the baseline equilibrium cannot be subgame perfect because no value of $\theta^{ast,b}$ can satisfy equation (19).

Consider second the case when $\delta^b < 1$. There are two possibilities: either $\hat{\delta}^b(\theta) < 1$ for all θ in which case it is immediate the the baseline equilibrium is subgame perfect; or, there exists $\theta^m \leq 1$ such that $\hat{\delta}^b(\theta^m) = 1$, and $\hat{\delta}^b(\theta) < 1 \forall \theta \in [0, \theta^m[$. As θ increases from 0 to θ^m, $\mathbf{P}(\delta \geq \hat{\delta}^b(\theta))$ then strictly decreases from $\mathbf{P}(\delta \geq \delta^b)$ to 0. The value $\theta^{*,b}$ derived using (18) will thus necessarily be in the interval $[0, \theta^m[$—and will therefore satisfy equation (19). This proves the proposed equivalence.

S5 Implementation into Mathematica

S5.1 Motivation and algorithm

The institution equilibrium characterized in section S3 is under-determined.

Given a specific distribution of discount factors, specific parameters, and a specific allocation of incentives performed by the institution (i.e. a specific institution), this section provides two algorithms for determining each equilibrium, which we detail below.

We implement these algorithms into the software Mathematica, to compute characteristics of either equilibrium when it exists, and obtain graphical representations.

A pdf print of the Mathematica Notebook is provided at the end of this document.

Note that we consider a specific case for the baseline equilibrium—namely, we will compare our results in the institution equilibrium with the results for the baseline equilibrium when $p_2 = 0$. This can be interpreted as the case in which we completely 'turn off' the institution—when $p_2 = 0$, there can be no reputational benefit for second-order cooperation, and only the baseline equilibrium can be possible. Note that while p_2 does not affect the actor's equilibrium strategy (since it does not appear in the formulas of section S4), it does affect the actor's average payoff defined below. For instance, when $\theta^{*,b} = 1$, the actor is certain to be trusted each time the institution game is drawn and her reputation is reset to $\emptyset$—in such a case, adding the possibility of second-order cooperation can only decrease her average payoff.

S5.1.1 Algorithm for the baseline equilibrium

The baseline equilibrium characterized in section S4 depends on the distribution of discount factors. We assume from here on a truncated normal distribution, of mode $\mu \in [0, 1]$ and standard deviation $\sigma > 0$.

Given μ and σ, and specific values of c_1, p_1, q and r, the unique equilibrium value of θ can be deduced using condition (18), which we solve numerically for any set of parameters using Mathematica.

We can then deduce whether a specific set of parameters can yield a (necessarily unique) baseline equilibrium using condition (19). Under this condition, we can compute the equilibrium's characteristics, e.g. the level of cooperation and the expected payoff, which are defined in section S5.2.

S5.1.2 Algorithm for the institution equilibrium

The institution equilibrium characterized in section S3 depends on the specific allocation of incentives performed by the institution, as well as the distribution of discount factors.

Using Mathematica, we consider different allocation of incentives—i.e. specific weights attributed to rewards, punishment and monitoring. For instance, the graphs presented in the main document are determined by considering a punishing-monitoring institution, which equally allocates contributions to punishment of defectors and increasing the probability of observation. Such an institution is characterized by $\gamma = \frac{1}{2}(\rho f_2 c_2)\frac{q}{1-q}$, $\pi_1 = \frac{1}{2}(\rho f_2)\frac{q}{1-q}$ and $\beta = 0$; the equilibrium value of f_2 being determined through the algorithm described here.

Given a specific allocation of incentives like the one described above, as well as specific values of other parameters, the equilibrium value of θ can be deduced using condition (12), which we solve numerically for any set of parameters using Mathematica. The parameters that must be specified are: μ, σ, c_1, p_1, q, r (as in the baseline equilibrium), as well as c_2, p_2 and ρ.

We then deduce whether a specific parameter set can yield a (necessarily unique) institution equilibrium using conditions (13-16), and compute the equilibrium's characteristics using the formulas detailed in section S5.2.

S5.1.3 Additional assumptions

In the main article, we argue that institutions can be viewed as a technology to amplify the beneficial effects of reputational incentives. For this reason, we wish to retain an equilibrium in which second-order cooperation remains 'easier' than first-order cooperation, that is such that $\frac{c_2}{p_2} \leq \frac{\mathbf{c_1}}{\mathbf{p_1}}$. To guarantee that, we assume that simultaneously have $c_2 \leq \mathbf{c_1}$ and $p_2 \geq \mathbf{p_1}$ in equilibrium, by assuming that these inequalities are true for the baseline values c_1 and p_2, and that:

$$\beta + \gamma \leq c_1 - c_2$$
$$\pi_1 \leq p_2 - p_1$$

As a result, in each example under consideration, $\hat{\delta}_2(\theta) = \frac{c_2}{q[p_2 \mathbf{r_D} - \mathbf{p_1}\theta c_2]}$.

S5.2 Other useful formula

S5.2.1 A note

Throughout section S5.2, we use the generic σ to refer to a strategy profile which is either the institution or the baseline equilibrium, and use the generic θ to refer to the probability that the chooser trusts according to σ. We use these generic notations in all the intermediate steps leading to the final formulas, which we designate by a numbered equation.

In the final formulas, we specify whether they are valid for $\theta = \theta^*$ (institution equilibrium) or $\theta = \theta^{*,b}$ (baseline equilibrium).

S5.2.2 Level of cooperation

We rely primarly on the level of cooperation in each case. In either the baseline or the institution equilibrium, the *level of cooperation LC* is defined as the probability that the chooser trusts given empty reputation, and that the signaler then reciprocates. Put differently, LC is the probability of dyadic cooperation in the initial round given that the trust game is drawn.

In the institution equilibrium in which the chooser trusts given empty reputation with probability θ, we immediately have:

$$LC = \theta^* \times \mathbf{P}(\delta \geq \hat{\delta}_1(\theta^*)) \tag{21}$$

In the corresponding baseline equilibrium, we have:

$$LC = \theta^{*,b} \times \mathbf{P}(\delta \geq \hat{\delta}^b(\theta^{*,b})) \tag{22}$$

S5.2.3 Average lifetime payoff of the actor

We use the notations $R_\delta(\theta)$, $U_\delta^\emptyset$, U_δ^G and U_δ^B introduced in section S3.1. Our goal here is to characterize $\bar{U}_\delta \equiv U_\delta^\emptyset(1-\delta)$, which is the average lifetime payoff of an actors whose discount factor is δ; that is, the lifetime payoff of null standing averaged over the infinite rounds of the repeated game by multiplying by $(1-\delta)$.

Reminder:

$$U_\delta^\emptyset = \theta U_\delta^G + (1-\theta)U_\delta^B \tag{4}$$

$$U_\delta^\emptyset - U_\delta^B = \theta R_\delta(\theta) \tag{5}$$

$$U_\delta^\emptyset - U_\delta^G = -(1-\theta)R_\delta(\theta) \tag{6}$$

These formulas are valid in both the institution and baseline equilibrium.

In the baseline equilibrium (assuming $p_2 = 0$). When $p_2 = 0$, we have: $\bar{U}_\delta^B = \delta\bar{U}_\delta$. When in bad standing, the actor either is not trusted, or plays in the institution game and is not observed—in the next round, she achieves null standing. Replacing using condition (4) and the definition of $\bar{U}_\delta$, we deduce that in the chosen baseline we always have:

$$\bar{U}_\delta = \theta^{*,b} R_\delta(\theta^{*,b}) \tag{23}$$

In the institution equilibrium. In contrast, in the institution equilibrium, an actor in bad standing can achieve null or good/bad standing in the next round depending on her action in the institution game. We calculate U_δ^B in two cases.

We consider first the case in which the actor's discount factor is $\delta_- < \hat{\delta}_2(\theta)$. We then have:

$$U_{\delta_-}^B = q[0 + \delta_- U_\delta^\emptyset] + (1-q)[0 + p_2\delta_- U_\delta^B + (1-p_2)\delta_- U\emptyset_\delta]$$

Using equation (5), we deduce:

$$U_{\delta_-}^\emptyset - \theta R_{\delta_-}(\theta) = \delta_- U_{\delta_-}^\emptyset + (1-q)p_2\delta_-(-\theta)R_{\delta_-}(\theta)$$

And therefore:

$$\bar{U}_{\delta_-} = (1-\delta_-)U_{\delta_-}^\emptyset = [1-(1-q)p_2\delta_-]\theta R_{\delta_-}(\theta)$$

In contrast, given a discount factor $\delta_+ \geq \hat{\delta}_2(\theta)$, we have:

$$U_{\delta_+}^B = q[0 + \delta_+ U_{\delta_+}^\emptyset] + (1-q)[-c_2 + p_2\delta_+ U_{\delta_+}^G + (1-p_2)\delta_+ U\emptyset_{\delta_+}]$$

From which we deduce, using equation s (5-6)

$$\bar{U}_{\delta_+} = U^{\emptyset}_{\delta_+} = = (1-q)(-c_2) + [\theta + (1-q)p_2\delta_+(1-\theta)]R_{\delta_+}(\theta)$$

Putting all this together, the average lifetime payoff of the actor is:

$$\bar{U}_{\delta} = \begin{cases} [1-(1-q)p_2\delta]\theta^* R_{\delta}(\theta^*) & \delta < \hat{\delta}_2(\theta^*) \\ (1-q)(-c_2) + [\theta^* + (1-q)p_2\delta(1-\theta^*)]R_{\delta}(\theta^*) & \delta \geq \hat{\delta}_2(\theta^*) \end{cases} \tag{24}$$

S5.2.4 Expected payoff of the chooser in the initial round

Recall that the chooser's payoff in a given round is defined assuming that the trust game is drawn. We note u her payoff in the initial round $t = 0$ (this was previously noted $u(\sigma)$).

In the institution equilibrium in which the chooser trusts given empty reputation with probability θ, we immediately have:

$$u = \theta^* \times (\mathbf{P}(\delta \geq \hat{\delta}_1(\theta^*))b - k) \tag{25}$$

In the corresponding baseline equilibrium, we have:

$$u = \theta^{*,b} \times (\mathbf{P}(\delta \geq \hat{\delta}^b(\theta^{*,b}))b - k) \tag{26}$$

S5.2.5 Long-run payoff of the chooser

Strictly speaking, the payoff of the chooser varies throughout the repeated game, as the actor's chances of being in null, good, or bad standing is a function of the round being played—as we will see just below, the actor's reputation follows a Markov chain.

We calculate here the long-run payoff of the chooser u^{∞}, defined as the expected payoff of a chooser who interacts with an infinite population of actors, whose reputation is taken in the steady state.

Note that this is not the quantity that is optimized by choosers. The baseline and institution equilibria are defined assuming that the informational value of each state is constant, in keeping with the idea of a myopic chooser (hence the quantity being optimized is closer to the payoff defined just above).

In the baseline equilibrium (assuming $p_2 = 0$). A cooperative actor's ($\delta \geq \delta_1^b(\theta)$) reputation alternates between null standing $\emptyset$ and good standing G. In each round t in which the actor is in null standing, the actor stays in null standing in round $t+1$ with probability $P_C^{\emptyset\to\emptyset} = (1 - q\theta p_1)$, and otherwise switches to good standing (when she faces a trust game, the chooser trusts, and she is observed). When the actor is in good standing, she stays in good standing with probability $P_C^{G\to G} = qp_1$, and otherwise switches to null standing.

In other words, the cooperative actor's reputation follows a Markov chain. In the steady state, the probabilities $P_C^{\infty,\emptyset}$ and $P_C^{\infty,G}$ that the actor is respectively in state $\emptyset$ or state G verify:

$$\begin{aligned} P_C^{\infty,\emptyset} &= (1 - q\theta p_1)P_C^{\infty,\emptyset} + (1 - qp_1)P_C^{\infty,G} \\ 1 &= P_C^{\infty,\emptyset} + P_C^{\infty,G} \end{aligned}$$

We obtain:

$$P_C^{\infty,\emptyset} = \frac{(1 - qp_1)}{q\theta p_1 + (1 - qp_1)} \tag{27}$$

$$P_C^{\infty,G} = \frac{q\theta p_1}{q\theta p_1 + (1 - qp_1)} \tag{28}$$

Similarly, the cheating actor's reputation follows a Markov chain, with: $P_D^{\emptyset\to\emptyset} = (1 - q\theta p_1)$ and $P_D^{B\to B} = 0$. The steady state probabilities must verify:

$$\begin{aligned} P_D^{\infty,\emptyset} &= (1 - q\theta p_1)P_D^{\infty,\emptyset} + P_D^{\infty,B} \\ 1 &= P_D^{\infty,\emptyset} + P_D^{\infty,B} \end{aligned}$$

From which we deduce:

$$P_D^{\infty,\emptyset} = \frac{1}{q\theta p_1 + 1} \tag{29}$$

$$P_D^{\infty,B} = \frac{q\theta p_1}{q\theta p_1 + 1} \tag{30}$$

Knowing the value of these probabilities in equilibrium, as well as the fraction of cooperative actors $f_1 = \mathbf{P}(\delta \geq \hat{\delta}^b(\theta^{*,b}))$, we can deduce the long-run payoff of the chooser in the baseline equilibrium:

$$u^\infty = f_1 P_C^{\infty,G}(b-k) + f_1 P_C^{\infty,\emptyset}\theta^{*,b}(b-k) + (1-f_1)P_D^{\infty,\emptyset}\theta^{*,b}(-k) \tag{31}$$

In the institution equilibrium. We assume that second-order cooperation is easier than first-order cooperation (i.e. $\delta_2(\theta) \leq \delta_1(\theta)$). There are now three types which we call: low L ($\delta < \delta_2(\theta)$), who always play D_1 and D_2; middle M ($\delta_2(\theta) \leq \delta < \delta_1(\theta)$), who always play D_1 and C_2; and high H ($\delta_1(\theta) \leq \delta$), who always play C_1 and C_2.

Low type. An actor of low type alternates between null and bad standing. She stays in bad standing with probability $P_L^{B\to B} = (1-q)p_2$, and otherwise switches to null standing. She switches from null standing to bad standing with probability $P_L^{\emptyset\to B} = q\theta\mathbf{p_1} + (1-q)p_2$, and otherwise stays in null standing.

The steady state probabilities must verify:

$$\begin{aligned} P_L^{\infty,B} &= (1-q)p_2 P_L^{\infty,B} + (q\theta\mathbf{p_1} + (1-q)p_2)P_L^{\infty,\emptyset} \\ 1 &= P_L^{\infty,B} + P_L^{\infty,\emptyset} \end{aligned}$$

We deduce:

$$P_L^{\infty,B} = \frac{q\theta\mathbf{p_1} + (1-q)p_2}{1 + q\theta\mathbf{p_1}} \tag{32}$$

$$P_L^{\infty,\emptyset} = \frac{1 - (1-q)p_2}{1 + q\theta\mathbf{p_1}} \tag{33}$$

High type. An actor of high type alternates between null and good standing. She stays in good standing with probability $P_H^{G\to G} = q\mathbf{p_1} + (1-q)p_2$, and otherwise switches to null standing. She switches from null standing to good standing with probability $P_H^{\emptyset\to G} = q\theta\mathbf{p_1} + (1-q)p_2$, and otherwise stays in null standing.

The steady state probabilities must verify:

$$\begin{aligned} P_H^{\infty,G} &= (q\mathbf{p_1} + (1-q)p_2)P_H^{\infty,G} + (q\theta\mathbf{p_1} + (1-q)p_2)P_H^{\infty,\emptyset} \\ 1 &= P_H^{\infty,G} + P_H^{\infty,\emptyset} \end{aligned}$$

We deduce:

$$P_H^{\infty,G} = \frac{q\theta\mathbf{p_1} + (1-q)p_2}{1 - q(1-\theta)\mathbf{p_1}} \tag{34}$$

$$P_H^{\infty,\emptyset} = \frac{1 - q\mathbf{p_1} - (1-q)p_2}{1 - q(1-\theta)\mathbf{p_1}} \tag{35}$$

Middle type. An actor of middle type alternates between null, good *and* bad standing. Note however that, whatever her current standing, the actor of middle type reaches good standing when the institution game is drawn and she is observed, with probability $(1-q)p_2$. Indeed, her ability to contribute to the institution does not depend on her standing, and she never achieves good standing when the trust game is drawn since she plays D_1 when trusted (hence the actor will reach bad or null standing after a trust game is drawn). In other words, we necessarily have:

$$P_M^{\infty,G} = (1-q)p_2 \tag{36}$$

The actor stays in good standing with probability $P_M^{G\to G} = (1-q)p_2 = P_M^{\infty,G}$, switches to bad standing with probability $P_M^{G\to B} = q\mathbf{p_1}$, and otherwise switches to null standing.

The actor stays in bad standing with probability $P_M^{B\to B} = 0$, switches to good standing with probability $(1-q)p_2$, and otherwise switches to null standing.

When in null standing, she switches to good standing with probability $(1-q)p_2$, switches to bad standing with probability $P_M^{\emptyset\to B} = q\theta\mathbf{p_1}$, and otherwise stays in null standing.

The other two steady state probabilities must verify:

$$P_M^{\infty,B} = q\mathbf{p_1}P_M^{\infty,G} + 0 + q\theta\mathbf{p_1}P_M^{\infty,\emptyset}$$
$$1 - P_M^{\infty,G} = P_M^{\infty,B} + P_M^{\infty,\emptyset}$$

We obtain:

$$P_M^{\infty,B} = q\mathbf{p_1}\frac{P_M^{\infty,G} + \theta(1 - P_M^{\infty,G})}{1 + q\theta\mathbf{p_1}}$$

Replacing using equation (36), we deduce:

$$P_M^{\infty,B} = q\mathbf{p_1}\frac{\theta + (1-\theta)(1-q)p_2}{1 + q\theta\mathbf{p_1}} \tag{37}$$

$$P_M^{\infty,\emptyset} = \frac{1 - (1-q)p_2(1 + q\mathbf{p_1})}{1 + q\theta\mathbf{p_1}} \tag{38}$$

Noting f_L, f_M and f_B the equilibrium fraction of actors of low, middle and high type respectively, we obtain:

$$u^\infty = f_H P_H^{\infty,G}(b-k) + f_M P_M^{\infty,G}(-k) + \theta^*[f_H P_H^{\infty,\emptyset}(b-k) + (f_M P_M^{\infty,\emptyset} + f_L P_L^{\infty,\emptyset})(-k)] \tag{39}$$

S5.3 Mathematica output

We illustrate our results in fives cases: one case is the baseline equilibrium (or the case of no institution), and the other four are the institution equilibrium, for four different types of institution. More precisely, we compute the institution equilibrium for a purely rewarding institution (where all multiplied contributions are affected to increasing the payoff of cooperators by β), for a purely punishing institution (invest solely in γ), for a purely monitoring institution (invest solely in π_1), and for a monitoring-punishing institution, which equally divides its resources between increasing the likelihood of observation and punishing defectors—this is the example considered in the main article, that this document supplements.

S5.3.1 Level of cooperation

Baseline equilibrium. In Figure 2, we plot the level of cooperation obtained in the baseline equilibrium:

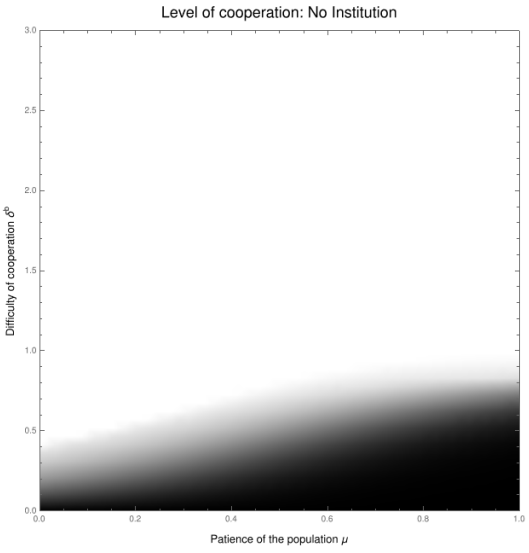


Figure 2: Level of cooperation in the baseline equilibrium, as a function of μ and δ^b.

In this graph—as in all graphs indicating the level of cooperation—the shade of gray indicates the level of cooperation at a given point: black indicates a level of cooperation of 1, and white indicates a level of cooperation of 0. We obtain this graph, as well as all other graphs presented in this supplementary document using using Mathematica's DensityPlot function, with 1 level of recursion and 30 points—plots used for the figures of the main text are computed with higher precision.

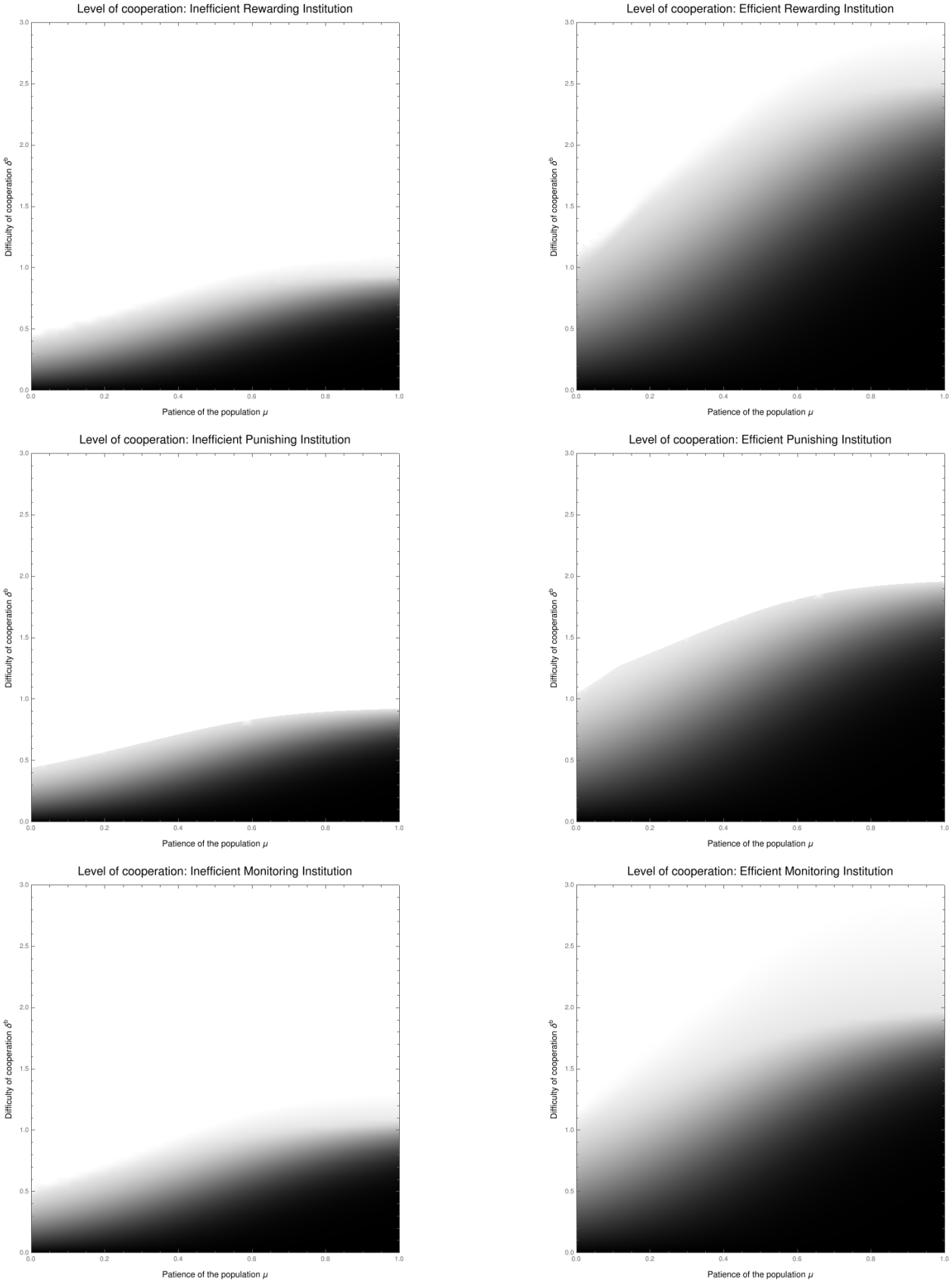


Figure 3: Level of cooperation in the institution equilibrium, for a purely rewarding institution (top row), a purely punishing institution (middle row), and a purely monitoring institution (bottom row). In each case, results are computed as a function of μ and δ^b, for $\rho = 1/3$ (inefficient institution, left column), and for $\rho = 3$ (efficient institution, right column).

To obtain this graph, we fix $\sigma = 1/4$, and vary μ between 0 and 1 on the x-axis. We refer to μ, the mode distribution of the actor population's discount factors as the *patience of the population*.

On the y-axis, we vary the (intrinsic) *difficulty of cooperation* $\delta^b = c_1/(p_1 qr)$ between 0 and 3. To do so, we fix the following parameter values: $q = 1/2$, $r = 2$, $p_1 = 1/3$, and vary c_1 between 0 and 1. These parameter values are chosen so as to normalize the maximum expected actor payoff qr. By fixing $p_1 = 1/3$, we ensure that the difficulty of cooperation verifies $\delta^b = 3c_1$. In other words, in the baseline equilibrium, the maximum attainable cost of cooperation is $c_1 = 1/3$; this leaves plenty of 'room for improvement' for the various examples of institution, which is visible for values of c_1 above 1/3, i.e. for $\delta^b > 1$.

We also fix: $b = 1$, $k = 1/10$. We do this do obtain equilibria for a large set of parameter values, and a more continuous variation of our measures of interest—since k/b is small, choosers can trust even given a small amount of cooperators, and the attained level of cooperation smoothly increases when the patience of the population μ increases and/or the difficulty of cooperation δ^b decreases. Results are similar but less smooth with larger values of k/b.

Institution equilibrium. In Figure 3, we plot the level of cooperation obtained in the institution equilibrium for the first three cases, i.e. the rewarding (top), punishing (middle), and monitoring (bottom) institutions. In each case, we consider the inefficient and efficient variant of the institution, by fixing $\rho = 1/3$ (left column) and $\rho = 3$ (right column).

To generate these six graphs (two per case; one case equals one row), we fix all parameters as above, and again vary δ^b between 0 and 3 by varying c_1 between 0 and 1.

In addition, we fix $p_2 = 1$, and assume that $c_2 = c_1/3$. This allows us to vary the cost of second-order cooperation along with the cost of first-order cooperation, whilst retaining the property that second-order cooperation is easier (i.e. ensuring that $c_2/p_2 \leq \mathbf{c_1}/\mathbf{p_1}$).

Finally, we plot the level of cooperation obtained for a monitoring-punishing institution in Figure 4, again for $\rho = 1/3$ and $\rho = 3$.

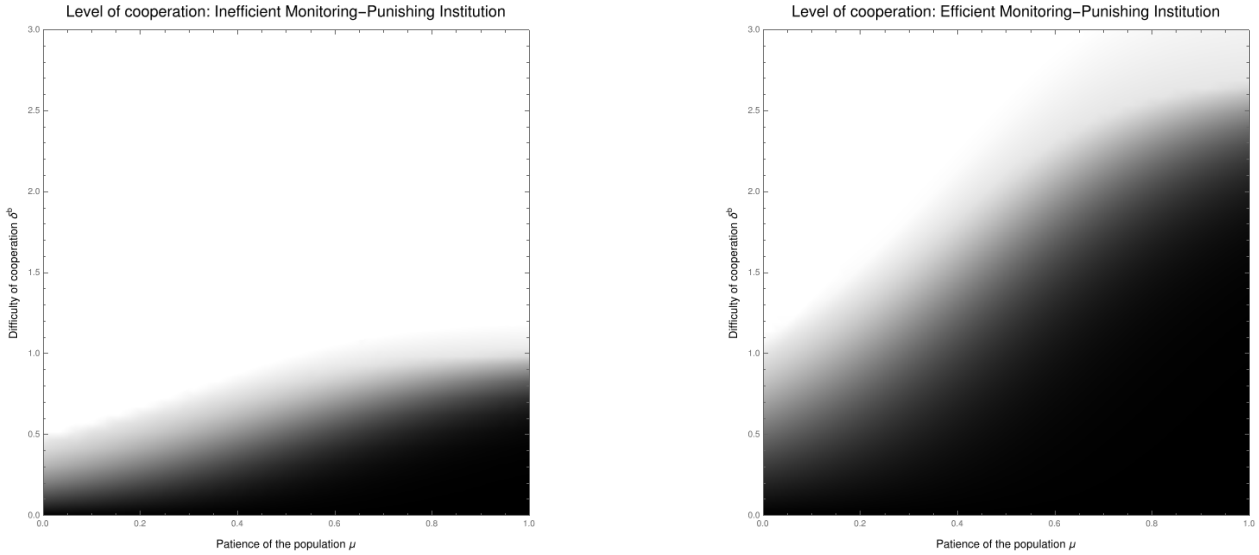


Figure 4: Level of cooperation in the institution equilibrium for a monitoring-punishing institution, as a function of μ and δ^b. Left: $\rho = 1/3$ (inefficient institution); right: $\rho = 3$ (efficient institution).

S5.3.2 Comparison between the monitoring-punishing institution and no institution

Increase in the level of cooperation. In each case, the level of cooperation is higher in the institution equilibrium than it is in the baseline equilibrium—and the difference is starker when the institution is more efficient (high ρ). To illustrate the effect of an institution on cooperation, we subtract the level of cooperation in the baseline equilibrium to the level of cooperation in the institution equilibrium in the case of a monitoring-punishing institution. We plot the resulting increase in the level of cooperation in Figure 5, for $\rho = 1/3$ (left) and $\rho = 3$ (right).

This time, the shade of gray indicates the increase in the level of cooperation at a given point: black indicates an increase of 1 (hence that the level of cooperation must be 1 in the institution equilibrium and 0 in the baseline equilibrium), and white indicates an increase of 0.

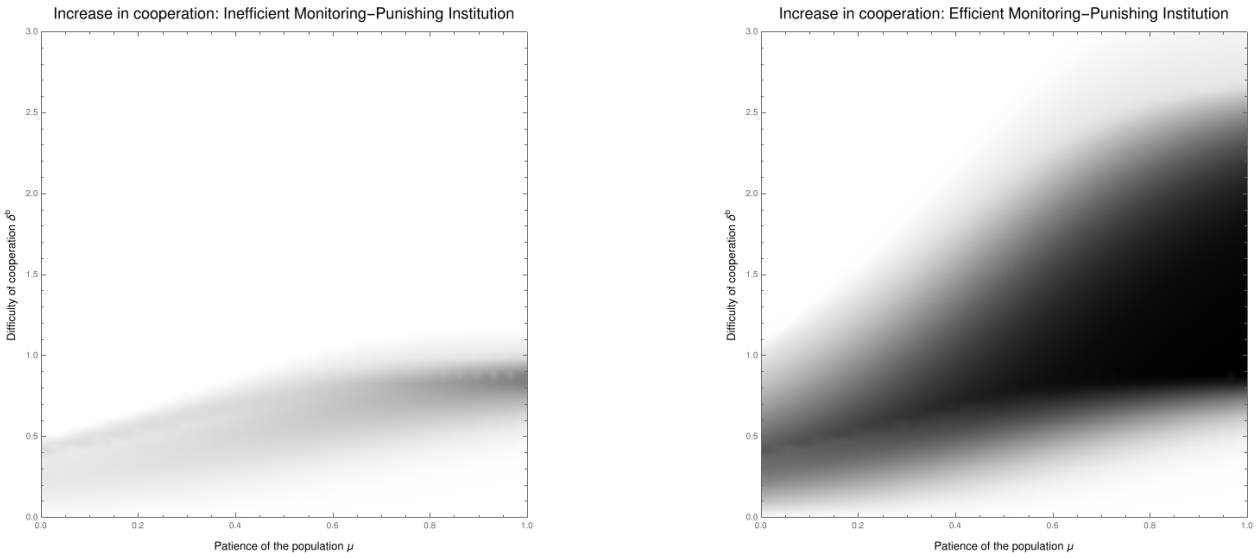


Figure 5: Increase in the level of cooperation due to the institution, as a function of μ and δ^b. To compute this value, we subtract results plotted in Figure 2 to those plotted in Figure 4 (monitoring-punishing institution), for $\rho = 1/3$ (left) and $\rho = 3$ (right).

Change in expected actor payoff ($\rho = 3$). Even an efficient institution (right of Figure 5) leads to only a marginal increase in the level of cooperation when δ^b is low and μ is high—in such a case, the level of cooperation is already high without an institution.

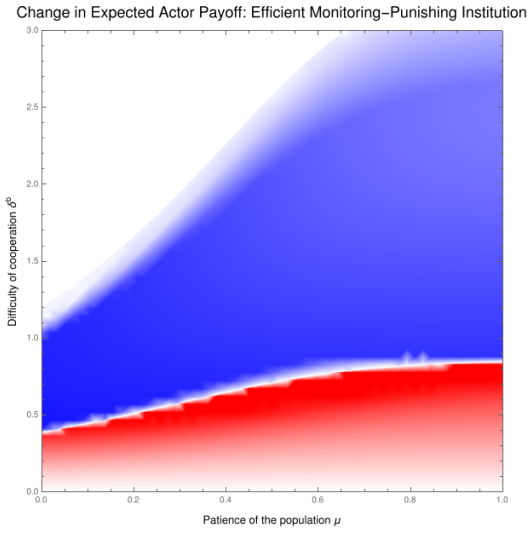


Figure 6: Change in actor expected payoff as a function of μ and δ^b, in the case of the monitoring-punishing institution for $\rho = 3$. Blue: increase, on a scale of 0 to 1, i.e. 100% of the maximum value qr. Red: decrease in absolute value, on a scale of 0 to 0.05.

The institution may in fact lead to a decrease in actors' payoffs in a similar parameter space, as actors then pay the cost of second-order cooperation c_2 in a context where costly enforcement of first-order cooperation is largely unnecessary. To illustrate this, we subtract the actor expected payoff in the baseline equilibrium to the actor expected payoff in the institution equilibrium, for $\rho = 3$, in the case of the monitoring-punishing institution.

We compute the actor expected payoff by taking the expected value of the average lifetime payoff of the actor (calculated above) in the relevant equilibrium, knowing the distribution of discount factors (i.e. knowing the value of μ), and normalize it (by dividing by qr). We plot our results in Figure 6, again as a function of μ and δ^b.

Shades of blue indicate an increase in actor expected payoff: dark blue indicates an increase of 1, and white an increase of 0. Shades of red indicate a decrease in actor expected payoff: dark red indicates an increase of 0.05 or more. The maximum decrease is small—just over 0.06 with our parameters. Note that with our assumptions, the cost of cooperation $c_2 = c_1/3$ is very small when c_1 is small and μ high, i.e. in those points of the parameter space

in which the institution appears unnecessary.

Change in expected payoff ($\rho = 3$). Finally, we carry out the same computation for the expected payoff; that is, the expected payoff of a random individual, by averaging between the actor expected payoff used above and the long-run payoff of the chooser (calculated in the previous section; we normalize it by dividing by $(b-k)$). Note that we weigh the actor expected payoff by $1/(1+q)$ and the chooser long-run payoff by $q/(1+q)$ to capture the fact that the actor plays in both games and the chooser only in the trust game.

Since the level of cooperation can only increase, the long-run payoff of the chooser can only increase in the institution equilibrium as compared to the baseline equilibrium.

We plot the result in Figure 7. In regions in which the actor expected payoff decreases, this decrease is partially compensated by an increase in chooser long-run payoff. As a result, the maximum net decrease for the expected payoff is just under 0.01, or 1% of the maximum value. We have to even further shrink the red scale in order to see decreases—dark red now indicates a decrease of 0.01 or more. As before, dark blue indicates an increase of 1.

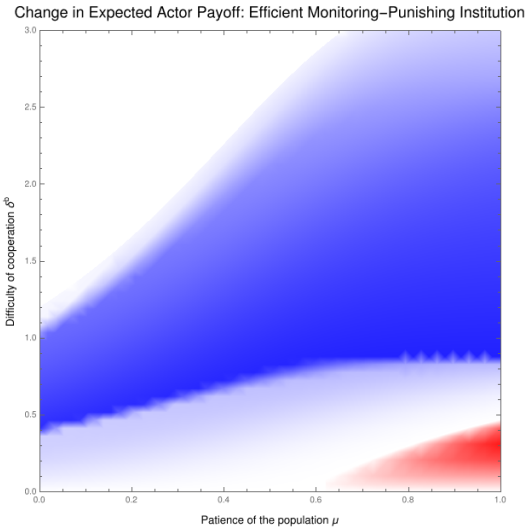


Figure 7: Change in expected payoff as a function of μ and δ^b, in the case of the monitoring-punishing institution for $\rho = 3$. Blue: increase, on a scale of 0 to 1, i.e. 100% of the maximum value qr. Red: decrease in absolute value, on a scale of 0 to 0.01.

Commentary on Glowacki (forthcoming). The evolution of peace.
Behavioral and Brain Sciences

Peace is a form of cooperation, and so are the cultural technologies which make peace possible

Julien Lie-Panis & Jean-Baptiste André
Institut Jean Nicod, Paris

Correspondence: Julien Lie-Panis, jliep@protonmail.com

Abstract While necessary parts of the puzzle, cultural technologies are insufficient to explain peace. They are a form of second-order cooperation---a cooperative interaction designed to incentivize first-order cooperation. We propose an explanation for peace-making cultural technologies, and therefore peace, based on the reputational incentives for second-order cooperation

This is an insightful analysis of the evolution of peace, using the lens of game theory. We propose to complement it, by exposing the cooperative dilemma underlying peace-making cultural technologies. While necessary parts of the puzzle, cultural technologies are insufficient to explain peace—they replace one cooperative dilemma with another. We propose a solution based on prosocial reputation. Cultural technologies, such as informal leadership, may be designed to amplify reputational incentives—in which case they replace a difficult cooperative dilemma with one which is easier. This is not just theoretical nitpicking. Taken together, the author's account and our complement can generate testable predictions regarding the conditions under which peace-making cultural technologies, and therefore peace, may evolve.

As the author rightfully points out, peace is the solution to a cooperative dilemma. In small-scale societies as well as in decentralized urban gangs, war, like defection, exacts a toll on the entire group; yet it is beneficial for certain individuals. If nothing keeps these individuals in check, war is the only Nash equilibrium.

Implicit in this account however, is that peace cannot be explained by reputation—or other canonical explanations for cooperation, such as kin altruism (Hamilton, 1963) and reciprocity (Axelrod & Hamilton,1981). In the iterated prisoner's dilemma that the author considers, cooperation is a Nash equilibrium when the benefit of a prosocial reputation exceeds the temptation to cheat (Nowak & Sigmund, 1998; Panchanathan & Boyd, 2003). War ends up being the only Nash equilibrium because certain individuals find it beneficial to cheat *even* when considering the reputational cost of deviating from peaceful behavior. In other words, peace can be characterized as the solution to a *hard-to-solve* cooperative dilemma—a cooperative dilemma for which reputation provides insufficient incentives.

To achieve peace, humans need to create additional incentives. The author rightfully insists on the central role played by cultural technologies—norms, social structures, mechanisms and institutions, which change the underlying incentive structure (North, 1990; Ostrom, 1990; Powers, Schaick, & Lehmann 2016; Henrich & Muthukrishna, 2021). Humans rely on cultural technology to change the rules of the game, and invent peace. To quote the author, peace becomes a possible solution when "decentralized societies develop internal social structures, including age or status groups, or informal but powerful leadership".

Yet, the author does not mention that cultural technologies are themselves the solution to a cooperative dilemma. Age, status groups, and informal leaders need not necessarily work towards the objectives of the group. Instead, they can advance their own objectives. As the author acknowledges, even though they often promote cooperation within the group (Garfield, Syme, & Hagen, 2020), e.g. by working towards peace (Fry et al., 2021; Glowacki & Gonc, 2013), informal leaders sometimes use their power and influence to promote their self-interest at the expense of the collective (Singh, Wrangham, & Glowacki, 2017).

Cultural technologies are a form of *second-order cooperation* —a cooperative interaction aimed at promoting cooperation (Yamagishi, 1986; Ostrom, 1990; Persson, Rothstein, & Teorell, 2013). In and of themselves, they are insufficient to explain peace. Cultural technologies allow humans to solve the first-order cooperative dilemma. Yet, they introduce another, second-order cooperative dilemma in its place. It seems we are back to square one.

Our solution is to view cultural technologies as technologies specifically designed to leverage reputation. Cultural technologies need not lead us back to our starting point, because second-order cooperation need not be as hard-to-solve as first-order cooperation. Humans can design cultural technologies which: (i) provide sufficient incentives for the hard-to-solve cooperative dilemma, and (ii) are themselves underlain by an *easy-to-solve* cooperative dilemma, that *can* be stabilized by reputation. When this is the case, cultural technologies (and reputation) are sufficient to explain peace (see Figure).

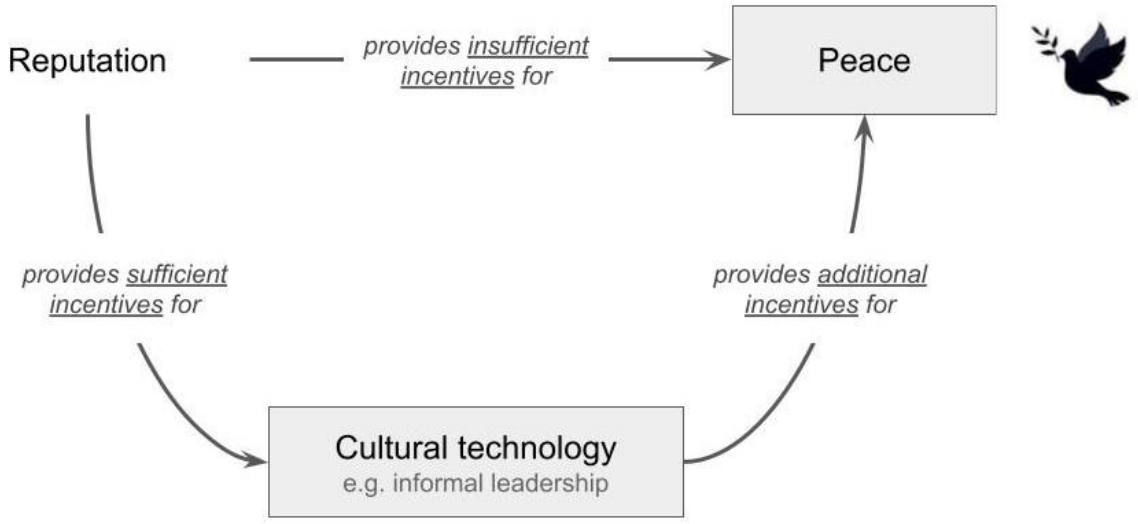


Figure: An explanation for peace through cultural technology.

Informal leaders, for instance, seem decidedly incentivized by reputation. Across small-scale societies, leadership is associated with social status and prestige (Garfield et al., 2021). Leaders tend to enjoy high social capital (Glowacki & von Rueden, 2015), and high social and material

benefits (Garfield et al., 2020; Gurven et al., 2000; Sugiyama, 2004; von Rueden, 2014). Leaders have a lot to lose by defecting. If they cheat, and promote self-serving warfare at the expense of the collective, they stand to lose their very position, and all its accompanying benefits.

In line with the author's account, there is nothing specific about peace or peace-making cultural technologies. Cultural technologies allow humans to scale up cooperation—beyond the limited scope of what can be achieved with reputation alone. Our complement further clarifies the "ironic" logic of peace uncovered by the author. Peace with another group is just one instance of large-scale cooperation. War along that group against another coalition is another such instance. Both depend on the ability to stabilize cultural technologies, that is to solve a second-order cooperative dilemma.

We can derive testable predictions from this idea. Cooperation is not infinitely scalable, because second-order cooperation cannot be made infinitely cheap and still provide sufficient incentives for first-order cooperation. We expect higher ability to establish peace-making cultural technologies, and therefore peace, when individuals have a stronger incentive to invest in their prosocial reputation—e.g. in longstanding communities, in which the shadow of the future looms large (Axelrod & Hamilton, 1981; Ostrom 1990), or in contexts of material security, in which individual's immediate needs are already met (Lie-Panis & André 2022, Mell, Baumard & André 2021).

References

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, *211*(4489), 1390-1396

Fry, D. P., Souillac, G., Liebovitch, L., Coleman, P. T., Agan, K., Nicholson-Cox, E., ... & Strauss, S. (2021). Societies within peace systems avoid war and build positive intergroup relationships. *Humanities and Social Sciences Communications*, *8*(1)

Garfield, Z. H., Schacht, R., Post, E. R., Ingram, D., Uehling, A., & Macfarlan, S. J. (2021). The content and structure of reputation domains across human societies: a view from the evolutionary social sciences. *Philosophical Transactions of the Royal Society B*, *376*(1838), 20200296

Garfield, Z. H., Syme, K. L., & Hagen, E. H. (2020). Universal and variable leadership dimensions across human societies. *Evolution and Human Behavior*, *41*(5), 397-414

Glowacki, L., & Gonc, K. (2013). Customary institutions and traditions in pastoralist societies: neglected potential for conflict resolution. *conflict trends*, *2013*(1), 26-32

Glowacki, L., & von Rueden, C. (2015). Leadership solves collective action problems in small-scale societies. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1683), 20150010

Gurven, M., Allen-Arave, W., Hill, K., & Hurtado, M. (2000). “It's a wonderful life”: signaling generosity among the Ache of Paraguay. *Evolution and Human Behavior*, *21*(4), 263-282.

Hamilton, W. D. (1963). The evolution of altruistic behavior. *The American Naturalist*, *97*(896), 354-356

Henrich, J., & Muthukrishna, M. (2021). The origins and psychology of human cooperation. *Annual Review of Psychology*, *72*, 207-240

Lie-Panis, J., & André, J. B. (2022). Cooperation as a signal of time preferences. *Proceedings of the Royal Society B*, *289*(1973), 20212266

Mell, H., Baumard, N., & André, J. B. (2021). Time is money. Waiting costs explain why selection favors steeper time discounting in deprived environments. *Evolution and Human Behavior*, *42*(4), 379-387

North, D. C. (1991). Institutions. *Journal of economic perspectives*, *5*(1), 97-112

Nowak, M. A., & Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. *Nature*, *393*(6685), 573-577

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

Panchanathan, K., & Boyd, R. (2003). A tale of two defectors: the importance of standing for evolution of indirect reciprocity. *Journal of theoretical biology*, *224*(1), 115-126

Persson, A., Rothstein, B., & Teorell, J. (2013). Why anticorruption reforms fail—systemic corruption as a collective action problem. *Governance*, *26*(3), 449-471

Powers, S. T., Van Schaik, C. P., & Lehmann, L. (2016). How institutions shaped the last major evolutionary transition to large-scale human societies. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1687), 20150098

von Rueden, C. (2014). The roots and fruits of social status in small-scale human societies. *The psychology of social status*, 179-200

Singh, M., Wrangham, R., & Glowacki, L. (2017). Self-interest and the design of rules. *Human Nature*, *28*, 457-480

Sugiyama, L. S. (2004). Illness, injury, and disability among Shiwiar forager‐horticulturalists: Implications of health‐risk buffering for the evolution of human life history. *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists*, *123*(4), 371-389

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and social Psychology*, *51*(1), 110

Chapter 6

Discussion

All social science research must do some violence to reality in order to reveal simple truths.

(Lazer & Friedman, 2007)

Simple models can be invaluable without being "right," in an engineering sense. Indeed, by such lights, all the best models are wrong. But they are fruitfully wrong. They are illuminating abstractions.

(Epstein, 2008)

An old adage holds that it is better to stay silent and be thought a fool than to speak and remove all doubt. As scientists, our goal is not to save face, but in fact to remove as much doubt as possible. Formal models make their assumptions explicit, and in doing so, we risk exposing our foolishness to the world. This appears to be the price of seeking knowledge. Models are stupid, but perhaps they can help us to become smarter. We need more of them.

(Smaldino, 2017)

6.1 Build on previous work . . . 171
6.2 Use analogies . . . 172
6.3 Embrace your passion for stupid models . . . 173
6.4 Seek smart friends (from other disciplines) . . . 175

I opened this dissertation with a semi-fictitious conversation. I tell people at a cocktail party that my work involves explaining why we, as humans, cooperate so much. They're intrigued. They might disagree with the premise (Have you seen the state of the world?), but they'll have lots of their own theories—or rather, their own, non-mathematical models (Epstein, 2008)—to contribute to what promises to be a lively conversation. I then proceed to mention game theory, and the conversation immediately dies out (Yes, I'm quite fun at parties).

This story makes two omissions. First omission: I used to lead with biology. I have learned that this is neither the most productive way to start a conversation, nor, I believe, one that gives an accurate vision of my work—I'll come back to these issues in section 6.4.2.

I've also omitted the many people who aren't rebuffed by mathematics. Many of my acquaintances are quite happy to entertain the idea of mathematical modeling as a valid mode of inquiry—some of them encounter models in their day to day, or were even trained in mathematics, as I was. What's always struck me is that, even with the same exact training, we have very different ideas of what a mathematical model should look like. What does your model predict? Have you thought about adding this variable? OK, this is a nice sounding idea, but when are you going to fit your model to some actual data?

As the opening quotes suggest, I'm not the first in my brand of mathematical modeling to encounter this issue. Models are 'wrong', if not 'stupid', and, like all other methods of inquiry, they do 'violence to reality'. More importantly, they're designed that way on purpose. Their purpose is to be 'fruitfully wrong'—to reveal the simple truth hiding behind a class of complex phenomena (Epstein, 2008). In that respect, the stupider the model, the better.

This fruitful stupidity is what I like the most about game theory. But I'm probably biased. For whatever reason, I have both: (a) a proclivity for elegant and simple math, and (b) a deep-rooted curiosity for the social sciences—bordering on an unhealthy admiration for those in it. I've been attracted to stupid models ever since my formative years in a French 'classe préparatoire'.[1] And I've been trying to hijack my way into the social sciences ever since leaving that program, for more years that I care to mention (for a long time, I thought this meant having to abandon math).

The following contains a few thoughts on the other purposes of stupid models, and making them more useful—but not less stupid.[2] These thoughts are not meant to be exhaustive or original (many others have covered the merits of stupid models).[3] These thoughts are

[1] This is an intensive undergraduate program (for me, focusing in math), where 30-40 students spend 2-3 years preparing for competitive exams. I owe a lot to my many great teachers, notably Nicolas Choquet and his love of math and weird puns (specifically, spoonerisms), and the many geeky friends who continue to accompany me.

[2] I'm thinking of the saying, 'All models are wrong, but some are useful', which is famously attributed to the statistician George Box (1979).

[3] Besides the references quoted at the beginning of this discussion (the quote from Lazer & Friendman is taken from Smaldino 2017), I've taken insights from a short blog series by Tiokhin, 2021, and a book by M. Hoffman and Yoeli, 2022. All of these references are quite accessible—I haven't ventured into uncharted territory (i.e., for me, something more philosophy-of-science-like), although I'm happy to take a recommendation from anyone reading this footnote.

also relatively personal; they might only apply if, like me, you have both traits (a) and (b). I've woven them together with elements from my PhD experience, including the scientific questions I addressed in the previous chapters, and the scientists whose thoughts I've encountered during this whole process, whether in verbal or written form.

6.1 Build on previous work

6.1.1 Make the most of reputation (as others have understood it)

Originally, I wanted to call this dissertation 'Making the most of reputation'. Although this might just be because of its vagueness—thankfully, my advisor, Jean-Baptiste André, caught this vagueness in time—I think this title/catchphrase encompasses the various projects I've worked on during my PhD, and its general approach. I've learned to make the most of reputation; that is, of the ways in which game theory typically apprehends reputation. To do so, my advisors and I have developed simple models using reciprocity, signaling, or both.

Chapter 3 shows that a lot can be done with reciprocity alone. Based on a repeated game and general principles of game theory, we are able to derive qualitative predictions about revenge. Strategies that are dominated in the stage game can only occur when they are incentivized in the repeated game: therefore, revenge must serve to deter future transgressions. Coordination depends on common knowledge (for a general demonstration, see Yoeli et al., 2022): therefore, revenge must be conditioned on publicly shared information (e.g., was someone hurt?), and not on private observation (e.g., what was the intent?).

In the other chapters, I relied on signaling, to ground my models in reputation-based decisions—in particular, the decision to trust (chapters 2, 5). There is nothing particularly new about this approach either. Amotz Zahavi, the founder of the evolutionary approach to honest signaling, proposed applying his framework to prosocial behavior (Zahavi, 1995). Whether or not they came to this idea through the prism of honest signaling, modelers studying the evolution of cooperation integrated trust and partner choice early on (Gintis et al., 2001; Leimar, 1997; Sherratt and Roberts, 1998). Much more recently, many models of cooperation, including the one presented in chapter 2, have explicitly relied on a signaling framework. These models help us understand the building blocks of trust and our cooperative reputations—such as stake in a partner (Barclay et al., 2021), lack of outside options (Quillien, 2020), or ability to invest in one's reputation (Lie-Panis & André, 2022).

6.1.2 Extend your own work (to extend reputation)

simple models provide a basis for further inquiry (Smaldino, 2017). Once we have a potential building block of reputation, it is possible, and tempting, to extend to more than one interaction. This is an area where signaling is very useful—or other, newer ways to model the inferences we make (see section 6.4.1). Although it is possible to study a form of cross-contextual reciprocity (Donahue et al., 2020, see also: Panchanathan and Boyd, 2004), in which individuals learn to coordinate behavior across games for mutual benefit, it is more

satisfying to ground this in cross-contextual inferences.

When can we trust in one context based on cooperation in another? When can we trust someone who punishes antisocial behavior, and what happens if we have more pertinent information (Jordan et al., 2016)? What about behavior in more competitive realms: when might we prefer a vicious friend (Krems et al., 2023)? What about other pieces of information not directly tied to behavior such as group membership (Arai et al., 2023; Imada et al., 2023)?

In real life, cooperation and reputation extend far beyond the simple give and take of a repeated game. Chapter 4 provides an answer to the first question based on time preferences: when contribution to a public good is sufficiently costly, it reveals high preference for the future, and the ability to also pay the cost of dyadic help to achieve the same reputational benefits—Barclay & Barker (2020) provide similar answers using two models, one with stake and the other with time preferences. While these models are very far from answering the many other questions we could ask (I stopped myself from adding other questions above), they illustrate the usefulness of signaling as a framework for studying reputation across different interactions.

6.2 Use analogies

Chapter 5 suggests that my previous catchphrase can be used to understand institutions. In our model, institutions are technologies for transforming contributions to a public good into incentives for cooperation in another context (e.g., into monitoring of behaviors in a trust game). When contributing is cheap, reputation is enough to stabilize the institution; when the institution is efficient in producing incentives, expensive forms of cooperation become possible. In other words, an efficient institution allows individual to make the most of reputation—to build the most mutually beneficial social organization that can be sustained by reputation alone. Just as a pulley system helps lift heavy loads with minimal effort, well-designed institutions may maximize the potential of limited reputational incentives, helping humans achieve extended levels of cooperation.

I really like this analogy. Maybe it's just me (I did study a bit of physics), but I can really see the machine. Some reputation and some cooperation enter, a few gears grind, and incentives and a lot more cooperation exit. It's not my idea. The general idea—investigating complex social phenomena, such as institutions, religion (Fitouchi & Singh, 2022) or nationalism (Sijilmassi, 2021), through the prism of social technology—owes entirely to Jean-Baptiste André and Nicolas Baumard and their inspirations (e.g., Lienard, 2016). It permeates the thinking of many of the lab's current students (two examples are cited just above). The back and forth between model, writing and discussions with my coauthors (in particular, Léo Fitouchi) that come with the research process have honed this analogy in my mind, helping me see a facet of institutions.

It's tempting to extend the analogy—in fact, that's one of the purposes of mathematical models (Epstein, 2008). Like contributions to the institution in our model, truthful third-party information about others' cooperative behavior (honest gossip) is a public good,

which extends the possibility of cooperation (Giardini, Balliet, et al., 2021). It is fragile, as individuals may be tempted to lie for their own interests (e.g., to hurt a rival Hess, 2016). Tentatively, a similar two game framework to the one we developed could be applied to gossip. This could explain, for instance, why cooperators tend to also tell the truth (Giardini, Vilone, et al., 2021)—relatedly, Wu et al. (2021) show that honest gossip is optimal when the gossiper has a stake in the receiver's welfare. The analogy with technology could then help understand other things that humans do in order to extend their cooperative groups (e.g., build age sets or other nested groupings Lienard, 2016).

6.3 Embrace your passion for stupid models

6.3.1 A neurotic interlude

You might reasonably think the model presented in chapter 2 superfluous. In 1971 already, Robert Trivers, the biologist credited for the idea of reciprocal cooperation (he called it reciprocal altruism) had highlighted that it came with a present-future trade-off. According to him (p. 39), reciprocal cooperation "can be viewed as a symbiosis, each partner helping the other while he helps himself. The symbiosis has a time lag, however; one partner helps the other and must then wait a period of time before he is helped in turn."

From there, it's logical that patient individuals should be more motivated by reciprocal cooperation. Even more worrying for chapter 2, time preferences can be directly deduced from Axelrod and Hamilton's 1981 model. All you need to do is replace the probability of repetition r with a discount factor δ. If you assume that interactions are repeated with certainty but that individuals discount the future according to δ (such that a payoff in two rounds is worth δ^2 in this round), then you get an analogous result to the one we proved in section 1.2: cooperation is possible when δ is greater than c/b, the cost to benefit ratio of cooperation.

Thankfully, this does not make our stupid model useless—and I have Epstein (2008), Smaldino (2017) and Tiokhin (2021) to back me up (I'll use them in a second). Yes, reciprocal cooperation entails a present-future trade-off—this isn't a new insight. So will reputation-based cooperation for that matter. Yes, if you assume that some individuals are patient and others impatient, then, all else being equal, the patient ones will more often be able to cooperate. Unless you make a mistake, your model's results will always follow from your assumptions.[4] Yes, our model is basically Axelrod and Hamilton plus different partners plus signaling (or alternatively, Leimar, 1997 with a different quality). But this does not mean the model is useless.

Even when a model isn't particularly illuminating, it's still a useful exercise. Formal models are just part of the scientific process. 'Patient people cooperate more with strangers'. OK, nice theory, let's formalize it. What do you mean by patience? The ability to pass the Marshmallow test? A life-long struggle against one's earthly passions, to elevate the soul and move closer to God? Low levels of serotonin signaling? Just like any experiment,

[4]Like many before me, I was reproached this in review.

formal models force us to make explicit assumptions, and remove doubt from our theories (Smaldino, 2017). What is the minimum set of assumptions that is needed to generate patient cooperation and impatient defection? Just like simple experiments conducted in completely unrealistic settings (the lab), stupid models create a controlled world with no unnecessary details to demonstrate a conceptual point (Tiokhin, 2021)—we might then reasonably take issue with the point, but to do that we first need to understand it. And, just like experimentation, modeling stems from a position of methodological doubt (Epstein, 2008). Will time preferences (an individual trait) follow from repetition probability (a feature of the relationship or social group)? Probably, but to be sure and learn new things, let's try it out.

6.3.2 Another analogy

I'm tempted to use another analogy, in a strange attempt to justify designing the model based on the model itself (bear with me). Like cooperation, research, starting with PhD training, also involves present costs. It takes a lot of time to learn the craft, based on the work of those who preceded you (section 6.1). When you start a PhD with lofty interdisciplinary goals like I did, the costs of investing into your training can seem exacerbated: all the necessary skills and practical knowledge are contained within the confines of a discipline (Nettle, 2018, chapter 11). Before you can declare disciplines dead and pave the way for a new era of science, you need to immerse yourself in the technical debates of your disciplinary predecessors (plus, you might have to dive into work from other disciplines).

As with cooperation, the benefits of research are (mostly) social and faraway. Research does not offer much in terms of immediate, material rewards; but research offers plenty in terms of later social benefits—we will meet plenty of interesting people, and we may hope that one day they will find us interesting too.

With this incentive structure, it pays to be (very) patient. In our model, patient individuals cooperate even when rewards are many rounds away (e.g., because on average it takes many rounds to be observed). These individuals also cooperate when costs are high and rewards low; all these factors affect the trade-off of cooperation. Patient individuals, we argue in the discussion, can then be seen as individuals with high intrinsic prosocial motivation (Benabou & Tirole, 2003)—as individuals who will help others in a large chunk of the parameter space, and that we, as onlookers, may reasonably treat as genuinely prosocial.

How does this justify designing the model? Well, what I'm saying is that, unsurprisingly, it pays to be passionate (for a discussion of the costs and benefits of passion, see M. Hoffman & Yoeli, 2022, chapter 15—I'll come back to this in section 6.4.2 as well). And (here's the leap), it pays to recognize that passion. For me at least (and not everybody needs to be passionate about this line of work), recognizing my passion, and the privilege I have to be working in a job I'm passionate about has helped me adjust my trade-offs. It has helped me see the joy and near-null costs associated with time spent learning—among other things by reminding me of the alternatives (I've tried jobs with less passion and more money). It has helped me set limits, by recognizing the tasks in which I don't need to invest much time. It has helped me engage in a sort of conscious self-deception, whereby instead

of giving huge value to the uncertain future, I chose to completely ignore it.[5]

6.4 Seek smart friends (from other disciplines)

6.4.1 What is the discipline of game theory?

I started this dissertation by distinguishing between economic and evolutionary game theory. Game theory was invented by mathematicians, and used by economists and political scientists to study interactions between firms and governments. It relied on strong assumptions of individual rationality (economic game theory). This was all and good until these scientists decided to extend the study to people, whose decisions processes very clearly didn't follow the strong assumptions. Thankfully, biologists stepped in, and showed that these unrealistic assumptions were unnecessary (evolutionary game theory).

Of course, this story is simplistic. Economists have long recognized two interpretations of the Nash equilibrium (Osborne & Rubinstein, 1994) (these interpretations can be traced to Nash himself). On the one hand, the deductive interpretation is the one we've relied on the most: a few rational individuals engage in a specific interaction, and solve the game in their heads in the same manner that we do on paper. On the other hand, we engaged with the steady state interpretation when we defined games: we defined them as models for classes of similar situations, not as isolated interactions. If we had continued with the steady state interpretation, we would have said that individuals form expectations based on past experiences with similar situations (they "know" the equilibrium, and therefore "know" what to expect from others), and behave so as to increase their payoffs given their expectations. Over time, people should converge towards a Nash equilibrium—this interpretation paves the way for the evolutionary interpretation, in which no knowledge is required.

More largely, the disciplines of game theory do not operate in isolation. The framework (Mailath & Samuelson, 2006) I relied on in chapters 3 and 5 was originally developed by economists to study the reputation of firms. These economists use it to explain why firms may wish to sell better products as a matter of principle. Others have borrowed the framework to explain the psychology of principled behavior (Singh & Hoffman, 2021). Conversely, economists are aware of critiques coming from psychology or the evolutionary sciences. They have long incorporated our cognitive biases into their models, discovering new ways to model utility and preferences. More recently, some economists have advocated for an evolutionary approach, in order to explain these biases (Page, 2022).

I think therein lies a fundamental strength of game theory: it doesn't belong to a single discipline. Of course, this can be a weakness, particularly while the training wheels are still on. More than once, in an interdisciplinary lab with little math and lots of freedom, I've felt the need for stronger disciplinary training (to paraphrase the title of an essay on interdisciplinarity in the sciences: Nettle, 2018, chapter 11, see also the previous section). But, in an interdisciplinary lab with little math and lots of freedom, it is easy to find exciting

[5] I'm told that there there is a link between patience and Triver's theory of self-deception but I have not found a reference for that.

and fruitful collaborations, with scientists used to working with modelers and evolving at the same level of abstraction.

Under such conditions, game theory and interdisciplinary collaborations mesh really well. These collaborations can produce models that address useful questions for the social sciences, by relying on simplistic but sensible assumptions—when the natural inclination for a mathematician is to seek technical solutions to technical problems. These collaborations can help investigate a problem via different methods, and, when they involve controlled experiments as well, take advantage of the similarities between experiments and models (Tiokhin, 2021, part 2) to hone both methods, as well as the research question. Once a model has revealed a candidate to a simple truth, it is easier to design relevant experimental conditions, and test the theory—and then, depending on the result, refine the model or the theory.

6.4.2 Can we dispense with biology?

To conclude, I'll return to my fake conversation. As I mentioned, I used to lead with biology. I used to tell people about ultimate explanations, and illustrate the problem of cooperation from a fitness perspective: it's costly to your genes, you see? This was not the right approach, and led me frequently feeling misunderstood (for a discussion of why that might be, see Nettle, 2023).

Does this mean we should dispense with evolution? The short answer is of course no. The evolutionary interpretation of game theory is parsimonious and productive. By looking at things through the evolutionary lens, we can 'rescue' game theory without having to make strong assumptions (section 1.2). We can even improve the predictive power of game theory by grounding our models in a better understanding of biological evolution, so as to restrict our analysis to a more realistic set of possible endpoints (André, 2023). Ultimately, evolution is the appropriate framework for understanding psychological function: for theoretical scientists interested in the big why questions, explaining cooperation by assuming that individuals have prosocial preferences is unsatisfying, to say the least (Page, 2022).

The slightly longer answer is also no, but it takes slightly longer to get there (two paragraphs instead of one). A lot of the things we, as theoretical social scientists, are interested in do not require evolution in first approximation. Optimization through learning processes rather than evolution also leads to evolutionary endpoints (M. Hoffman & Yoeli, 2022). If you spend your formative years surrounded by people who value complex problems in algebraic topology, you will probably learn to like abstract mathematics with a passion. But you might gradually lose your passion after exiting the mathematical bubble—in particular, if you find other things you're better at, or if you're pushed towards a less ethereal pursuit because you have to make ends meet.

Social learning can then account for the speed at which passions are gained and lost, and allows us to make a prediction: you should be more passionate about what others around you value. Of course, we can do better. How can we also account for the role played by other pursuits, and material needs? To do that, we need to understand the function of passion—and for that, the appropriate framework is, once again, evolution. An evolutionary account

suggests that the function of passions is to motivate individuals to invest time and effort into becoming better at pursuits that are likely to bring future rewards (M. Hoffman & Yoeli, 2022, chapter 15). This explains why others' opinion matters: it's better to be good at things that are valued. It also explains why being good at other valued pursuits and having more pressing material needs similarly have a similar effect: both increase the cost of investing in this specific pursuit.

It would be quite strange to conclude a dissertation in evolutionary game theory by striking out the evolutionary. Evolution uses game theory to its full potential, and can provide functional answers to the very abstract questions I'm interested in. These questions can be related to any of the social sciences. For that purpose, simple models and simple analogies (section 6.2) are powerful tools: they aid trans-disciplinary communication by conveying clear messages (for those social scientists who adopt a similarly abstract level of analysis).

I just worry sometimes that biology may disrupt that communication, and prevent us from using the full interdisciplinary potential of game theory. I am interested in evolution for the lens it provides—not for its own sake. Leading with evolution might send the wrong message about my interests, when the same model with learning might do just fine (as I tried to convey with my passions analogy above). Conversely, even though I am interested in ultimate abstractions, it would be a shame to miss out on all the work that has already been conducted elsewhere, simply because it relies on more proximate abstractions (e.g., preferences). Many of the more innovate ideas in game theory—ideas that go beyond the traditional frameworks I've explored during my PhD—are coming from trans-disciplinary locations. At the frontier between evolution and economics, models look at the evolution of preferences (Alger, 2023), and the evolution of reference points and loss aversion (Kubitz & Page, n.d.). At the frontier with the cognitive sciences, models look at the evolution of theory of mind, the ability to impute mental states to others (Kleiman-Weiner, 2018; Qi & Vul, 2022). To do so, they go beyond inferences about behavior, to look at the ability to infer elements of a partner's utility function. This requires going beyond the stable social environment of repeated games (and beyond other more 'traditional' refinements, e.g., Hilbe et al., 2018), to look at situations in which the game changes in each time step, in a completely unpredictable manner.

Bibliography

Abeler, J., Calaki, J., Andree, K., & Basek, C. (2010). The power of apology. *Economics Letters*, *107*(2), 233–235. https://doi.org/10.1016/j.econlet.2010.01.033

Alexander, R. D. (1987). *The biology of moral systems*. Transaction Publishers.

Alger, I. (2023). Evolutionarily stable preferences. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *378*(1876), 20210505. https://doi.org/10.1098/rstb.2021.0505

Anderson, R. A., Crockett, M. J., & Pizarro, D. A. (2020). A theory of moral praise. *Trends in Cognitive Sciences*, *24*(9), 694–703. https://doi.org/10.1016/j.tics.2020.06.008

André, J.-B. (2010). The evolution of reciprocity: Social types or social incentives? *The American Naturalist*, *175*(2), 197–210. https://doi.org/10.1086/649597

André, J.-B. (2023, June 30). *Adaptive parsimony as an evolutionary solution to the equilibrium selection problem* (preprint). Social and Behavioral Sciences. https://doi.org/10.32942/X26K6M

André, J.-B., & Day, T. (2007). Perfect reciprocity is the only evolutionarily stable strategy in the continuous iterated prisoner's dilemma. *Journal of Theoretical Biology*, *247*(1), 11–22. https://doi.org/10.1016/j.jtbi.2007.02.007

Arai, S., Tooby, J., & Cosmides, L. (2023, July 7). *Group as a biological market: Eliminating reputation concern decreases ingroup-favouring cooperation and punishment* (preprint). PsyArXiv. https://doi.org/10.31234/osf.io/yjhw2

Aumann, R., & Brandenburger, A. (1995). Epistemic conditions for nash equilibrium. *Econometrica (1986-1998)*, *63*(5), 1161–1180. Retrieved July 30, 2023, from https://www.proquest.com/docview/214864335/abstract/DF9CFC6415154657PQ/1

Axelrod, R. M. (2006). *The evolution of cooperation* (Rev. ed). Basic Books. (Original work published 1982)

Axelrod, R. M., & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, (4489), 1390–1396.

Balliet, D., Tybur, J. M., & Van Lange, P. A. M. (2017). Functional interdependence theory: An evolutionary account of social situations. *Personality and Social Psychology Review*, *21*(4), 361–388. https://doi.org/10.1177/1088868316657965

Barclay, P. (2013). Strategies for cooperation in biological markets, especially for humans. *Evolution and Human Behavior*, *34*(3), 164–175. https://doi.org/10.1016/j.evolhumbehav.2013.02.002

Barclay, P., & Barker, J. L. (2020). Greener than thou: People who protect the environment are more cooperative, compete to be environmental, and benefit from reputation. *Journal of Environmental Psychology*, *72*, 101441. https://doi.org/10.1016/j.jenvp.2020.101441

Barclay, P., Bliege Bird, R., Roberts, G., & Számadó, S. (2021). Cooperating to show that you care: Costly helping as an honest signal of fitness interdependence. *Philosophical*

Transactions of the Royal Society B: Biological Sciences, *376*(1838), 20200292. https://doi.org/10.1098/rstb.2020.0292

Barker, J. L., Power, E. A., Heap, S., Puurtinen, M., & Sosis, R. (2019). Content, cost, and context: A framework for understanding human signaling systems. *Evolutionary Anthropology: Issues, News, and Reviews*, *28*(2), 86–99. https://doi.org/10.1002/evan.21768

Beekman, G., Bulte, E., & Nillesen, E. (2014). Corruption, investments and contributions to public goods: Experimental evidence from rural liberia. *Journal of Public Economics*, *115*, 37–47. https://doi.org/10.1016/j.jpubeco.2014.04.004

Benabou, R., & Tirole, J. (2003). Intrinsic and extrinsic motivation. *Review of Economic Studies*, *70*(3), 489–520. https://doi.org/10.1111/1467-937X.00253

Bersch, K. (2019, January 31). *When democracies deliver: Governance reform in latin america* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781108559638

Biernaskie, J. M., Perry, J. C., & Grafen, A. (2018). A general model of biological signals, from cues to handicaps. *Evolution Letters*, *2*(3), 201–209. https://doi.org/10.1002/evl3.57

Bliege Bird, R., & Power, E. A. (2015). Prosocial signaling and cooperation among martu hunters. *Evolution and Human Behavior*, *36*(5), 389–397. https://doi.org/10.1016/j.evolhumbehav.2015.02.003

Bliege Bird, R., Ready, E., & Power, E. A. (2018). The social significance of subtle signals. *Nature Human Behaviour*, *2*(7), 452–457. https://doi.org/10.1038/s41562-018-0298-3

Boon-Falleur, M., Baumard, N., & André, J.-B. (2022, June 24). Optimal resource allocation and its consequences on behavioral strategies, personality traits and preferences. https://doi.org/10.31234/osf.io/2r3ef

Box, G. E. P. (1979, January 1). Robustness in the strategy of scientific model building. In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in statistics* (pp. 201–236). Academic Press. https://doi.org/10.1016/B978-0-12-438150-6.50018-2

Boyd, R. (2006). Reciprocity: You have to think different. *Journal of Evolutionary Biology*, *19*(5), 1380–1382. https://doi.org/10.1111/j.1420-9101.2006.01159.x

Carlsmith, Darley, & Robinson. (2002). Why do we punish? deterrence and just deserts as motives for punishment. *Journal of personality and social psychology*, *83*(2). https://doi.org/10.1037/0022-3514.83.2.284

Cimino, A. (2011). The evolution of hazing: Motivational mechanisms and the abuse of newcomers. *Journal of Cognition and Culture*, *11*(3), 241–267. https://doi.org/10.1163/156853711X591242

Currie, T. E., Campenni, M., Flitton, A., Njagi, T., Ontiri, E., Perret, C., & Walker, L. (2021). The cultural evolution and ecology of institutions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1828), 20200047. https://doi.org/10.1098/rstb.2020.0047

Cushman, F., Dreber, A., Wang, Y., & Costa, J. (2009). Accidental outcomes guide punishment in a "trembling hand" game (L. Santos, Ed.). *PLoS ONE*, *4*(8), e6699. https://doi.org/10.1371/journal.pone.0006699

Delton, A. W., Jaeggi, A. V., Lim, J., Sznycer, D., Gurven, M., Robertson, T. E., Sugiyama, L. S., Cosmides, L., & Tooby, J. (2023). Cognitive foundations for helping and harming

others: Making welfare tradeoffs in industrialized and small-scale societies. *Evolution and Human Behavior*. https://doi.org/10.1016/j.evolhumbehav.2023.01.013

Densley, J. A. (2012). Street gang recruitment: Signaling, screening, and selection. *Social Problems*, *59*(3), 301–321. https://doi.org/10.1525/sp.2012.59.3.301

Dessalles, J.-L. (2014). Optimal investment in social signals. *Evolution*, *68*(6), 1640–1650. https://doi.org/10.1111/evo.12378

Dessalles, J.-L. (2018). Self-sacrifice as a social signal. *Behavioral and Brain Sciences*, *41*. https://doi.org/10.1017/S0140525X18001590

Dhaliwal, N. A., Martin, J. W., Barclay, P., & Young, L. L. (2022). Signaling benefits of partner choice decisions. *Journal of Experimental Psychology: General*, *151*(6), 1446–1472. https://doi.org/10.1037/xge0001137

Donahue, K., Hauser, O. P., Nowak, M. A., & Hilbe, C. (2020). Evolving cooperation in multichannel games. *Nature Communications*, *11*(1), 3885. https://doi.org/10.1038/s41467-020-17730-3

Dugatkin, L. A. (1988). Do guppies play TIT FOR TAT during predator inspection visits? *Behavioral Ecology and Sociobiology*, *23*(6), 395–399. https://doi.org/10.1007/BF00303714

Dumas, M., Barker, J. L., & Power, E. A. (2021). When does reputation lie? dynamic feedbacks between costly signals, social capital and social prominence. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1838), 20200298. https://doi.org/10.1098/rstb.2020.0298

Epstein, J. M. (2008). Why model? *Journal of Artificial Societies and Social Simulation*, *11*(4), 12. https://www.jasss.org/11/4/12.html

Fitouchi, L., André, J.-B., & Baumard, N. (2022). Moral disciplining: The cognitive and evolutionary foundations of puritanical morality. *Behavioral and Brain Sciences*, 1–71. https://doi.org/10.1017/S0140525X22002047

Fitouchi, L., & Singh, M. (2022). Supernatural punishment beliefs as cognitively compelling tools of social control. *Current Opinion in Psychology*, *44*, 252–257. https://doi.org/10.1016/j.copsyc.2021.09.022

Fitouchi, L., & Singh, M. (2023). Punitive justice serves to restore reciprocal cooperation in three small-scale societies. *Evolution and Human Behavior*. https://doi.org/10.1016/j.evolhumbehav.2023.03.001

Fudenberg, D., & Maskin, E. (1986). The folk theorem in repeated games with discounting or with incomplete information. *Econometrica*, *54*(3), 533–554. https://doi.org/10.2307/1911307

Fudenberg, D., & Maskin, E. (1990). Evolution and cooperation in noisy repeated games. *The American Economic Review*, *80*(2), 274–279. Retrieved July 11, 2023, from https://www.jstor.org/stable/2006583

Gächter, S., & Falk, A. (2002). Reputation and reciprocity: Consequences for the labour relation. *The Scandinavian Journal of Economics*, *104*(1), 1–26. https://doi.org/10.1111/1467-9442.00269

Gächter, S., & Schulz, J. F. (2016). Intrinsic honesty and the prevalence of rule violations across societies. *Nature*, *531*(7595), 496–499. https://doi.org/10.1038/nature17160

Garfield, Z. H., Syme, K. L., & Hagen, E. H. (2020). Universal and variable leadership dimensions across human societies. *Evolution and Human Behavior*, *41*(5), 397–414. https://doi.org/10.1016/j.evolhumbehav.2020.07.012

Gelfand, M. J., Caluori, N., Jackson, J. C., & Taylor, M. K. (2020). The cultural evolutionary trade-off of ritualistic synchrony. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1805), 20190432. https://doi.org/10.1098/rstb.2019.0432

Geoffroy, F., Baumard, N., & André, J.-B. (2019). Why cooperation is not running away. *Journal of Evolutionary Biology*, *32*(10), 1069–1081. https://doi.org/10.1111/jeb.13508

Giardini, F., Balliet, D., Power, E. A., Számadó, S., & Takács, K. (2021). Four puzzles of reputation-based cooperation. *Human Nature*. https://doi.org/10.1007/s12110-021-09419-3

Giardini, F., Vilone, D., Sánchez, A., & Antonioni, A. (2021). Gossip and competitive altruism support cooperation in a public good game. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1838), 20200303. https://doi.org/10.1098/rstb.2020.0303

Gintis, H., Smith, E. A., & Bowles, S. (2001). Costly signaling and cooperation. *Journal of Theoretical Biology*, *213*(1), 103–119. https://doi.org/10.1006/jtbi.2001.2406

Glowacki, L. (2022). The evolution of peace. *Behavioral and Brain Sciences*, 1–100. https://doi.org/10.1017/S0140525X22002862

Grafen, A. (1991). Modelling in behavioural ecology. In J. Krebs & N. Davies (Eds.), *Behavioural ecology* (3rd, pp. 5–31). Blackwell Scientific Publications.

Grafen, A. (1990). Biological signals as handicaps. *Journal of Theoretical Biology*, *144*(4), 517–546. https://doi.org/10.1016/S0022-5193(05)80088-8

Gutiérrez, N. L., Hilborn, R., & Defeo, O. (2011). Leadership, social capital and incentives promote successful fisheries. *Nature*, *470*(7334), 386–389. https://doi.org/10.1038/nature09689

Henrich, J., & Muthukrishna, M. (2021). The origins and psychology of human cooperation. *Annual Review of Psychology*, *72*(1), 207–240. https://doi.org/10.1146/annurev-psych-081920-042106

Hess, N. H. (2016, October 5). *Informational warfare* (M. L. Fisher, Ed.; Vol. 1). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199376377.013.15

Hilbe, C., Šimsa, Š., Chatterjee, K., & Nowak, M. A. (2018). Evolution of cooperation in stochastic games. *Nature*, *559*(7713), 246–249.

Hoffman, E., Mccabe, K. A., & Smith, V. L. (1998). Behavioral foundations of reciprocity: Experimental economics and evolutionary psychology. *Economic Inquiry*, *36*(3), 335–352. https://doi.org/10.1111/j.1465-7295.1998.tb01719.x

Hoffman, M., & Yoeli, E. (2022). *Hidden games: The surprising power of game theory to explain irrational human behavior* (First edition). Basic Books.

Imada, H., Romano, A., & Mifune, N. (2023). Dynamic indirect reciprocity: When is indirect reciprocity bounded by group membership? *Evolution and Human Behavior*. https://doi.org/10.1016/j.evolhumbehav.2023.05.002

James, W. (2007, April 1). *The principles of psychology*. Cosimo, Inc. (Original work published 1890)

Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature*, *530*(7591), 473–476. https://doi.org/10.1038/nature16981

Jordan, J. J., & Rand, D. G. (2017). Third-party punishment as a costly signal of high continuation probabilities in repeated games. *Journal of Theoretical Biology*, *421*, 189–202. https://doi.org/10.1016/j.jtbi.2017.04.004

Jordan, J. J., Sommers, R., Bloom, P., & Rand, D. G. (2017). Why do we hate hypocrites? evidence for a theory of false signaling. *Psychological Science*, *28*(3), 356–368. https://doi.org/10.1177/0956797616685771

Kleiman-Weiner, M. (2018). *Computational foundations of human social intelligence* (Doctoral dissertation). Massachusetts Institute of Technology.

Klein Teeselink, B., van Dolder, D., van den Assem, M. J., & Dana, J. (2023, April 25). High-stakes failures of backward induction. https://doi.org/10.2139/ssrn.4130176

Korndörfer, M., Egloff, B., & Schmukle, S. C. (2015). A large scale test of the effect of social class on prosocial behavior. *PLOS ONE*, *10*(7), e0133193. https://doi.org/10.1371/journal.pone.0133193

Krems, J. A., Hahnel-Peeters, R. K., Merrie, L. A., Williams, K. E. G., & Sznycer, D. (2023). Sometimes we want vicious friends: People have nuanced preferences for how they want their friends to behave toward them versus others. *Evolution and Human Behavior*. https://doi.org/10.1016/j.evolhumbehav.2023.02.008

Kubitz, G., & Page, L. (n.d.). *If you can, you must. the evolutionary foundation of reference point choice and loss aversion*. https://www.uts.edu.au/sites/default/files/2019-07/Gregory%20Kubitz.pdf

Lazer, D., & Friedman, A. (2007). The network structure of exploration and exploitation. *Administrative Science Quarterly*, *52*(4), 667–694. https://doi.org/10.2189/asqu.52.4.667

Lehmann, L., Powers, S. T., & van Schaik, C. P. (2022). Four levers of reciprocity across human societies: Concepts, analysis and predictions. *Evolutionary Human Sciences*, *4*, e11. https://doi.org/10.1017/ehs.2022.7

Leimar, O. (1997). Reciprocity and communication of partner quality. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *264*(1385), 1209–1215. https://doi.org/10.1098/rspb.1997.0167

Leimar, O., & Hammerstein, P. (2001). Evolution of cooperation through indirect reciprocity. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *268*(1468), 745–753. https://doi.org/10.1098/rspb.2000.1573

Lienard, P. (2016). Age grouping and social complexity. *Current Anthropology*, *57*, S105–S117. https://doi.org/10.1086/685685

Lie-Panis, J., & André, J.-B. (2022). Cooperation as a signal of time preferences. *Proceedings of the Royal Society B: Biological Sciences*, *289*(1973), 20212266. https://doi.org/10.1098/rspb.2021.2266

Lie-Panis, J., & André, J.-B. (2023, June 29). Peace is a form of cooperation, and so are the cultural technologies which make peace possible. https://doi.org/10.31234/osf.io/nr6ek

Lie-Panis, J., & Dessalles, J.-L. (2023, January 9). Runaway signals: Exaggerated displays of commitment may result from second-order signaling. https://doi.org/10.48550/arXiv.2301.03388

Lie-Panis, J., Fitouchi, L., Baumard, N., & André, J.-B. (2023, July 13). A model of endogenous institution formation through limited reputational incentives. https://doi.org/10.31234/osf.io/uftzb

Mailath, G. J., & Samuelson, L. (2006). *Repeated games and reputations: Long-run relationships*. Oxford University Press.

Maynard Smith, J., & Price, G. R. (1973). The logic of animal conflict. *Nature*, *246*(5427), 15–18. https://doi.org/10.1038/246015a0

McCloskey, D. N. (2016). *Bourgeois equality: How ideas, not capital or institutions, enriched the world*. The University of Chicago Press.

McCullough, M. E. (2008). *Beyond revenge: The evolution of the forgiveness instinct* (1st ed). Jossey-Bass.

McCullough, M. E., Kurzban, R., & Tabak, B. A. (2013). Cognitive systems for revenge and forgiveness. *Behavioral and Brain Sciences*, *36*(1), 1–15. https://doi.org/10.1017/S0140525X11002160

McCullough, M. E., Pedersen, E. J., Tabak, B. A., & Carter, E. C. (2014). Conciliatory gestures promote forgiveness and reduce anger in humans. *Proceedings of the National Academy of Sciences*, *111*(30), 11211–11216. https://doi.org/10.1073/pnas.1405072111

Mell, H., Baumard, N., & André, J.-B. (2021). Time is money. waiting costs explain why selection favors steeper time discounting in deprived environments. *Evolution and Human Behavior*, *42*(4), 379–387. https://doi.org/10.1016/j.evolhumbehav.2021.02.003

Muthukrishna, M., Francois, P., Pourahmadi, S., & Henrich, J. (2017). Corrupting cooperation and how anti-corruption strategies may backfire. *Nature Human Behaviour*, *1*(7), 1–5. https://doi.org/10.1038/s41562-017-0138

Nagel, T. (1979). *Mortal questions*. New York: Cambridge University Press.

Nannicini, T., Stella, A., Tabellini, G., & Troiano, U. (2013). Social capital and political accountability. *American Economic Journal: Economic Policy*, *5*(2), 222–250. https://doi.org/10.1257/pol.5.2.222

Nash, J. (1951). Non-cooperative games. *Annals of Mathematics*, *54*(2), 286–295. https://doi.org/10.2307/1969529

Nettle, D. (2018, October 15). *Hanging on to the edges: Essays on science, society and the academic life*. Open Book Publishers. https://doi.org/10.11647/obp.0155

Nettle, D. (2023, June 21). Innateness is for animals: Intuitive biology, intuitive psychology, and the folk concept of innateness. https://doi.org/10.31234/osf.io/n3qt4

Nettle, D., Colléony, A., & Cockerill, M. (2011). Variation in cooperative behaviour within a single city (Y. Moreno, Ed.). *PLoS ONE*, *6*(10), e26922. https://doi.org/10.1371/journal.pone.0026922

Noë, R., & Hammerstein, P. (1994). Biological markets: Supply and demand determine the effect of partner choice in cooperation, mutualism and mating. *Behavioral Ecology and Sociobiology*, *35*(1), 1–11. https://doi.org/10.1007/BF00167053

Nowak, M. A. (2006). *Evolutionary dynamics: Exploring the equations of life*. Harvard University Press. https://doi.org/10.2307/j.ctvjghw98

Nowak, M. A., & Sigmund, K. (1992). Tit for tat in heterogeneous populations. *Nature*, *355*(6357), 250–253. https://doi.org/10.1038/355250a0

Nowak, M. A., & Sigmund, K. (1993). A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game. *Nature*, *364*(6432), 56–58. https://doi.org/10.1038/364056a0

Nowak, M. A., & Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. *Nature*, *393*(6685), 573–577. https://doi.org/10.1038/31225

Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, *437*(7063), 1291–1298. https://doi.org/10.1038/nature04131

Ohtsubo, Y., & Watanabe, E. (2009). Do sincere apologies need to be costly? test of a costly signaling model of apology. *Evolution and Human Behavior*, *30*(2), 114–123. https://doi.org/10.1016/j.evolhumbehav.2008.09.004

Ohtsuki, H., & Iwasa, Y. (2006). The leading eight: Social norms that can maintain cooperation by indirect reciprocity. *Journal of Theoretical Biology*, *239*(4), 435–444. https://doi.org/10.1016/j.jtbi.2005.08.008

Osborne, M. J., & Rubinstein, A. (1994). *A course in game theory*. MIT Press.

Ostrom, E. (1990, November 30). *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press.

Page, L. (2022, November 3). *Optimally irrational: The good reasons we behave the way we do* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781009209175

Paldam, M., & Gundlach, E. (2008). Two views on institutions and development: The grand transition vs the primacy of institutions. *Kyklos*, *61*(1), 65–100. https://doi.org/10.1111/j.1467-6435.2008.00393.x

Panchanathan, K., & Boyd, R. (2003). A tale of two defectors: The importance of standing for evolution of indirect reciprocity. *Journal of Theoretical Biology*, *224*(1), 115–126. https://doi.org/10.1016/S0022-5193(03)00154-1

Panchanathan, K., & Boyd, R. (2004). Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature*, *432*(7016), 499–502. https://doi.org/10.1038/nature02978

Peysakhovich, A., Nowak, M. A., & Rand, D. G. (2014). Humans display a 'cooperative phenotype' that is domain general and temporally stable. *Nature Communications*, *5*(1), 4939. https://doi.org/10.1038/ncomms5939

Pinsof, D. (2023, March 1). The evolution of social paradoxes. https://doi.org/10.31234/osf.io/avh9t

Powers, S. T., van Schaik, C. P., & Lehmann, L. (2016). How institutions shaped the last major evolutionary transition to large-scale human societies. *Philosophical Transactions of the*

Royal Society B: Biological Sciences, *371*(1687), 20150098. https://doi.org/10.1098/rstb.2015.0098

Putnam, R. D., Leonardi, R., & Nanetti, R. (1994). *Making democracy work: Civic traditions in modern italy* (5. print., 1. Princeton paperback print). Princeton Univ. Press.

Qi, W., & Vul, E. (2022). The evolution of theory of mind on welfare tradeoff ratios. *Evolution and Human Behavior*, *43*(5), 381–393. https://doi.org/10.1016/j.evolhumbehav.2022.06.003

Quillien, T. (2020). Evolution of conditional and unconditional commitment. *Journal of Theoretical Biology*, *492*, 110204. https://doi.org/10.1016/j.jtbi.2020.110204

Raihani, N. J. (2021). *The social instinct: The hidden logic of cooperation.* JONATHAN CAPE LTD.

Raihani, N. J., & Bshary, R. (2015). Why humans might help strangers. *Frontiers in Behavioral Neuroscience*, *9*. https://doi.org/10.3389/fnbeh.2015.00039

Richter, X.-Y. L., & Lehtonen, J. (2023). Half a century of evolutionary games: A synthesis of theory, application and future directions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *378*(1876), 20210492. https://doi.org/10.1098/rstb.2021.0492

Roberts, G. (1998). Competitive altruism: From reciprocity to the handicap principle. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *265*(1394), 427–431. https://doi.org/10.1098/rspb.1998.0312

Roberts, G. (2020). Honest signaling of cooperative intentions. *Behavioral Ecology*, *31*(4), 922–932. https://doi.org/10.1093/beheco/araa035

Roberts, G., Raihani, N., Bshary, R., Manrique, H. M., Farina, A., Samu, F., & Barclay, P. (2021). The benefits of being seen to help others: Indirect reciprocity and reputation-based partner choice. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1838), 20200290. https://doi.org/10.1098/rstb.2020.0290

Selten, R. (1965). Spieltheoretische behandlung eines oligopolmodells mit nachfrageträgheit: Teil i: Bestimmung des dynamischen preisgleichgewichts. *Zeitschrift für die gesamte Staatswissenschaft / Journal of Institutional and Theoretical Economics*, *121*(2), 301–324. Retrieved July 4, 2023, from https://www.jstor.org/stable/40748884

Selten, R. (1983). Evolutionary stability in extensive two-person games. *Mathematical Social Sciences*, *5*(3), 269–363. https://doi.org/10.1016/0165-4896(83)90012-4

Sherratt, T. N., & Roberts, G. (1998). The evolution of generosity and choosiness in cooperative exchanges. *Journal of Theoretical Biology*, *193*(1), 167–177. https://doi.org/10.1006/jtbi.1998.0703

Sigmund, K., De Silva, H., Traulsen, A., & Hauert, C. (2010). Social learning promotes institutions for governing the commons. *Nature*, *466*(7308), 861–863. https://doi.org/10.1038/nature09203

Sijilmassi, A. (2021). *La psychologie derrière les constructions nationales : Comment les technologies culturelles favorisent la coopération à grande échelle et façonnent les "communautés imaginaires"*. (These en préparation). Université Paris sciences et lettres. Retrieved July 27, 2023, from https://www.theses.fr/s298983

Singh, M., & Hoffman, M. (2021, January 13). *Commitment and impersonation: A reputation-based theory of principled behavior*. PsyArXiv. https://doi.org/10.31234/osf.io/ua57r

Smaldino, P. E. (2017, May 25). Models are stupid, and we need more of them. In R. R. Vallacher, S. J. Read, & A. Nowak (Eds.), *Computational social psychology* (1st ed., pp. 311–331). Routledge. https://doi.org/10.4324/9781315173726-14

Spadaro, G., Molho, C., Van Prooijen, J.-W., Romano, A., Mosso, C. O., & Van Lange, P. A. M. (2023). Corrupt third parties undermine trust and prosocial behaviour between people. *Nature Human Behaviour*, *7*(1), 46–54. https://doi.org/10.1038/s41562-022-01457-w

Spence, M. (1974). Competitive and optimal responses to signals: An analysis of efficiency and distribution. *Journal of Economic Theory*, *7*(3), 296–332. https://doi.org/10.1016/0022-0531(74)90098-2

Számadó, S., Zachar, I., Czégel, D., & Penn, D. J. (2023). Honesty in signalling games is maintained by trade-offs rather than costs. *BMC Biology*, *21*(1), 4. https://doi.org/10.1186/s12915-022-01496-9

Tinbergen, N. (2010). On aims and methods of ethology. *Zeitschrift für Tierpsychologie*, *20*(4), 410–433. https://doi.org/10.1111/j.1439-0310.1963.tb01161.x (Original work published 1963)

Tiokhin, L. (2021, May 10). *Blog series: Modeling for metascientists (and other interesting people) - introduction*. https://doi.org/10.13140/RG.2.2.35351.50086

Tooby, J., & Cosmides, L. (1992). The psychological foundations of culture. In J. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 19–136). Oxford University Press.

Traulsen, A., & Glynatsi, N. E. (2023). The future of theoretical evolutionary game theory. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *378*(1876), 20210508. https://doi.org/10.1098/rstb.2021.0508

Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology*, *46*(1), 35–57. https://doi.org/10.1086/406755

Veblen, T. (1973). *The theory of the leisure class: With an introd. by john kenneth galbraith*. Houghton Mifflin. (Original work published 1899)

Wang, Q., He, N., & Chen, X. (2018). Replicator dynamics for public goods game with resource allocation in large populations. *Applied Mathematics and Computation*, *328*, 162–170. https://doi.org/10.1016/j.amc.2018.01.045

Whitehouse, H., & Lanman, J. A. (2014). The ties that bind us: Ritual, fusion, and identification. *Current Anthropology*, *55*(6), 674–695. https://doi.org/10.1086/678698

Wu, J., & Axelrod, R. (1995). How to cope with noise in the iterated prisoner's dilemma. *Journal of Conflict Resolution*, *39*(1), 183–189. Retrieved July 11, 2023, from https://econpapers.repec.org/article/saejocore/v_3a39_3ay_3a1995_3ai_3a1_3ap_3a183-189.htm

Wu, J., Számadó, S., Barclay, P., Beersma, B., Dores Cruz, T. D., Iacono, S. L., Nieper, A. S., Peters, K., Przepiorka, W., Tiokhin, L., & Van Lange, P. A. M. (2021). Honesty and dishonesty in gossip strategies: A fitness interdependence analysis. *Philosophical Trans-*

actions of the Royal Society B: Biological Sciences, *376*(1838), 20200300. https://doi.org/10.1098/rstb.2020.0300

Yoeli, E., Burum, B., Dalkiran, N. A., Nowak, M., & Hoffman, M. (2022, November 22). *The emergence of categorical norms* (preprint). In Review. https://doi.org/10.21203/rs.3.rs-2050019/v1

Zahavi, A. (1975). Mate selection—a selection for a handicap. *Journal of Theoretical Biology*, *53*(1), 205–214. https://doi.org/10.1016/0022-5193(75)90111-3

Zahavi, A. (1995). Altruism as a handicap: The limitations of kin selection and reciprocity. *Journal of Avian Biology*, *26*(1), 1. https://doi.org/10.2307/3677205

Zwirner, E., & Raihani, N. (2020). Neighbourhood wealth, not urbanicity, predicts prosociality towards strangers. *Proceedings of the Royal Society B: Biological Sciences*, *287*(1936), 20201359. https://doi.org/10.1098/rspb.2020.1359

Liste des éléments retirés

A la fin du chapitre 3, le matériel relatif à des expériences en cours a été retiré de ce document pour sa version en ligne (16 pages en tout). Il s'agit du texte présenté à des sujets recrutés sur Prolific pour trois séries de vignettes visant à tester les prédictions de ce chapitre (séries: (i) prédictions du modèle de base, (ii) prédictions du modèle avec excuses, (iii) prédictions du modèle avec connaissances privées vs. publiques).