Negative Indirect Reciprocity

Theory and Evidence

by

Matthew Chudek

B.A. University of Melbourne, 2003
M.A. University of British Columbia, 2009

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL
STUDIES
(Psychology)

The University Of British Columbia
(Vancouver)

August 2013

© Matthew Chudek, 2013
Abstract

Explanations of humans’ evolutionary origins that invoke the ratchet of cumulative cultural learning must confront the ‘cooperative dilemma of culture’. Adaptive cultural knowledge is a widely shared but easily degraded public good. How did our ancestors avoid the temptation to hoard valuable knowledge and to deceive and manipulate each other, before the advent of complex social institutions? I present one possible solution: negative indirect reciprocity (NIR). I use a series of mathematical models to reason about how our ancient ancestors’ dispositions to gainfully exploit one another could have supported more complex forms of cooperation, providing a foundation for our rapidly evolving corpus of shared cultural know-how. Together these models show how reputation-based, opportunistic exploitation can play a pivotal role in sustaining cooperation in small scale societies, even before the advent of complex institutions.

I also present two empirical tests of the assumptions made by these models. First, I measure contemporary reputational judgements in circumstances that the models predict are relevant. In the process I also map my participants’ judgements to the full set of first and second-order reputation assessment rules described by indirect reciprocity theory. Second, I test whether a recently observed peculiarity of people’s moral reasoning—our tendency to ascribe blame to those who profit from others suffering because of mere good fortune—is consistent with the constraints assumed by NIR. The results of both empirical studies support the assumptions made by NIR.
Preface

Chapters three, four and five of this thesis are being prepared for publication. Chapters 3 and 4 are coauthored by Joseph Henrich.

Chapter 3 is a theoretical model initially suggested by Joseph Henrich. I developed and refined numerous variants of this model. Henrich has provided in depth conceptual feedback throughout and has also provided detailed feedback of many versions of this manuscript.

The initial idea for the empirical study in chapter 4 emerged from discussions between Henrich and myself. I designed the details of the study, built a web-survey system capable of administering it, gathered and analysed the data, produced figures and prepared the manuscript. Henrich provided detailed feedback on the manuscript.

Henrich also provided all funding to pay participants in the studies described in chapters 4 and 5.

The data for study one in chapter 5 were gathered by Erik Thulin.
# Table of Contents

Abstract ............................................. ii

Preface ........................................... iii

Table of Contents ....................................... iv

List of Tables ........................................ vii

List of Figures ........................................ viii

Acknowledgments ....................................... ix

Dedication ........................................... x

1 Introduction ......................................... 1

2 The cooperative dilemma of an emerging cultural species 9
   2.1 The cooperative dilemma of culture (a.k.a. the evil teacher problem) ........................................ 14
      2.1.1 Why does culture require cooperation? ............... 16
      2.1.2 Why does (human-like) cooperation require culture? . 20
   2.2 Three kinds of solutions ................................ 22
      2.2.1 Route 1: The big step ............................... 23
      2.2.2 Route 2: The arms race ............................ 24
      2.2.3 Route 3: The ratchet ............................... 26
## 3 Negative indirect reciprocity

3.1 Mathematical model

3.1.1 Context and overview

3.1.2 Model definition

3.1.3 Reputational dynamics

3.1.4 Behavioural dynamics

3.1.5 Combining reputations and behaviour

3.1.6 Discrete strategy approach

3.1.7 Evolving continuous traits interpretation

3.2 Additional supplemental materials

## 4 Surveying indirect reciprocity: how do people assign reputations?

4.1 Methods

4.2 Results

4.3 Discussion

4.4 Supplemental

## 5 It is better to profiteer on the guilty: is moral condemnation sensitive to reputation?

5.1 Methods, study one

5.2 Results, study one

5.2.1 Do these questions index the same underlying construct?

5.2.2 Was condemnation of the lucky profiteer sensitive to the victims’ reputation?

5.3 Methods, study two

5.4 Results, study two

5.4.1 Do these questions index the same underlying construct?

5.4.2 Was the manipulation effective?

5.4.3 Did participants condemn the profiteer less for profiting on the suffering of the wicked?

5.4.4 Did participants reward the trader for taking a loss in miracle scenarios?
5.5 Discussion ................................................... 156

6 Conclusions .................................................. 161
  6.1 Near future directions ................................. 167
  6.2 Distant future directions ............................... 169

Bibliography .................................................. 172
List of Tables

Table 3.1 Summary of parameters ........................................... 54
Table 4.1 Glossary ................................................................. 92
Table 4.2 The cues used to establish the Target’s reputation .... 101
Table 4.3 A map from dichotomous reputations reputation changes
to their continuous interpretations ........................................ 108
Table 4.4 Correlations between our four measures of participants’
reputational judgments ....................................................... 113
Table 4.5 Summary of key results for Chapter 4 ....................... 118
Table 4.6 Frequencies of discretely categorised judgements of Targets 124
Table 4.7 Frequencies of discretely categorised judgements of Coop-
erators ................................................................................. 125
Table 4.8 Frequencies of discretely categorised judgements of Defectors126
Table 4.9 A discrete partition of participants’ reputational judgements 130
Table 5.1 Likelihood-maximising linear regression parameters ... of
ratings of the lucky profiteer ............................................... 146
Table 5.2 Correlations among study two dependant variables ... 150
Table 5.3 Participants’ ratings of Floret residents in Study 2 ........ 152
Table 5.4 Participants’ ratings of the lucky profiteer in disaster sce-
narios .................................................................................. 154
Table 5.5 Participants’ ratings of the lucky profiteer in miracle sce-
narios .................................................................................. 156
List of Figures

Figure 3.1 The NIR decision tree ................................. 35
Figure 3.2 NIR-P basins of attraction and equilibrium reputations . 39
Figure 3.3 NIR’s logic, testable assumptions and predictions .... 48
Figure 3.4 Simulations showing the accuracy of analytic predictions about Resident and Invader equilibrium reputations . . . 61
Figure 3.5 Simulations showing the accuracy of analytic predictions about equilibrium reputations of three discrete strategies. 64
Figure 3.6 Invasability plots for the three discrete strategies in NIR . 69
Figure 3.7 Meta-barycentric plots of NIR boundary equilibria . . . 76
Figure 3.8 Selection gradients for three version of NIR ........... 84
Figure 3.9 Relative volunteering rates for NIR-V and NIR-M .... 87

Figure 4.1 American Participants’ absolute judgements of Targets and relative judgements of Actors .................. 115
Figure 4.2 Indian Participants’ absolute judgements of Targets and relative judgements of Actors ..................... 116

Figure 5.1 Likelihood maximising estimates of the means of participants’ ratings ........................................ 145
Figure 5.2 Means of participants’ ratings of Floret residents .... 151
Figure 5.3 Means of participants’ ratings of the profiteer ......... 155
Acknowledgments

First and foremost I’d like to thank Joseph Henrich for support, advice and stimulating discussion throughout my time at the University of British Columbia. This would not have been possible without him.

I would also like to thank my committee members (Ara Norenzayan, Susan Birch and Michael Doebeli) and the amazingly bright and diverse faculty of the University of British Columbia’s psychology department for advice, support, collaboration and feedback.

I would like to acknowledge and thank the University of British Columbia, and especially the Faculty of Graduate Studies, for funding in the form of grants, fellowship and scholarships. I would like to thank the Department of Psychology for several travel grants and a terrific education.

I would especially like to thank Faye d’Eon-Eggertson for all her care and support, and for her careful proofreading. Your reading experience would be far less polished and far more jarring were it not for her heroic efforts.

Finally, I would like to thank my friends, family and loved ones for their support during these interesting times.
Dedication

I dedicate this work to my brother (Mark) and sister (Georgina). I’ve been so far away from them and missed so much of their lives to undertake this apprenticeship into the academy.

They’re amazing people and I miss them.
Chapter 1

Introduction

“What am I?”

Like “why is the sky blue?” and “why is my puppy sick?”, it is an entirely reasonable questions for a curious child to ask. But unlike those others, it is not a question we have a good answer to.

Our understanding of modern humans lacks a central explanatory paradigm. We know of no simple set of laws or principles which integrates and makes sense of the staggering range of things people do, make, say, think and become. Much of our understanding of humans is, by Thomas Khun’s\(^1\) definition, pre-scientific.

The explanations offered by sociologists—who focus on the institutions that emerge from our interactions—are entirely disconnected and often contradict those offered by social psychologists—who focus on how we, as individuals, are influenced by our social institutions. Economists’ consistent preference optimising rational agents have little or nothing to do with historians’ tension between Marxist or internalist interpretations of the causes of societal change. Anthropologists’ relativism sits at stark odds with the robust universalism implied by psychologists’ broad generalisations from undergraduate student samples. A psychologist studying crime who stumbles into a criminology conference would be just as perplexed by symbolic interac-

\(^1\) a twentieth-century philosopher of science, who’s descriptions of scientific progress (Kuhn, 1996) still inform many scientists’ understanding
tionism as a criminologist would be, were their situations reversed, by subtle
distinctions between attitudes, attributions, appraisals, attachments, acti-
vations, arousals, affects, automatic processes, beliefs, bias, construals, cog-
nitions, dissonances, deindividuations, emotions, ego-threats, perceptions,
representations, self-verbs, and the panoply other functionally-defined enti-
ties psychologists suppose occupy human minds.

It is rarer still for any of these disciplines’ explanations to explicitly
connect with the principles of biology. Even though we know that humans
everolved by natural selection, most disciplines proceed as though humans
play by entirely different rules from other animals, without ever explicitly
accounting for why or how this came to be.

Even within psychology the theoretical picture is little more integrated.
Cognitive psychologists, personality psychologists, developmentalists and so-
cial psychologists are engaged in conversations that rarely overlap. When
they do it is by the idiosyncratic connections made by individual researchers,
rather than due to the a priori, formally derived predictions of a shared cen-
tral theory of psychology. Rather than systematically testing and building
a shared paradigm, most psychological research consists of thorough but
disconnected local descriptions of particular phenomena and exciting novel
effects.

Within social psychology in particular, the theoretical landscape is par-
ticularly cacophonous. Many disconnected mini-theories cluster closely around
the empirical effects they describe and the methods used to investigate
them. Though individual researchers strive to make connections between
these mini-theories, these connections are typically ad hoc and rely on re-
searchers’ intuitive and common-sense understandings of their explanatory
concepts rather than precise and formal definitions.

When asked “what am I?”, we can offer many reasonable answers to little
questions. Why do we sometimes help each other and sometimes not? Why
are we sometimes willing to spend more money than other times? Why do
we forget things when we’re old? Why did Rome fall? Why do people from
that other country act so strangely?

However we lack a shared set of central explanatory principles that could
1. say how something like us can emerge from biology

2. predict *a priori* what we would be like; individually, collectively and historically

3. begin to integrate (and, sometimes, outright reject) the many very different, sometimes contradictory explanations for different aspects of our behaviour.

This absence of core theoretical principles might be a malfunction of the sociology of our science, a consequence of how we incentivise and motivate social scientists. Walter Miscel, then president of the Association for Psychological Science, implied as much when he turned his peers’ attention as the ‘toothbrush problem’ (Mischel, 2009):

> Psychologists treat other peoples’ theories like toothbrushes, no self-respecting person wants to use anyone else's.

It is also possible that our scientific institutions (including those that motivate researchers to invest primarily in their own novel, distinct micro-theories) are merely shaping themselves to the contours of our object of study. Central explanatory principles of the human phenotype may be perversely difficult to discover, even with the full arsenal of modern empiricism, or worse, there may not be any principles to find.

It is entirely conceivable that no such central explanatory principles exist. Perhaps the best answer we will ever be able to give to ‘What am I?’ is a scattering of local descriptions of disparate human-phenomena. One for cognitive dissonance. Another, that invokes very different theoretical assumptions, for demographic transitions. A third for extraversion, and so on. In such a world, our current sociology of isolated, idiosyncratically connected inquiry might be ideal.

I believe that, difficult though they may be to untangle, such principles do exist and are worth searching for. The argument for their existence is simple and I find it persuasive. Our species evolved by natural selection.
It did so very recently (on evolutionary timescales) and in a sudden, unprecedented explosion of complex novel behaviour. What’s more, no other species did the same. This suggests that, though some aspects of human psychology may have accreted independently over millennia, there were also some key recent ingredients that sparked an inferno of positive feedback and new dynamics. If there were simple, discrete causes then chances are we can find a simple explanatory principles that reconstruct them and predict their manifold consequences.

The value of searching for these principles, let alone the means of doing it, are harder to argue for. One could make a good case that our current hodge-podge of cross-disciplinary inquiry is the best way of discovering them. By careful, if initially disconnected, empiricism we might gradually establish a solid foundation of accurate descriptions of human-phenomena. This would accumulate, bottom-up, until they converged on the central explanatory principles. The counter-point to this argument is that our domain of inquiry—human behaviour—may be complex enough that this process is non-terminal. We could go on making empirical discoveries about humans forever. Or rather, like the proverbial blind men and their elephant\textsuperscript{2}, we would merely rediscover too-simple, low-dimensional redescriptions of the same underlying higher-dimensional phenomena, cloaked in different, incompatible theoretical jargon (for a philosophical explication of this possibility see Dennett, 1991).

For instance, one regularity that will be relevant later in this dissertation is ‘negativity bias’. People seem to have a robust tendency to respond more strongly to subjectively bad or unpleasant events than good ones. In 2001, Roy Baumeister and his colleagues documented a terrific diversity of domains and investigations in which this same pattern had been rediscovered (Baumeister et al., 2001). In each domain—from emotion research, to morality, to the trajectory romantic relationships, to economic choices about pecuniary profit and loss—it was interpreted as a novel insight, a local and distinct piece of evidence for a distinct theory, couched in distinct theoretic-\textsuperscript{2}\footnote{if you’re unfamiliar with the proverb, it can be readily found on the internet}
cal terms and entailed by different assumptions. Identifying these seemingly disconnected effects as facets of a single pattern was a great achievement.

Even though negativity bias is a very simple pattern—a linear relationship between the valence of an event and its effect\(^3\)—it was by no means simple or obvious to spot. If some or most of the real patterns that explain human behaviour are more complex than that—say non-linear relationships, or interactions between multiple factors, or patterns that spread across cultural timescales—then we may well be pessimistic about whether an industry of bottom-up empiricism will ever converge upon them.

If such concerns have you worried, you may see the worth of simultaneously pursuing a top-down strategy. That is, you might consider trying to derive the explanatory patterns \textit{a priori}. To do so, you would start from either our understanding of the biological principles that launched our species, or another sufficiently abstract, theoretically rigorous description of humans. You would then postulate as few principles, and as simple principles as you could and attempt to derive from them the gambit of human-phenomena (individual, collective, psychological, historical, etc.). Successful theories would be those that were simple, fit seamlessly with biology, and accurately predicted the greatest breadth of human-phenomena, especially previously undiscovered phenomena.

Such top-down attempts are valuable because, if successful, they can integrate disconnected empirical efforts and push our understanding forward in leaps and bounds. But they are also risky. You can invent an almost limitless diversity of such grand theories, but only a few are accurate or useful. Decades of valuable researcher time could be squandered fruitlessly pursuing them, at the price of empirical phenomena that could have been better documented in the meantime. This is, perhaps, why such ambitious and unlikely speculation is often frowned upon by psychologists.

However the difficulty of hitting upon an accurate top-down theory is also

\(^3\)Actually, given the null-hypothesis testing inferential framework in which most of this work is done, the relationship being formally tested is often magnitude-free and so simpler still: a dichotomy. More bad means more strong, regardless of how much more of either or the specific functional relationship between them.
an asset: a low false-positive rate. Top-down theories are unlikely to broadly predict novel contemporary empirical phenomena unless they are accurate. By ‘broadly’, ‘novel’ and ‘predict’ I mean that the theory is consistent with a wide range of phenomena that it was not explicitly designed to explain. The measure of top-down theories is the breadth of their explanatory power, their ability to easily integrate phenomena that previously seemed disconnected.

Relax. I will not propose a grand top-down theory in this dissertation. Instead, I hope to contribute to a broader paradigm of already successful top-down explanation. I will refer to this paradigm, which I review below, as ‘culture-gene coevolution’, even though scholars from many disciplines have converged on this same theoretical space and called it by different names. Specifically, I hope to contribute by refining the theoretical terrain on which culture-gene coevolutionary theories contend.

In chapter 2, I spotlight a set of top-down explanations of human behaviour that share these traits: 1) they are grounded in biology, and 2) they give a central role to ‘culture’—the social transmission of complex, encoded, phenotype-shaping information. I argue that though these attempts are already somewhat successful and very exciting, they are plagued by an under-recognised theoretical challenge: the cooperative dilemma of culture. I review existing solutions to this dilemma.

In the remainder of this dissertation I derive and test a novel solution to the cooperative dilemma of culture.

In chapter 3, I pick out one of the most plausible and commonly cited socio-ecological dynamics that may have established the cooperative foundation on which ancestral culture thrived: reputation. I argue that existing formal models of reputation do not address the cooperative dilemma of culture and offer a new model that does. This new model, which I call Negative Indirect Reciprocity (NIR), shifts the theoretical focus from the mutual-aid emphasised in most approaches to human cooperation (positive cooperation), to our countless unrealised opportunities to gainfully exploit one another (negative cooperation).

Many scholars have assumed that positive and negative cooperation are two sides of the same coin, and are equally well represented by our existing
models of positive cooperation. In chapter 3 argue that they are not, detail some of the important asymmetries between them and argue that negative cooperative dilemmas likely shaped our cultural species’ early evolution more strongly than positive ones did.

Succinctly, NIR is a model of “indirect reciprocity”—reputation-based cooperation. Unlike existing models, NIR minimises the cognitive prerequisites of reputation-based cooperation by assuming that cooperation or defection by ‘inaction’ (i.e., having an opportunity to act, but choosing not to) does not change reputations. That is, it does not assume that communities are able to coordinate their understanding of abstract opportunities, roles, responsibilities and so on. NIR show that, in the context of cooperation by ‘not exploiting’, such simple reputational systems can form the foundation of more complex societies.

‘Negative cooperation’ and ‘reputation without meaningful inaction’ have a special synergy. Together they create selective pressure for individuals to do whatever it takes to improve their reputation—that is, to please their peers on average. This pressure to attend to one’s community’s behavioural expectations can help explain how more sophisticated forms of cooperation arose.

NIR tells a specific, theoretically-rigorous story of how human reputations may have first emerged and how they formed the substrate of other forms of cooperation. It entails several clear empirical predictions about contemporary humans psychology. In the remaining chapters I test some of these predictions.

In chapter 4, I attempt to provide the first theoretically-informed catalogue of people’s reputation-based intuitions by simply asking them. I find that the valence of cooperation (whether it is helping or exploiting) matters a lot (as predicted by NIR) and that negative non-cooperation causes particularly strong reactions across cultures (as predicted by NIR).

Chapter 5 documents a study that, while simple, is important because it tests NIR against a novel theoretical phenomenon that it was not designed to explain. I find that this phenomenon, of people disliking those who incidentally profit while others suffer, fits quite nicely to the psychological biases
that NIR predicts we should have.
Chapter 2

The cooperative dilemma of an emerging cultural species

Somewhere between our split from chimpanzees (circa 6-3 mya, Patterson et al., 2006, but cf. Yamamichi et al., 2012) and the emergence of fully anatomically modern humans (circa 200 kya, Day, 1969; McDougall et al., 2005) our ancestors began behaving very strangely. Their new ways of thinking and behaving, and the social dynamics these engendered, were ultimately rooted in the same (relatively) well understood processes that shape the behaviour of our many non-human relatives: evolution by natural selection. Yet somehow this species experience an unprecedented explosion of complex and diverse behavioural forms like opera, foot-binding and hopscotch. On the face of it, it seems as though some new dynamics began interacting with natural selection.

Psychologists and other social scientists work to understand the staggering behavioural and cognitive diversity that has emerged in our species. For instance, cultural psychologists document deep cognitive differences between humans raised among different communities (e.g., Nisbet, 2003), evolutionary psychologists try to identify behavioural similarities which can be explained as adaptations to a shared ancestral environment (e.g., Barkow et al., 1992), while social psychologists tend to treat social environments as extrinsic and investigate how individuals respond to them, often implicitly
assuming that the breadth of diversity is well represented by North American university undergraduates (Henrich et al., 2010).

Alongside this industry of bottom-up inquiry, some scholars also confront the top-down challenge of disentangling the circumstances that first launched this process and the core laws or principles that continue to drive it forward. Scholars from across disciplines have proposed mechanism and dynamics at the core of humans’ distinct evolutionary trajectory, including archaeologists (Mithen, 1996), anthropologists (Boyd & Richerson, 2005; Deacon, 1998; Richerson & Boyd, 2004), primatologists (Tomasello, 1999) psychologists (Csibra & Gergely, 2009; Tomasello, 1999), biologists (Laland, 2004; Laland et al., 2001; Wilson, 2012), mathematicians (Nowak & Sigmund, 2005), linguists (Pinker, 2010) and philosophers (Sterelny, 2012), among others.

Refining the glut of theory is not easy. Early human societies no longer exist so we cannot directly observe their behaviour, experiment on their cognition nor watch them change. We cannot directly measure the dynamics that gave rise to our species. However, we may have some chance of reconstructing them from the convergence of several indirect sources of evidence.

Close inspection of our genome offers hints of ancestral speciation events (Garrigan & Hammer, 2006) and recent rates of genetic selection (Laland et al., 2010). However much variability among contemporary humans seems to be cultural, not genetic (Bell et al., 2009; Cavalli-Sforza & Cavalli-Sforza, 2000). Archaeological remains help fill this gap, but are limited to materials that survive decomposition and more sparse the further into the past we wish to gaze. Cross-species comparisons with our nearest cousins (e.g., Dean et al., 2012; Herrmann et al., 2007, 2010) provide a source of direct and even experimental evidence about evolved behavioural and cognitive differences. But while contemporary primates share our ancestors, they are not our ancestors and have experienced several million years of distinct evolutionary pressures.

Observation of contemporary forager societies can help here (e.g., Bell et al., 2009; Henrich et al., 2006), since their ecology, group size and residence patterns may be more similar to our ancestors’. However the mere existence
of contemporary foragers is reason to suspect their societies are unlike our
earliest ancestors’. Contemporary foragers somehow escaped the explosive
spread of agriculture (Gignoux et al., 2011), while typical foragers did not.

A final empirical avenue for testing these theories is to carefully derive
their implications for contemporary cognition, and compare these to how
people think and behave today. This is the method I pursue in this disser-
tation.

We can also make purely theoretical progress by narrowing the window
of plausible theories. One way to do this is to reject willy-nilly post-hoc
explanations of particular empirical phenomena (e.g., Frankenhuis, 2010;
Navarrete & Fessler, 2005). Another way is to notice and define theoretical
puzzles (e.g., Rogers, 1988) whose solutions (e.g., Boyd & Richerson, 2005;
Enquist et al., 2008; Laland, 2004) push our theories towards assumptions
more likely to be true. This dissertation highlights one such puzzle, the
cooperative dilemma of culture, for a subset the of these hypotheses: culture-
gene coevolutionary theories.

Culture-gene coevolutionary theories explain humans by emphasising cu-
mulative cultural learning. When they say ‘culture’ these theories are re-
ferring to behaviour-shaping information that is transmitted across genera-
tions socially, not genetically. While many species show evidence of some
cultural transmission (Brown & Laland, 2003; Laland, 2004; Rendell et al.,
2010; Whitehead et al., 2004; Whiten et al., 1999), only among ours did this
cultural information begin accumulating into ever more complex forms, that
eventually included ‘calculus’, ‘maps’ and ‘dancing the Macarena’.

The revolutionary importance of this transition is worth spending a few
paragraphs on. To give it some more intuitive, less jargony traction, I’ll
borrow a metaphor from cognitive psychology. Our minds, and other an-
imals’, are information processing systems; software instantiated upon the
hardware of our brains. They are computers designed to take sensory input,
interpret it, make some sense of the world outside, and make decisions about
how to behave. To do this, they must make assumptions about that world.

For instance we can, in a sense, directly perceive some of the assump-
tions our visual system makes when we look at a visual illusion and see
something we know is not really there. Developmental scientists have attempted to identify other, more conceptual assumptions that children use to make sense of the world. Proposed assumptions range from intuitions about the existence and properties of objects, the geometry of space, the existence and behaviour of goal-directed agents and even mathematical principles (see Spelke & Kinzler, 2006, for a review). They even include complex intuitions about the relative importance different kinds of information (e.g., Barrett & Broesch, 2012), the existence of ‘minds’ that are distinct from bodies (Chudek et al., forthcoming) and something functionally equivalent to Bayesian priors about the nature of causality (Griffiths et al., 2011; Kalish et al., 2007; Yeung & Griffiths, 2011).

Many of the assumptions we use to make sense of the world, like those that process vision, are encoded genetically. Our minds also seem to be able to infer other assumptions as they develop. For instance, rats rapidly make strong assumptions about which olfactory cues signal toxic food merely by observing other rats’ reactions (Galef & Whiskin, 2008). Though we don’t know whether rats explicitly represent this information in the same way humans do, we can observe their minds translating sensory information into behaviour as though they were assuming this of the outside world.

Though our minds (and rats’) can make some sense of the world beyond that encoded in our genes, the accuracy and complexity of these representations is limited by the scant sensory experience we are exposed to during our short lives. The advent of cumulative cultural learning was a revolution in the potential complexity of these representations. It let our minds also tap into the experiences of our conspecifics. In addition to learning about the world from our own experience and by observing others, we began to tap into the accumulated wisdom of minds long dead.

Imagine all the different assumptions our minds could make about the outside world as a vast space that stretches off in as many directions as you can conceive. Before the revolution of cumulative culture our individual journeys through this space were like sparks flying off of the slow-burning fuse of genetic evolution. Even today this fuse continues to endlessly wend its way through the space of possible external realities, guided by the selective
retention of more successful variants. Like other animals, our brains were constrained to perceive the world in whichever ways had led their ancestors to be more robust, fecund or in other ways fitter.

With the advent of cumulative culture something entirely new happened. Our ancestors detached from the fuse. They began incorporating others' representations of the outside world into their own, building on others’ assumptions, sparking into the unknown from a new position far from their genetically imbued starting point. We are still on this journey together. We each aggregate the experiences, knowledge and assumptions of our peers and each new generation inherits this amalgam. To continue the metaphor, our cultures are like great fireballs travelling through the space of ‘possible assumptions our minds could make about the world outside’. We blaze most brightly around an evolving core of shared assumptions, practices and understandings. From there our individual sparks still fly in all directions, innovating their own unique sense of the world outside.

Culture-gene coevolutionists work to understand this phenomenon of cumulative cultural learning; to theoretically and empirically describe it and discover any laws or principles that govern it. The questions they ask include:

- How did this process start, and why did it seemingly happen only in our phylogenetic line (e.g., Boyd & Richerson, 1988, 1996; Richerson & Boyd, 2004)?

- What cognitive adaptations and behavioural traits make it possible, and what new cognitive adaptations does it favour (e.g., Boyd & Richerson, 2005; Henrich, 2009; Henrich & Gil-White, 2001; Laland, 2004; Nakahashi et al., 2012)?

- Are there rules or principles that govern the trajectories such conflations of cultural learning will take through the space of all possible culture (e.g., Boyd & Richerson, 1988; Cavalli-Sforza & Feldman, 1981; Laland et al., 2001; Richerson & Boyd, 2004)?

- Can our evolving cultural corpus cause us to fracture into distinct,
symbolically marked ethnicities or social classes (e.g., Henrich & Boyd, 2008; McElreath et al., 2003)?

- How and why does this process produce cross-generational institutions, like religion (e.g., Gervais et al., 2011; Norenzayan & Shariff, 2008) or marriage (e.g., Henrich et al., 2012), and how do these evolve?

- How does cultural evolution redirect genetic evolution (e.g., Laland et al., 2010); how does it reshape our ecology and adaptive landscape (e.g., O’Brien & Laland, 2012; Rendell et al., 2011)?

- Does this process ever break down, how and why (e.g., Henrich, 2004)?

- How does this our evolving culture interface with our capacity for cooperation (e.g., Boyd et al., 2011a; Chudek & Henrich, 2011, and this dissertation)?

This culture-gene coevolutionary explanatory story is rapidly maturing, and has already had considerable success predicting many emergent features of contemporary society and psychology. However there is something troubling about it. Why did it make sense—from the perspective of our genes—for our genome to give up so much control over our adult phenotype, and let cultural information shape us instead?

2.1 The cooperative dilemma of culture (a.k.a. the evil teacher problem)

The cooperative dilemma of culture, which I also like to call the ‘evil teacher problem’, is pervasive. One need not agree with any of the specifics of any particular CGC theory to be trouble by it. It should concern you whatever your stance in controversies over group selection (e.g., Abbot et al., 2011; Nowak et al., 2010; West et al., 2011) and other evolutionary processes. To meet it, you merely need to accept the following premises:

1. **Humans are cultural** To understand us you need to explain the corpus of technology, ideas and knowledge that we have accumulated over generations, non-genetically.
2. **Humans are strangely cooperative** To understand us you need to explain how we became so cooperative. We regularly make choices that entail a relative cost (or forfeiture of a possible benefit) for us and relative benefit for others. Those others regularly include non-kin, people we know distantly or only by reputation, and even complete strangers who we are unlikely to meet again. The ways we cooperate are often unique to particular societies, and in any case vary dramatically across short stretches of time and space. No other species cooperates like we do (for detailed arguments for the peculiarity of human cooperation see Chudek et al., 2013b). Though many species cooperate in their own unique ways, human cooperation is distinct in its variety and rate of change. Humans in some places build houses together, in others they queue, and in others they construct and police far flung trade networks whose purpose is taking other humans as slaves. They also cooperate on vastly different scales. In some regions, villages are permanently at war with each other. In others, huge empires maintain relative internal harmony. Furthermore the rates at which the scale and form of human cooperation changes—some regions have moved from villages to nations within a single lifetime—is unprecedented and cannot be explained by genetic mutation and natural selection alone.

3. **Culture requires cooperation** For a species to accumulate a corpus of complex cultural knowledge, its members must accurately share valuable, fitness-relevant, phenotype-shaping information. Each individual acquires this information from the minds of others, who could easily distort it to their advantage or keep it to themselves but, typically, do not. When we share cultural knowledge, we cooperate. This is especially true for the complex, hard-to-observe ecological mastery that allows small scale foragers to thrive in some of the worlds harshest ecologies.

4. **Explanations of human cooperation invoke culture** Culture typically changes far more quickly than genomes do. Any mechanism that could keep us cooperatively sharing valuable cultural knowledge,
across its staggering breadth of domains, would need to *keep up with culture*. For example, a set of genetically adapted intuitions that managed to keep us honestly sharing food would do us little good when we began culturally transmitting warfare techniques, inheritance systems, queuing etiquette or intellectual property laws. Models of powerful, domain-general, human-specific cooperation-sustaining mechanisms do exist; I discuss some below and in Chapter 3. However these mechanisms typically presuppose that social groups are able to rapidly converge on and even enforce behavioural norms. That is, they presuppose that we are already a fairly cultural species.

5. **There is an explanatory regress** Accounts of human emergence face the challenge of explaining the origins of human culture and cooperation simultaneously, without assuming each to explain the other.

I will assume you agree with the first two claims, and will try to convince you of the others.

### 2.1.1 Why does culture require cooperation?

I am aware of three interrelated cooperative challenges posed by the emergence of human culture.

The first challenge: for complex culture (e.g., behavioural patterns no a-cultural individual could plausibly devise alone, like ‘trebuchets’ and ‘opera’) to accumulate, some individuals must share valuable information with others when they could do nothing. The second: those same individuals must not distort the cultural information to their own benefit, even though others must (to some degree) trust them. The third, language makes lying cheap.

For a clearer sense of these three arguments, lets imagine two individuals: a learner and a teacher. They are unrelated and the teacher has some information that would be useful to the learner. Each of us plays both these roles in our lives, sometimes both simultaneously.

At very early stages of the emergence of culture, when only simple behaviours are transmitted, it is plausible that the teacher has no choice but
to transmit their information. For example, imagine the information is a foraging technique. Unless the teacher is willing to not forage at all, they might have no choice but to demonstrate their technique to the onlooking learner. As long as cultural information is simple enough to be transmitted by mere looking-on, learners are in a position of power. However as soon as culture becomes complex enough that it requires any deliberate demonstration of a technique (e.g., stone-tool making), or a process must be performed precisely and rarely (e.g., making tools for making tools, house building), or can be done out of view of other (e.g., food processing) or requires specialised knowledge (e.g., tracking, foraging, medicinal plant use), the power changes hands. As soon as culture gets really interesting, interesting enough that some specialised teaching is required, teachers must cooperatively transmit valuable information to unrelated others.

I have heard two counter-arguments to this first thesis. The first is that perhaps such teaching could begin within the family, piggy-backing on kin-recognition mechanisms for cooperation. I have seen no formal models of such a process, and whether it could allow enough horizontal transmission for cultural accumulation to ever get off the ground. I do not think such models are necessary for two reasons:

- Humans don’t just behave that way. We do not show a widespread proclivity for limiting our cultural transmission to family members (e.g., Henrich & Broesch, 2011) and certain eye- and body-cues seem to put children into a ‘pedagogical stance’ where they are highly credulous to information transmitted by any adult, even stranger-scientists studying them (Csibra & Gergely, 2009).

- Even if cultural transmission started within families, the cooperative dilemma is merely pushed back (or forward in time) to whenever humans started sharing enough of this information with non-kin for our complex societies to begin developing (Chudek et al., 2013b).

The second counter-argument is usually voiced after people reflect on why they themselves are not evil teachers. A selfish teacher would soon
find that others refused to teach them too. The long-term costs of others’ exclusion would be too great. More formally, this intuition suggests that direct, pairwise reciprocity might have solved the cooperative dilemma of culture. Below I argue that it cannot (section 2.1.2).

The dilemma runs deeper. Imagine that the learner and teacher pass the first hurdle and the teacher begins sharing valuable information with the learner. This puts the learner in a very vulnerable position. While cultural information is simple enough that the learner can readily apprehend the causal logic of what they are doing, they need not ‘trust’ the teacher in any fitness-relevant sense. However as soon as the corpus accumulates information complex enough that it is causally opaque to us (which, I suspect, need not be very complex at all), the learner must trust. That is, adaptations must enter their genome which cause it to give up some control over its adult phenotype. Instead, their behaviour starts to be shaped by information that has passed through others’ minds. This information must be, on average, adaptive for the learner or such adaptations would not be favoured.

The second challenge: complex cultural accumulation implies that some control over the learner’s behaviour is in the hands of the teacher. This should select for exploitative adaptations in the teacher. Teachers who find ways to mutate and distort cultural information to their own advantage should be favoured by selection.

Such evil teachers could, for instance, teach foraging techniques or dietary prohibitions that result in the teacher eating some of the learner’s food. While I have occasionally seen older children teaching younger children rules unabashedly biased in their interests, the endless possibilities for such Machiavellian manipulation of cultural information do not even occur to most adults. The more the first challenge is overcome, the more this second challenge is exacerbated. The more valuable information we share, the more culture-dependent learners come trust this causally-opaque information, the more opportunities there are for evil teachers to distort it to their benefit. The importance of this second challenge is underscored by just how trusting contemporary children are of non-kin adults (Harris, 2012).
The most common answers I have heard to this challenge are that manipulative teaching would result in a) retaliation and b) the teacher being ignored in the future. These are ways of restating the dilemma rather than solutions. How do learners know when to retaliate? Say I teach you a medicinal remedy but omit a scarce ingredient so I can gather more myself. How are you to know that your slower recovery times are not a consequence of some other aspect of your lifestyle, your poorer skill at preparing medicines or the ill will you earned of a sorcerer? How can I infer that you snare the big game more often than me because you deliberately taught me inferior techniques, and not due to the countless other differences between us?

If retaliation is error-prone (sometimes you weren’t deceiving me) how is it not more costly for the learner than blind trust or walking away? Why do the benefits of exploitation for the evil teacher not exceed the costs of potential retaliation or being ignored? Answering these questions requires specific hypotheses about ancestral socio-ecological dynamics and how the retaliation or ignoring affected them. It requires detailed models of how these ecologies interacted with genetic and cultural evolution. It requires a solution to the cooperative dilemma of culture.

The third challenge of the cooperative dilemma of culture: language makes lying cheap. One plausible way to meet the first two challenges is to argue that it is more costly for the teacher to withhold information, or figure out how to distort it exploitatively, than the benefits they could obtain by doing so. This argument fails, and the first two challenges are exacerbated, when language enters the picture. At some point in the history of our cultural corpus, it became encoded it in a semantically rich, combinatorial language.

I find persuasive the argument that given what a complex, specialised adaptation language is, there must have existed a strong need for it (i.e., a complex cultural corpus worth transmitting) before it was selected for. Regardless of your stance on this issue when language entered the picture, it became possible to transmit far more complex, subtle cultural information, but also to engage in subtler, more complex deceptions. There are mathematical models demonstrating that language makes lying powerful and easy.
(Lachmann & Bergstrom, 2004), but I suspect that this evolutionary argument is intuitive, simple and uncontroversial enough that it does not require them. When you have language, it is easy to lie.

Hopefully these arguments convince you that only a cooperative species could be cultural. But how do we know that the cooperative chicken did not precede the cultural egg?

2.1.2 Why does (human-like) cooperation require culture?

Many animals cooperate. For instance, some eusocial insects act as a single super organism by sequestering their reproductive line, just like the cells in our bodies do (Smith & Szathmáry, 1997). On a smaller scale, some social mammals provision one another with public goods, such as making self-endangering alarm-calls upon sighting a predator (Seyfarth et al., 1980).

However human culture could not have piggy-backed on these simpler forms of cooperation. First, humans are not eusocial, we are distantly related and each reproduce individually. Second, specific genetic adaptations for specific cooperative behaviours in specific domains could not sustain culture. Culture changes much faster than genes do. As culture accumulates, cooperative dilemmas arise rapidly in many different domains. Who should take the risks when hunting? How should we divide the spoils? Who should do the hard-labour and who should pray for rain? Who marries whom, who reproduces when and who contributes which resource to which resource-hungry child, during their long, unproductive juvenile period (Gurven et al., 2006; Kaplan & Robson, 2002; Lancaster et al., 2000; Walker et al., 2006, 2002)? By the time our slowly mutating genome develops a mechanism (e.g., a signal; Boyd et al., 2010) for solving one of these dilemmas, our rapidly changing cultural corpus has generated countless others.

If ancestral humans had genetic adaptations for sustaining cooperation that supported their cultural accumulation, they must have been domain-general adaptations that could rapidly stabilise arbitrary, new forms of cooperation whenever they arose. I am aware of no purely-genetic mechanisms for sustaining culture (e.g., kin-selection, limited dispersal, genetic group-
selection, etc.) that can satisfy this criterion.

A second class of mechanisms posit *institutions* for sustaining culture. That is, highly coordinated, sometimes socially enforced, behaviours that ensure that it is in individuals’ interest to cooperate in arbitrary domains. A simple example is a police force or ‘punishment pool’ (Sigmund et al., 2010). Individuals first put resources towards sustaining the existence of an impartial punisher (e.g., paying the sheriff) and then the punisher spends those resources on making defectors pay. These institutional mechanisms can ‘keep up with culture’. They can sustain arbitrary forms of cooperation without requiring slow, specific genetic adaptations. However most institutional mechanisms presuppose a species that is able to dynamically and rapidly coordinate such institutions. For instance, how do people establish punishment pools? How do they ensure the pool’s resources do effectively punish? How do they coordinate what specific behaviours the pool should punish. How do all these moving parts keep up with our rapidly evolving culture?

Even if such institutional mechanisms did not require language and sophisticated, culturally transmitted concepts, they would at a minimum require that we carefully attended to and trustingly imitated one-another’s behaviour. While institutional mechanisms undoubtedly played and continue to play a role in sustaining and managing the latter forms of human sociality, they cannot be invoked to explain the emergence of culture without stumbling into an explanatory regress.

A third class of models posits genetically-selected, individual-level behavioural strategies that, in aggregate when common-enough, sustain cooperation in arbitrary domains. The flagship of such promising solutions is reciprocity.

‘Direct reciprocity’ explores how individuals’ can benefit by conditioning their behaviour towards someone on how that someone treated them in the past. If enough individuals act this way, pairwise cooperation can thrive (Axelrod & Hamilton, 1981; Boyd & Lorberbaum, 1987; Doebeli & Hauert, 2005; Trivers, 1971; van Veelen et al., 2012). However direct reciprocity struggles to explain cooperation among many individuals simultaneously,
such as provisioning non-excludable goods to an entire group (e.g., sharing valuable cultural information).

‘Indirect reciprocity’ (Leimar & Hammerstein, 2001; Ohtsuki et al., 2006; Panchanathan & Boyd, 2004; Panchanathan et al., 2003; Sigmund, 2012) or reputation-based cooperation is a better candidate for catalysing the emergence of human culture. Models of indirect reciprocity (discussed further in chapter 3) explore how individuals can benefit by conditioning their behaviour towards someone on how that someone treated a third person in the past. Though most models of indirect reciprocity only consider pairwise cooperation, it is easy to see how they can be extended. Once individuals have reputations, and once those reputations coordinate how others treat them, selection will favour individuals who do whatever it takes (as long as it is not too costly) to improve their reputation. If provisioning public goods (e.g., freely sharing knowledge) improves one’s reputation, then indirect reciprocity can sustain that too.

However for indirect reciprocity to hold water as a solution to the cooperative dilemma of culture, one would need to demonstrate that it can get off the ground without presupposing the existence of a cultural species. In chapter three I argue that existing models do make this supposition and so fall prey to the explanatory regress we are trying to avoid. I then go on to provide an alternative that does not.

In short, I am not aware of any existing theories (or formal, evolutionary models) that show how the kind of domain-arbitrary, rapidly adapting cooperation necessary for culture could emerge with assuming sophisticated cultural pre-adaptations.

If the cooperative dilemma of culture is a real puzzle of CGC theories, then how can we solve it?

2.2 Three kinds of solutions

I am aware of three broad approaches to untangling the cooperative dilemma of culture.
2.2.1 Route 1: The big step

The first approach supposes that human culture was a fortuitous accident, the convergence of several other mature preadaptations. My impression is that this is many non-experts’ default model, though it has also been proposed by careful thinkers (e.g., Pinker, 2010).

Humans became bipedal, began using tools, developed language including specialised, localised cognitive adaptations, became quite intelligent and developed many of the specialised cognitive adaptations psychologists are documenting today. All this happened, without any substantial accumulation of a cultural corpus (certainly not enough to warrant worrying about the cooperative dilemma), during the five million years or so since we split from other primates. Finally, during the last fifty thousand years or so, a mature version of human culture emerged all at once, a consequence of these other adaptations.

In this scenario, there is no cooperative dilemma because culture doesn’t coevolve slowly with genetics. It comes into existence full-blown after humans are, more or less, anatomically and cognitively modern. We need not puzzle over how selection favoured genes that promoted reliance on culture simultaneously to selecting cooperative cultural content, because the genetic adaptations happened first and then culture came second. By the time culture was on the table, humans were intelligent enough to reason their way past the fitness trough of mutual exploitation to the distant fitness peak of mutual cooperation. This route out of the dilemma relies on us being too smart to culturally exploit one another, smart enough to share culture freely.

I can understand why this is most people’s default argument, and why it seems intuitive. Emerging evidence suggests that humans are intuitive dualists (Chudek et al., forthcoming), we think about our intentional, rational human minds and our animalistic bodies as different (and potentially separable) kinds of things. We intuitively imagine minds as inhabiting bodies. This makes it particularly easy to imagine that our gene-built brains and bodies came first, and then our culture-shaped minds began inhabiting
them, fully-formed, afterwards.

While I believe these accounts do not hold up to close inspection, this dissertation is not the place for that argument (see Boyd et al., 2011b, for recent arguments to this end). Notice only that for such big-step explanations to be plausible, they must specify a suite of adaptive benefits that carved out a ‘cognitive niche’ (e.g., Pinker, 2010) and drove our metabolically costly brain expansion (Aiello & Wheeler, 1995; Kotrschal et al., 2013), but did not do so for other species in our planet’s long history. They must also explain what favoured our fantastic penchant for language, which children not exposed to a language-community seemingly develop spontaneously (Senghas et al., 2004), without invoking its advantages for the transmission of complex, encoded information.

Explanations that put culture first can do this parsimoniously—the adaptive value of our accumulated cultural knowledge selects for brains ever better at accessing and using it. If we prefer this parsimony, we must pay for it by confronting the evil teacher problem.

### 2.2.2 Route 2: The arms race

An alternative is that, though teachers have always been incentivised to exploit, learners have evolved cognitive counter-measures, ways of sifting the cultural wheat from the deceptive chaff. On this account, there is an evolutionary arms race between evil-teachers, whom evolution is shaping to send biased information, and sceptical learners. Learners may avoid exploitative cultural information by, for instance, preferring information that is backed by action (Henrich, 2009), averaging information between models (Boyd & Richerson, 1988) or imitating others’ model choices (Henrich & Gil-White, 2001).

I find this solution more plausible but have several concerns. First, even if learners are not manipulable, it is not obvious why teachers would transmit cultural information at all (though plausible solutions do exist, such as the possibility that valued teachers accrue social benefits, termed deference; Henrich & Gil-White, 2001). Solving the second challenge merely exacer-
bates the first. It may be possible that a run-away evolutionary process could navigate this trade-off. Learners have an advantage, then teachers do, then learners, and so on. Cognitive adaptations for culturally exploiting and avoiding exploitation escalate quickly enough that it is always worth transmitting ever more valuable cultural information. However I have not yet seen (or developed) any clear models of whether and how such a process could work.

On the empirical front, if our were species were situated at the cutting end of such an arms-race, we ought to be suspicious learners and shrewdly manipulative teachers. Instead, when it comes to sharing cultural information, we are both credulous and honest. We happily send our children to schools where they diligently learn from often under-paid, selfless strangers. Children (Lyons et al., 2007; Whiten et al., 2009) and even adults (McGuigan et al., 2011) dutifully follow the strange and seemingly pointless instructions of strangers, even strangers who are patently from exotic out-groups (Nielsen & Tomaselli, 2010). Children even go out of their way to enforce rules learned from strangers on others (Rakoczy et al., 2009). Infants deliberately make patently incorrect choices when strangers suggest them with even very subtle pedagogical cues (Topál et al., 2008). We even find it hard to distinguish strangers’ subtle suggestions from our own memories (Loftus & Palmer, 1974), and infants trust knowledgeable strangers over their own mothers (Stenberg, 2009).

These high levels of trust contrast starkly with children’s impressive selectivity in choosing between potential social models. They prefer, for instance, to learn from their more confident (Birch et al., 2009), previously knowledgeable (Birch et al., 2008), self-similar (Buttelmann et al., 2012) and prestigious (Chudek et al., 2012) peers (see Chudek et al., 2013a, for a recent review). This suite of learning behaviour is more likely a product of a selective environment where communities shared knowledge freely, honestly and cooperatively children’s dilemma was distinguishing the higher quality model, than an arms race between Machiavellian deception and lie-detecting.
2.2.3 Route 3: The ratchet

A third possibility is that some co-evolutionary process ratcheted up human cooperation and culture simultaneously. This ratchet would start from small amounts of culture, little and simple enough that its transmission would not generate a cooperative dilemma. This early culture would generate evolutionary dynamics that sustained cooperation in the ways required to sustain more complex culture, which would sustain more cooperation and so on.

There are many possible ways that such culture-cooperation ratchets could have worked. For instance, one could posit that teaching began primarily among kin (Fogarty et al., 2011), allowing enough cultural sophistication to accumulate that complex cooperation sustaining mechanisms could come into play and allow broader information sharing. Or perhaps early teaching was promoted by its utility in improving the return of mutualistic endeavours like hunting (see Sterelny, 2012, for one such account).

Alternatively, culture could have given rise to new strategic niches (e.g., prestigious leaders Henrich & Gil-White, 2001) or new learning strategies such as conformism (Boyd & Richerson, 1988; Henrich & Boyd, 2001) and ‘credibility enhancing displays’ (Henrich, 2009) which changed the nature of the cooperative dilemmas faced by our ancestors.

I am keen to promote a clearer understanding of the cooperative dilemma of culture, and to see the many plausible solutions proposed, formalised, debated and evaluated. To help drive this process forward, here I will present a novel proposal which uses reputation to kick off the ratchet by solving the cooperative dilemma in a way that selects for conformity to arbitrary, culturally evolving norms.

Contemporary interactions between cooperation and cultural learning, especially non-conscious behavioural imitation, give prima facie grounds to suspect that they have a long and intertwined evolutionary history. People who are imitated by others tend to act more prosocially afterwards, both adults (for reviews of this extensive literature, see Chartrand & Bargh, 1999; Chartrand & van Baaren, 2009; Lakin et al., 2003) and children (Carpenter
et al., 2013). Also, children are more likely to trust information learned from individuals who have imitated them (Over et al., 2013). Reciprocally, people who are motivated to affiliate with others—for instance, those who have been ostracised (that is, excluded from the benefits of their peers’ sociality)—respond by imitating others more. Again, this is true of both adults (Chartrand & van Baaren, 2009; Lakin et al., 2008, 2003; Leighton et al., 2010) and children (Over & Carpenter, 2009).

In chapter 3, I argue for the plausibility of a previously unconsidered ratchet: Negative Indirect Reciprocity (NIR). NIR explains how, if small amounts of cultural learning brought proto-reputations into existence, communities could have coordinated their opportunities to exploit others in a way that enforced their shared social norms. The existence of such community-enforced norms, in turn, could support more sophisticated, institution-based forms of cooperation, facilitating the emergence of a complex, cooperative cultural corpus.

In the next chapter I present the mathematical theory that articulates NIR in a way that is accessible to psychologists. I focus on deriving clear, testable predictions about contemporary psychology. In the subsequent chapters, provide empirical tests of these predictions. In chapter 4, I survey people’s actual reputational intuitions and compare them to the predictions made by indirect reciprocity theory, including NIR. In chapter 5, I present a genuinely novel test of NIR. I test its fit to a recently observed psychological phenomenon that it was not designed to explain.
Chapter 3

Negative indirect reciprocity

The puzzling origins of both human cooperation and cultural learning are likely intertwined. Some aspects of human cooperation are shared with other species and were likely shaped by the same kinds of distal pressures (e.g., kin-based sociality; Daly & Wilson, 1988; Hamilton, 1964; Park et al., 2008; Smith, 1964; Stewart-Williams, 2007). Other aspects seem peculiar to humans, such as queuing, paying taxes, and sacrificing to all-powerful deities (Norenzayan & Shariff, 2008). Formal models of how such cooperation-sustaining institutions emerge and persist typically presuppose that we are a highly cultural species. For instance, important models assume that individuals can readily coordinate their cognitive representations related to identifying who is a deserving “recipient” and a responsible “donor” (Boyd et al., 2010; Leimar & Hammerstein, 2001; Panchanathan & Boyd, 2004), that people establish abstract institutions (Sigmund et al., 2010) and know how to interpret one-another’s signals of cooperative intention (Boyd et al., 2010).

These assumptions imply a deeper puzzle, since cognitive capacities for sophisticated cultural learning themselves pose a cooperative dilemma. They imply that natural selection favoured mutations that relaxed our genome’s control over its phenotype, and allowed fitness-relevant behaviour to be shaped by cultural information (Richerson & Boyd, 2004) acquired from unrelated conspecifics. For this to be plausible, something must ensure that
the information acquired from others is, on average, fitness-enhancing, especially if large, metabolically expensive brains are required to access it (Aiello & Wheeler, 1995).

But the more influence culturally-learned information has on one individual’s behaviour, the greater the selection pressure for others to exploit that dependence by distorting the information they transmit to their own benefit. At first, many learnable phenotypes (e.g., stone tools use techniques) may have been hard to conceal or distort. However as soon as cultural know-how became complex enough that it required language or pedagogy (Csibra & Gergely, 2009) for transmission, lying to exploit trusting others became cheap and almost limitless in its potential for lucrative deceptions (Henrich, 2009; Lachmann & Bergstrom, 2004). This ‘cooperative dilemma of culture’ is exacerbated by the fact that cultural information changes far more quickly than genetic information, making it unlikely that genetically-evolved intuitions alone could distinguish useful from exploitative cultural knowledge. Before language and the cognitive capacities for coordinating complex cultural institutions could have emerged, some mechanisms operating on the same timescale as cultural change must have reliably sustained the quality of the public good that is our shared corpus of adaptive cultural know-how.

To tackle this puzzle, I ask how mechanisms that rapidly ensure cooperation in arbitrary domains (e.g., hunting, sharing information, trade) can get off the ground without pre-existing capacities for socially coordinating complex institutions and cognitive representations. One promising possibility are models of ‘indirect reciprocity’ (IR). *Prima facie* most IR models merely assume that (a) individuals have opinions of one another, (b) that these opinions influence how they treat each other and (c) that communities somehow synchronise these opinions. These synchronised opinions are called ‘reputations’. Once reputations exist they can catalyse the emergence of more complex forms of cooperation (Chudek & Henrich, 2011; Panchanathan & Boyd, 2004).

Since many primates form coalitions with non-kin (Higham & Maestripieri, 2010; Langergraber et al., 2007; Perry & Manson, 2008; Silk, 2002;
Watts, 2002), the first two assumptions typical of IR models are plausibly preadaptations in our phylogenetic lineage. The third assumption implies some social coordination and suggests a cooperative dilemma of culture. One could accrue many fitness benefits by strategically manipulating reputations. However it is also plausible that early, pre-verbal reputations were transmitted by observing others interactions (i.e., rather than gossip) and could not be easily distorted. Cognitive biases initially evolved to increase the quality of general social learning (e.g., Boyd & Richerson, 2005; Henrich, 2009; Henrich & Gil-White, 2001; Laland, 2004; Nakahashi et al., 2012), may have incidentally also caused the transmission of social opinions (which are, after all, just another type of cultural information) bringing reputations into existence. If early cultural learners closely monitored one-another’s dietary, foraging or tool-use preferences, they may have picked up preferences concerning community members too.

However, even if we grant the plausibility of the third assumption, existing IR models implicitly assume an even stronger form of cultural coordination. Framed in the context of reciprocal helping, these models suppose that sometimes someone has an opportunity to help but does not, and that their reputation worsens due to their inaction. This seemingly innocuous assumption implies that their peers somehow coordinate their representations of both the abstract opportunity to act, and the significance of inaction. This is a sophisticated cognitive feat. It is especially impressive if we assume it emerged early enough to sustain the diverse, cooperative cultural corpus that allowed an African primate species to spread across the globe.

Noting this issue, Leimer and Hammerstein write that IR models assume “a reasonably fair and efficient mechanism of assigning donors and recipients...a well-organised society, with a fair amount of agreement between its members as to which circumstances define the roles of donor and recipient.” (Leimar & Hammerstein, 2001). Subsequent IR models have mirrored these assumptions without concern for their limitations.

Here I fill the gap between the emergence of human culture, cooperation and reputations by showing how IR can get off the ground without assuming coordinated reactions to ‘inaction’. Our model assumes that inac-
tion never changes reputations and demonstrates that even so IR can form
the substrate of more complex forms of cooperation. In fact, our model
is grounded in a relatively modest assumption about early cognition: that
individuals disliked (i.e., worsen their reputational representation of) those
who actively and observably exploited someone they liked (i.e., those with
a good reputation).

For several reason I focus on “negative cooperative dilemmas” where ‘de-
fecting’ means gainfully exploiting someone and ‘cooperating’ means seeing
such an opportunity to exploit someone but passing it up (‘doing nothing’).

**Substantial positive cooperation presupposes negative cooperation:**
Before more complex forms of mutual aid, defence and helping emerge,
the ubiquitous opportunities to exploit each other (particularly, the
old, weak and injured) must be brought under control. Otherwise, ex-
ploration and cycles of revenge will undermine positive cooperation.

**Positive cooperation exacerbates the negative dilemma**
*but not the reverse:*
The mutual aid of positive cooperation can create an abundance of
exploitable resources, both tangible (e.g., food caches) and intangible
(e.g., trust). If cooperation has not first been stabilized in negative
dilemmas, exploitation can quickly sap these benefits, sabotaging the
viability of positive cooperation.

**Escalating returns:** Prior to the emergence of complex institutions like
money and debt, if an individual with a good reputation is helped
multiple times (i.e., by multiple peers) they typically experience dimin-
ishing marginal returns. A little food when you are starving provides
a huge benefit; a lot of food when you are full provides only a small one. On the other hand, repeated exploitation (e.g., stealing someone’s
resources) can put victims in an ever more desperate situation with
ballooning fitness consequences. This suggests negative dilemmas may
have generated steeper selection gradients earlier in our evolutionary
history, and so had more influence on the direction of evolution.
**Built-in individual-level motivation:** In a positive cooperative dilemma with unobservable inaction (or lack of sufficient agreement about what constitutes ‘inaction’), an individual’s reputation can endogenously rise (by observably helping) but not fall. Though an individual’s reputation might fall accidentally, selection will never favour individuals who take deliberate costly actions to worsen their reputation. Reciprocally, negative dilemmas generate selective pressure for individuals to take deliberate, costly actions to improve their reputation. Positive dilemmas can’t accomplish this until sufficiently complex cultural institutions or cognitive abilities establish agreement about what constitutes ‘inaction’. This creates a chicken and egg situation for positive dilemmas, since substantial cooperation is required before sophisticated cultural-cognitive abilities can emerge.

**Relevance to culture:** The cooperative dilemma of cultural learning, the main hurdle to more sophisticated institutional forms of cooperation, is a fundamentally negative dilemma. Individuals must pass up opportunities to gainfully deceive their credulous conspecifics.

**Preadaptations are more plausible:** Negative dilemmas are not symmetric with positive cooperative dilemmas because they require that individuals notice, cognitively represent and respond to opportunities to profit by exploiting others, while positive cooperation requires they represent opportunities to pay costs to help others. The former abilities were likely better, earlier in the culture-cooperation coevolutionary process, since they yield direct, self-interested gains.

**Contemporary humans are more sensitive to harm than helping:** Harmful or aversive actions, events or stimuli are more likely to have effects, and typically have stronger effects on contemporary humans than their positive or beneficial counterparts (for extensive and influential reviews, see Baumeister et al., 2001; Cacioppo & Berntson, 1994; Rozin & Royzman, 2001). This pattern recurs across the gambit of human cognition, from sensation, to the experience of emotions and their
effect on cognitive processing, to trajectories of learning, to memory, impression formation, and the effect of feedback on self-perception. Of particular relevance is that negative information (i.e., about other’s harmful acts) seems to have a far more potent effect on diminishing someone’s reputation than positive information has on restoring it (Fiske, 1980; Rozin & Royzman, 2001; Skowronski & Carlston, 1987). Early antecedents of this negativity bias may even be apparent among three-month-old infants (Hamlin et al., 2010). In fact, people are more likely to judge that someone caused, and intended to cause, a negative outcome than a corresponding positive one, even if their actual action was identical (Knobe, 2003, 2010). The ubiquity of negativity biases in contemporary cognitions suggests they run deep and may have ancient evolutionary roots. If our ancestors were as negativity-biased as we are, the impact of negative cooperative dilemmas may dwarfed positive ones in determining the long-run distribution of their reputations.

**Contemporary humans are more sensitive to harm by commission than harm by omission:**

Contemporary humans tend to condemn others moral transgression more severely (Baron & Ritov, 2004; Cushman et al., 2006; Spranca et al., 1991) when they are the result of deliberate actions (*commissions*, which play a central role in NIR), than if they are the consequence of an equally intentional inactions (*omission*, which are absent from NIR). Correspondingly, people seem less-disposed to transgress by commission than omission (Ritov & Baron, 1999), especially if they might be punished by others (DeScioli et al., 2011). These effects, which seem peculiar to negative commissions (Spranca et al., 1991) not positive ones, support NIR’s emphasis on negativity cooperation by commission alone.
NIR-P: \( \kappa = 1 \); NIR-V: \( \zeta = 0 \);
NIR-M: \( \zeta = 1 \)

Evolving variables: \( v, s \)
Parameters: \( \rho, \kappa, \zeta \)
Consequences: \( d, t, k \)

- Opportunity to act
  - Opportunity to exploit
    - Victim has good reputation
      - Exploit
        - (Inflict \( d \)-amage, earn \( t \)-akings, reputation worsens)
    - Victim has bad reputation
      - Exploit
        - (Inflict \( d \)-amage, earn \( t \)-akings)
      - Do nothing
        - (Nothing changes)
  - Do nothing
    - (Nothing changes)
  - Peers apathetic
    - (Nothing changes)
  - Peers disappointed
    - (Reputation worsens)

Costless exogenous reputation improvement
(Reputation improves)

Pay for reputation improvement (volunteering)
(Pay costs \( k \), reputation improves)
Figure 3.1 (preceding page): The NIR decision tree. The probability of each branch is described by blue parameters and green variables ($v$: evolving disposition to pay reputation improvement costs; $s$: evolving disposition to exploit victims with good and bad reputations). Red text at terminal nodes describes the consequences of each outcome.
Models

To help resolve the puzzle of early human cooperation I unpack three stages of complexity that emerge from a more general model of Negative Indirect Reciprocity (NIR). These insights follow from two convergent thought experiments. One possibility is a ‘discrete strategy’ perspective, where we imagine interactions between very different kinds of individuals such as those who always cooperate with well reputed individuals (reputation-conditional cooperators; REPcoop) and obligate defectors (Exploiter). An alternative is a ‘continuous disposition’ perspective (based on the successive invasions assumptions of Adaptive Dynamics (Geritz et al., 1997; Waxman & Gavrilets, 2005)), where we imagine communities of individuals who share very similar, perhaps genetically endowed, dispositions to cooperate. In either case, we can reason formally about what kinds of individuals would be favoured by selection, and both perspectives lead to similar general conclusions.

Here I summarise these key qualitative insights and feature the succinct but informative discrete strategy invasion criteria: the conditions under which obligate defectors (Exploiter) cannot invade a population of obligate cooperators (REPcoop). Figure 3.1 depicts the logic of these models. The Mathematical Model section below contains full details.

Imagine a single, large population of individuals who each have a ‘reputation’—a community-wide opinion that influences others’ behaviour—which can be either ‘good’ or ‘bad’. I represent this reputation as a stochastic variable whose stationary distribution is the probability of being ‘good’ on average.

During their lifetimes these individuals encounter two kinds of opportunities. Sometimes (with frequency $1 - \rho$) they notice a way to exploit a conspecific, yielding some takings ($t$) to the exploiter while damaging ($d > t$) their target. This situation is error-prone: sometimes well-meaning individuals accidentally exploit (probability $\eta$), and sometimes exploitation attempts fail (probability $\phi$). By assuming accidental exploitation is vanishingly rare ($\eta \rightarrow 0$), I present simplified expressions that preserve the model’s essential insights as long as $\eta$ remain small($\eta \sim \frac{1}{20}$). The mathematical model section below provides the full expressions and robustness analyses.
I assume that individuals tend to dislike those who exploit someone they like. That is, exploiting someone with a good reputation causes one’s own reputation to worsen. However I assume that forgoing opportunities to exploit (i.e., cooperating by inaction) carries no consequences. Assuming otherwise would imply that individuals were recognizing that opportunities to exploit existed, assessing that another individual noticed them as well, coordinating their reactions to these counterfactuals as a community, and so on.

Consequently, not-exploiting badly-reputed individuals only ever yields costs and that poorly reputed individuals are always exploited. In this simple model, which focuses on active exploitation, unconditional cooperators never prosper.

Individuals also have opportunities to improve their reputation (with probability $\rho$). My first model—Pure Negative Indirect Reciprocity (NIR-P)—assumes that such improvement is costless and exogenous. Our earliest reputation-using ancestors had no awareness of their own reputation nor how to improve it. However their reputations may still have improved at random after some time, perhaps because their peers had limited memories and eventually forgot their old gripes or because they stumbled upon a non-excludable food resource that their peers gratefully shared.

All the models presented here, including NIR-P, are bistable. They have an uncooperative equilibrium—where selection favours exploiting anyone whenever the chance arises—and a cooperative equilibrium—where selection disfavours exploiting well-reputed individuals. At NIR-P’s cooperative equilibrium, a population of individuals who never exploit well-reputed peers ($\text{RepCoop}$) have, on average, higher fitness than rare individuals who always exploit everyone, so long as

$$\frac{\text{Exploitation inefficiency}}{\text{Exploiter reputation}} < \frac{\text{Exploiter reputation}}{1 - \frac{2\rho}{1 + \rho - \phi + \rho \phi}}$$

The left side of this inequality represents the inefficiency of exploitation; the ratio of benefits gleaned by an exploiter (e.g., thief) to harm caused
to their victim, which is often much greater (e.g., stealing from the weak). NIR is particularly stable in domains where exploitation opportunities are common and their consequences dire (e.g., stealing from the weak, sick and old).

The right represents the difference between cooperators’ average reputations (unity at the cooperative equilibrium when $\eta \to 0$) and the invading defector’s reputation. The possibility of accidental exploitation ($\eta > 0$) makes this inequality harder to satisfy by making cooperators’ reputations slightly worse on average, but does not change these qualitative insights (the details of this possibility are spelled out in the ‘Mathematical Model’ section, below).

Figure 3.2 shows the population frequency of reputation-respecting REP_COOP needed before selection favours more cooperation. When opportunities for exploitation are far more common than opportunities for reputation improvement ($\rho$ is small) and exploitation is inefficient ($\tau$ is small), a few cooperators are enough to trigger a cascade of ecological interactions that lead to a world where only poorly reputed individuals are exploited. A convergent result under a ‘continuous disposition’ perspective implies that, under these circumstances, selection will mould a population only slightly ill-disposed to exploit their better-reputed peers, into highly reputation-sensitive individuals loathe to exploit those with good reputations.

A key challenge for models of cooperation is explaining how a species initially composed of uncooperative individuals could arrive at the cooperative equilibrium’s basin of attraction. Here NIR has an easier time than most other approaches. It is plausible that preadaptations for friendship, coalition-formation and direct reciprocity gave early reputation-users some proclivity to dislike those who harmed their allies before social learning became strong enough to coordinate individual opinions into community-wide reputations. That is, this evolving system may have started within its cooperative basin of attraction, particular if inefficient exploitation opportunities (low $\frac{\tau}{\eta}$) were plentiful (low $\rho$).

Assume that the reputation-sensitive communities described by NIR-P did emerge. As reputations came to carry great fitness consequences,
Figure 3.2: NIR-P basins of attraction and equilibrium reputations. The location of the internal unstable equilibrium that divides NIR-P’s cooperative (above the lines) and uncooperative basins of attraction (left panel); and the equilibrium reputations (right panel) of cooperative RepCoop and Miser (higher, red lines) and uncooperative Exploiter (lower, blue lines) for $\frac{t}{q} = 0.1$ (darkest lines), 0.5 and .0.8 (lightest lines). All errors ($\epsilon, \phi, \eta$) set to $\frac{1}{20}$. When the proportion of non-exploiters is above the threshold demarked in the left panel, selection, on average, favours even less exploitation (i.e., more cooperation).
selection could begin to favour individuals who noticed costly opportunities to improve their reputation and were disposed to act on them. For instance, they might chose to share a resource they could have kept to themselves to make their peers’ sentiments towards view them more favourable. To model this I assume that if an opportunity to improve one’s reputation occurs \((\rho)\), it is sometimes costless and exogenous (probability \(\kappa\)), and sometimes \((1 - \kappa)\) requires the individual to pay a deliberate cost \((k)\). For brevity, I call this ‘volunteering’, since the most interesting cases are those in which these costs raised reputations by contributing to others’ fitness. These error prone volunteering attempts sometimes fail (probability \(\varepsilon\)), but are still costly. I continue to assume that these early reputation-users could not coordinate reactions to inaction, and so ‘not volunteering’ carries no consequences for reputations. NIR-P is a special case of this broader model (i.e., where \(\kappa = 1\)).

My next model—Voluntary NIR (NIR-V)—extends NIR-P by asking just how much costly volunteering the threat of reputation-based exploitation can sustain. I consider two distinct cooperative strategies; both never exploit well-reputed peers, but one always volunteers (RepCoop) and the other never does (Miser). In the formal model below I show that RepCoop have an advantage over Miser when exploitation is inefficient \((t_d \text{ small})\), opportunities for reputation improvement are relatively rare \((\rho \text{ small})\), costly opportunities relatively plentiful compared to costless exogenous improvement \((k \text{ small})\) and these costs are not too great \((k_d \text{ small})\). Here I show the condition for a cooperative, volunteering population (RepCoop) to do better than a rare, uncooperative, un-volunteering mutant (Exploiter). This is easiest to express in terms of the long-run probability that each strategy will be well-reputed \((\pi)\):

\[
\pi_{\text{RepCoop}}^{\text{NIR-V}} = 1
\]

\[
\pi_{\text{Exploiter}}^{\text{NIR-V}} = \frac{2\rho k + (1 - \rho)(1 - \phi)}{2\rho + (1 - \rho)(1 - \phi)}
\]

\[
\frac{t_d}{d} < \left(\pi_{\text{RepCoop}}^{\text{NIR-V}} - \pi_{\text{Exploiter}}^{\text{NIR-V}}\right) - \frac{k}{d} \frac{\rho}{1 - \rho} \frac{2(1 - \kappa)}{(1 - \phi)}
\]
The first term on the right again represents the difference between cooperators’ and defectors’ equilibrium reputations. Now a second term expresses the additional burden of costly reputation improvement. In general, this condition can be satisfied so long as the costs of contributing are not too great ($\frac{k}{d}$ is small) and opportunities to improve reputations are infrequent ($\rho, \kappa$ is small).

Continuous disposition perspectives on NIR-V models (see mathematical model, below) suggest that, if intermediate dispositions to contribute are possible they will be favoured by selection under NIR-V. Selection favours contribution rates that balance the costs of reputation improvement against the benefits of sometimes gainfully exploiting others.

To see why NIR-V is important, consider what ‘volunteering’ represents. Among a reputation-using community, selection favours doing *whatever it takes* to improve your reputation, up to a certain cost ($k$). This could include resource sharing, grooming, or chasing pests away from shared resources, but also includes conformity to others’ preferred behavioural standards and imitation of the best-reputed individuals. This selective pressure for conformity to whatever pleases one’s community could help sustain more sophisticated forms of socially coordinated cooperation. NIR-V provides a plausible cognitive foundation for the emergence of ‘social norms’ (Chudek & Henrich, 2011). Under NIR-V adhering to norms is rewarded, with a higher reputation, but failing to is not punished.

NIR-V also describes plausible cognitive and socioecological preconditions for the emerge of coordinated responses to inaction.

In societies at NIR-V’s cooperative equilibrium—where individuals are disposed to perform costly to please their peers—as individuals becomes ever more attentive to their own opportunities to volunteer they might also notice and respond to others’ volunteering opportunities. If volunteering typically pleases peers, these more reputation-savvy communities may (with probability $\zeta$) be disappointed when someone pass up a volunteering opportunity, causing their reputation to worsen.

Such reputation loss is not a deliberate attempt to punish deviance, it is an emergent consequence of prior selection for cognitive systems that attend
Nevertheless, since low reputations lead to exploitation by many peers, this disappointment coordinates community-wide sanctioning of failures to perform commonly expected behaviours.

My final model—Mandatory NIR (NIR-M)—asks how interactions change as volunteering gradually becomes a norm sanctioned by reputation loss ($\zeta \to 1$). Now NIR-V is a special case (i.e., where $\zeta = 0$). At NIR-M’s cooperative equilibrium, obligate cooperator-volunteers ($\text{RepCoop}$) resist invasions by obligate defector-nonvolunteers ($\text{Exploiter}$) when

$$p_{\text{nirm}}\text{RepCoop} = \frac{1-e(1-k)}{1-e(1-\zeta)(1-k)}$$

$$p_{\text{nirm}}\text{Exploiter} = \frac{2\kappa \rho}{2\kappa\rho + 2\zeta \rho (1-\kappa) + \frac{1}{1-e(1-\zeta)(1-k)}}$$

The conditions for cooperation and volunteering to thrive under NIR-M are similar to NIR-V, however notice that cooperator’s reputations are lower (as $\zeta$ becomes higher), a consequence of accidental failures to volunteer. These lowered reputations inversely weight the burden that volunteering costs place on well-intentioned cooperators, making NIR-M a more favourable environment for non-cooperators community expectations of volunteering rise ($k_d$ and $\zeta$ rise). From a continuous disposition perspective, as reputational enforcement of volunteering increases ($\zeta \to 1$) the equilibrium disposition to volunteer also increases since not doing so carries weightier reputational consequences.

NIR-M, in particular, is capable of sustaining high-levels of costly volunteering when defectors are rarely given free passes back to good reputations (low $k$) as nearly everyone must conform to behavioural standards to sustain their reputation (see Figure 3.9). At this equilibrium an observer would see an apparently naturally harmonious society (i.e., with little exploitation) and even high-levels of prosocial volunteering, or norm compliance. Meanwhile, hidden from view, such cooperation would be sustain by rare but costly exploitation of the poorly reputed. Detailed ethnographic work by
Henrich suggests this mechanism may be important in small-scale societies (Henrich & Henrich, forthcoming).

**Discussion**

Together these models map a path from minimal cognitive perquisites to larger-scale forms of human cooperation by first suppressing within-group exploitation—such as theft or rape—and then harnessing it to sustain cooperative contributions to public goods—such as meat sharing or communal defence. The logic of this process and testable predictions it implies are presented in Figure 3.3.

NIR-P describes dynamics when reputational systems first emerge: if community members are sufficiently reluctant to exploit their well-reputed peers, selective forces will sustain and enhance this reluctance, perpetuating harmonious (i.e., non-exploiting) communities. This is particularly likely if exploitation opportunities are common and benefit the perpetrator little relative to the harm they cause the victim.

Once harmonious communities exist NIR-V can emerge. That is, selective pressures can dispose individuals to make some costly, reputation-improving contributions to others welfare (which I have called ‘volunteering’). To do this they must develop an awareness of what behaviours would please others on average. This puts a community’s normative behavioural expectations on the selective landscape. Ironically, it is the central challenge of NIR—that ‘negative cooperation’ is typically unobservable and so cannot reliably improve reputations—that leads to pressure for the cognitive abilities assumed by models of latter forms of cooperation—that individuals can recognise and rapidly coordinate on arbitrary shared norms.

Once NIR-V dynamics lend potent fitness consequences to shared expectations, and cause individuals to sometimes (as long as it is not too costly) do whatever it takes to satisfy those expectations, NIR-M can push communities even closer to full-blown social norms and a psychology for navigating them. If individuals come to expect reputation-raising acts and are somewhat disappointed by their absence, failure to volunteer can actually lower
one’s reputation. NIR-M shows that this strengthens selective pressures for adherence to community expectations, by providing a larger reputationally-distributed stick for their enforcement.

To thrive in the social-ecology described by NIR-M individuals must be quick to perceive their community’s norms—the behaviours that please others on average, which could include generosity in times of plenty, sharing adaptive knowledge or wearing trendy running shoes—and be disposed to adhere to them. Communities meanwhile, come to wield a powerful means of enforcing compliance to these norms. This distributed mechanism for norm-enforcement emerges without any individuals intending it; they merely selfishly exploit friendless, low-status victims when the opportunity arises because they know they can get away with it. Indeed we may still witness these dynamics today, as the recurrent emergence of schoolyard bullying recapitulates the socio-ecological dynamics of early, pre-institutional human societies (Card et al., 2008; Merrell et al., 2008; Rodkin & Berger, 2008).

The ‘volunteering’ norms that emerge under NIR-M could be prosocial acts, but need not be. In fact any arbitrary and even maladaptive community norm could be sustained by this mechanism. There are two reasons to suspect that over time such volunteering would become increasingly prosocial. First, improving others’ welfare is particularly likely to raise their opinion of you. The creates a what cultural evolutionists have termed a content bias favouring prosocial reputational content. Second, by making deviation from local community expectations costly, NIR-M, favours migrants who adopt the norms of their new community rather than maintaining their old behaviours. This decreases behavioural variability within groups relative to variation between them, increasing the between-group selective pressures for norms that lead to success in intergroup competition, which may include contributions to defense, raiding, economic productivity, alliance building, trading and information sharing (Chudek & Henrich, 2011).

These cognitive and socio-ecological conditions make it far easier for even more potent, coordinated or institutional, forms of cooperation to emerge. By reputationally rewarding those who share valuable cultural information, NIR also untangles the cooperative dilemmas that would otherwise prevent
the emergence of a culture-sharing species.

The reputational system postulated by NIR imposes minimal cognitive demands on early reputational cooperators, since they can ignore (1) anything that happens to people in bad standing, (2) all ‘non-events’ (like not exploiting), and (3) the exploiter’s previous reputation.

In ancestral human societies, when individuals fell sick, were injured, or faced emergencies requiring them to rapidly leave camp, exploiters had opportunities to steal food, mating opportunities, allies, beads, and raw materials (skins, flint, ochre, and obsidian) with little chance of direct retribution—either because the victim could not pinpoint the perpetrator or was in no position to enact revenge. In times of distress (illness or injuries) exploitation is particularly easy and the loss of valuable resources is particularly damaging.

These exploitative opportunities were likely distributed more or less at random, as my models assume. In contrast, opportunities for positive cooperation likely covary with fitness, with fitter individuals having more chances to help.

The early stages of the emergence of generalised, high-fidelity cultural learning (Henrich & Henrich, 2007) may have provided the cognitive foundations for individual friendships, coalitions and opinions of others to transmit socially, becoming ‘reputations’. I hypothesise that these early reputations interacted with random exploitation opportunities to shape social cognition and lay the foundations for the latter evolution of human cooperation and culture.

Because the evolutionary story outlined has been carefully mathematically formalised to ensure that it is consistent with the constraints imposed by natural selection. One benefit of such formalisation is that it makes the assumptions and predictions of the model transparent, allowing them to be precisely tested. In Figure 3.3 I expound the testable claims entailed by each step in the NIR account. These come in two forms: assumptions and predictions.

NIR is motivated by and grounded in several assumptions about our ancestors’ lives and interactions. For instance, several critical assumptions are
instantiated by NIR’s reputation-assessment rules. Like other IR models, NIR assumes that people track the reputations of those they interact with, that the way they treat those people is influence by their reputation, and that those people’s reputations are affect by the way they treat others. If these facts were true of our ancestors then (ceteris paribus) they ought to be true of their modern descendants. If we didn’t observe these kinds of reputation dynamics in contemporary humans, we would have strong reason to doubt that NIR and other IR models are good descriptions of our ancestors’ interactions.

What sets NIR apart from other IR models is its emphasis on exploitation, that is: cooperation by deliberate, observable action (and not inaction) in negative cooperative dilemmas. NIR assumes that these kinds of interactions played a particularly important role in ancestral reputational and evolutionary dynamics. Contemporary populations do show robust negativity- (Baumeister et al., 2001; Cacioppo & Berntson, 1994; Rozin & Royzman, 2001) and omission-biases (Baron & Ritov, 2004; Cushman et al., 2006; DeScioli et al., 2011; Ritov & Baron, 1999; Spranca et al., 1991). That is, their reputations are particularly sensitive to defection in negative dilemmas.

NIR’s testable claims are still more specific. NIR expects these biases to manifest in second order reputational interactions in particular. That is, the difference in reputational consequences when one interacts with a well or poorly-reputed target individual ought to be more pronounced when one defects in negative dilemmas, than in other kinds of cooperative interactions. As far as I know, this more specific claim has not been tested prior to the evidence presented in chapters 4 and 5.

If we observe the psychological phenomena assumed by NIR, we can confidently claim that we have tested NIR. We cannot, however, claim to have explained the phenomena. NIR does not explain why these conditions exist, it merely depends on the assumption that they do.

On the basis of these assumptions (and others outlined in Figure 3.3) and the logic of natural selection, NIR also makes predictions about how individuals adapted to an NIR-ecology ought to behave. These behaviour and phenomena, if we observe them, are explained by NIR at an ultimate level.
NIR tells us, for instance, why individuals ought to be less inclined to exploit better-reputed peers (by showing the plausibility of the ecological conditions under which selection minimise the this disposition), pay arbitrary costs to improve their reputation, or conform to arbitrary community norms. However testing these claim in modern populations is tricky. First, because they are quite broad. Many other models also predict a world where reputations matter and people follow and enforce norms, and it many different processes may have contributed to this outcome. Second, because NIR provides the bedrock for more complex forms of cooperation and so anticipates that our contemporary patterns of cooperation have been altered by many thousand of years of evolving cultural institutions. That is, NIR is a model of how cooperation first began, but not necessarily of how (except in special cases) it is maintained today.

The psychology and behaviour of contemporary humans are better suited to testing NIR’s quite specific assumptions, than its more general predictions. That said, NIR could be clearly tested and its contribution to the foundations of human cooperation disambiguated from other mechanisms’, if we could identify contemporary societies that were not governed by long-evolved institutions and could recapitulate the socio-ecological dynamics played out by our ancestors. The schoolyard societies continually rediscovered by generations of contemporary children, foisted by modern educative institutions into the daily company of their equally young and naive peers, my provide just such an opportunity.

The remainder of this dissertation kick-starts the more tractable project of testing NIR’s specific assumptions. In chapter 4, I test whether people’s reputational assessments of others show the negativity- and omission-biased patterns of second-order indirect reciprocity predicted by NIR. In chapter 5, I assess whether a peculiar, recently-observed anomaly in people’s moral reasoning can be explained by these same assumed conditions.
Figure 3.3: NIR’s logic, testable assumptions and predictions
3.1 Mathematical model

In this section I will once again walk through the logic of NIR. However, this time as I do I will explicate a mathematical model of the ecological interactions it engenders and their long-term evolutionary consequences. At each step, I will clearly lay out the mathematical assumptions that underlie each piece of verbal reasoning and their implications.

3.1.1 Context and overview

The simultaneous emergence of human cooperation and culture is a puzzle. Humans are able to share complex, encoded information about their world. We call this information ‘culture’. For somewhere between a few hundred thousand and a few million years, this corpus of inter-generationally transmitted know-how has been accumulating and evolving. Today it contains impressive concepts and skills like ‘science’ and ‘opera’. Simultaneously, humans have begun cooperating with each other on scales rarely seen outside of eusocial species. We often chose to endure a relative individual cost to bring about a relative benefit for one or many others.

The emergence of cooperation is a well-known evolutionary puzzle. Among humans, culturally evolved institutions (such as police forces, reputations and library fines) make it much easier to explain how cooperation can evolve than in non-cultural species. A less well-know puzzle is how cultural-transmission itself emerges. Particularly central to this puzzle is the ‘cooperative dilemma of culture’. Members of a cultural species must trust information they receive from others. This however makes it easy for others to exploit them by distorting that transmitted information to their own benefit. These benefits should select for ever more cultural exploitation, until it is no longer adaptive to culturally learn from others. How, before the emergence of cooperation-sustaining cultural institutions, did the quality of cultural information stay high enough, for long enough, for our sophisticated cultural learning capabilities to evolve?

One step towards untangling this conundrum is establishing how simple, cooperation-sustaining reputational systems can emerge from simpler pri-
mate preadaptations. Several models have demonstrated how ‘reputations’ can sustain arbitrary forms of cooperation. However, these models implicitly assume a strong form of cultural learning. Specifically: that communities agree on how to interpret one another’s inactions. They assume that sometimes an individual faces an opportunity and chose to do nothing and that this will prompt a coordinated response from their social peers. This implies that others perceive the same opportunity and share a cognitive representation of what ‘should’ be done in such situations. This is quite a sophisticated capacity to ascribe to our early reputation-judging ancestors, and it is not obvious whether it could arise without a ‘cooperative dilemma of culture’.

In this model I try to establish whether a reputational system for sustaining cooperation could emerge even when inaction (i.e., not helping in positive dilemmas, not exploiting in negative ones) does not change reputations.

We focus on the negative cooperation context (i.e., cooperating by ‘not exploiting’ someone, rather than ‘helping’ them) for the reasons detailed in the main text.

To avoid the potential confusion and ambiguity of talking about negative costs and benefits, I describe these dilemmas in terms of the damage ($d$) done to victims and the takings ($t$) earned by exploiters.

We define Negative Indirect Reciprocity (NIR) as a system of interacting individuals, with these properties:

### Assumptions.

1. Individuals have regular opportunities for negative cooperation (i.e., exploitation).

2. Individuals have reputations. That is, others assign them a status of good or bad.

3. Inaction (i.e., cooperating by not-exploiting) does not change reputations.
4. Exploiting ‘good’ individuals worsens one’s reputation.

The final assumption seems to be a requirement common to any reputational system capable of sustaining cooperation (Ohtsuki et al., 2006).

There are three possible reputational rules we could add to this: exploiting bad individuals could either improve, worsen or not change one’s reputation. Here I consider what I consider the case most plausible given primate pre-adaptations: exploiting bad individuals does not change one’s reputation. This is the most plausible starting point since it requires no special cognitive adaptations, just ignorance or apathy about whatever happens to ‘bad guys’.

This opens up the possibility of interpreting ‘good reputation’ as meaning ‘people you care about’, since all actions towards badly reputed targets are ignored. This interpretation provides the explanatory bridge between reputations and primate cognitive preadaptations for forming coalitions, as described in the main text.

Assumptions.

5. Exploiting bad individuals does not change reputations.

But how do reputations improve? We consider two possibilities (see decision trees in Figure 3.1). First, reputations could improve at random (with probability $\kappa$). Second, they could improve when an individual makes a deliberate, costly effort to improve them (at cost $k$).

3.1.2 Model definition

Imagine a large population of individuals; call them a ‘community’. Each individual is born, interacts many times with their peers, reproduces and
dies. Their interactions are shaped (on average, in the long run) by their (and others’) genetically transmitted behavioural dispositions. These dispositions can be characterised as either discrete strategies (Discrete Strategy approach, below), or as continuously varying traits, whose average value is subject to natural selection (Adaptive Dynamics approach, below). Both yield convergent insights.

To express the continuous interpretation we use lower case latin variable names (specifically $s, v$) to talk about a focal individual’s disposition, and upper case ($S, V$) to refer to the population average of that same disposition.

Individuals vary in their disposition to a) exploit a well-reputed peer (variables $s/S$), and b) to take costly actions to make others like them more (i.e., to improve their reputation; variables $v/V$). It is easy to show that, within this model, selection will always maximise dispositions to exploit badly-reputed individuals. Verbal logic suffices: not exploiting a ‘bad guy’ involves sacrificing a material benefit ($t$), but never yields any reputational benefit. By assuming that badly-reputed individuals are always exploited, our model focuses on the less immediately obvious dynamics of exploiting ‘good guys’.

To express the discrete interpretation, we merely consider individuals who’s dispositions are set to their boundaries. Specifically, obligate defectors (exploiter: $s \rightarrow 1, v \rightarrow 0$); miserly cooperators (miser: $s \rightarrow 0, v \rightarrow 0$) who do not exploit others but do not take costly actions to improve their reputation (i.e., by improving others’ welfare) either; and reputation-based cooperators (RepCoop: $s \rightarrow 0, v \rightarrow 1$) who never willingly exploit and always try to improve their reputation. As in the continuous version, we do not consider strategies that cooperate with poorly-reputed individuals since their payoff must be strictly smaller than RepCoop’s.

The outcomes of these interactions cumulate to influence the relative frequency at which individuals genetically transmit these dispositions to the next generation. Specifically, all else being equal, an individual’s reproductive fitness is proportional to the sum of a) their takings ($t$) when exploiting others, b) the damage ($d$) they incur from others’ exploitation and c) the costs ($k$) they pay to improve their reputation.
We assume that ‘opportunities to improve reputations’ and ‘opportunities to exploit others’ occur with relative frequency $\rho : (1 - \rho)$. We assume the relative rate of these two opportunities is an independent parameter in our model, thereby implicitly assuming that the dispositions we are modelling do not affect it (e.g., Exploiter is not more likely to face an opportunity to exploit others than RepCoop is).

**Assumptions.**

6. *Behavioural dispositions that lead to higher fitness become more frequent over time.*

7. *Individuals can have a different disposition to exploit peers who they like and peers who they don’t like. That is, reputation influences cooperative behaviours.*

8. *The probabilities of having an opportunity to exploit, be exploited or improve reputations are uncorrelated with anything else in our model.*

We take as our starting point a world where individuals observe and imitate one another’s opinions of one another, such that these opinions quickly converge on a community-wide consensus. We call this the individual’s ‘reputation’. An unanswered and interesting question is how a community of individuals can evolve reputations in the first place. We do not attempt to answer that here, but suspect that it involves early adaptations for cultural learning, such as prestige (Henrich & Gil-White, 2001) or conformity (Boyd & Richerson, 1988) biases. These can produce substantial asymmetries in how likely the most influential individuals’ or the majority’s opinions are of being transmitted to others. To keep the present model simple, we consider the limiting case where there is community consensus on reputations.

To keep our model as simple as possible (without sacrificing insight), we
Rates

\[ \frac{\rho}{1 - \rho} \in (0, 1) \]  Relative frequency of \( \frac{\text{reputation improvement}}{\text{exploitation}} \) opportunities.

\( \kappa \in (0, 1) \)  Probability of reputations improving costlessly

Errors

\( \phi \in (0, 1) \)  Accidental cooperation (doing nothing)

\( \eta \in (0, 1) \)  Accidental defection (exploiting)

\( \varepsilon \in (0, 1) \)  Failure to improve reputation, despite paying costs

Community Reactions

\( \zeta \in (0, 1) \)  Probability that ‘not helping’ worsens reputation.

Table 3.1: Summary of parameters

assume that reputations can be in two states (good or bad). This makes sense, because though the actual underlying psychological representation of an individual’s reputation may be continuous, when individuals face a decision about exploitation (at a given \( t \) and \( d \)) or costly reputation improvement (at a given \( k \)), they must ultimately make a binary choice.

By reasoning about who has higher fitness when (i.e., the sum consequences of these interactions), we can draw inferences about which dispositions are likely to evolve under which circumstances. To do this we need to figure out what an individual with certain dispositions would experience in a population of individuals with other dispositions.

3.1.3 Reputational dynamics

Our first step is to lay out all the circumstances under which an individual’s reputation might change, weighted by their probabilities of occurring. This
is depicted visually in Figure 3.1.

First we define the probability, when a reputational change happens, that a focal individual’s reputation will become good ($P_g$) or become bad ($P_b$). Then, we can derive the stationary distribution of this two-state stochastic process as the relative probability that the focal individual’s reputation becomes good ($G = \frac{P_g}{P_g + P_b}$).

To do this, we will need to define parameters that represent all the different things that we’ve said can occur. We have already represented the relative rate at which opportunities to improve reputations arise, relative to opportunities for exploitation ($r_1/\rho$). We also know the probability that an individual’s reputation improves for extrinsic reasons, without them needing to take any action at all ($k$). If their reputation does not improve by such good fortune ($1 - k$), we have also assumed that the (average) probability that an individual will pay a cost to improve it is a consequence of their evolved disposition ($v$). But what are the chances that this costly attempt hits it mark? Let’s assume that deliberate attempts to improve reputations fail with probability ($\epsilon$). Putting this all together, the relative probability of one’s reputation improving is

$$P_g = \rho (k + (1 - k)v(1 - \epsilon))$$

What about worsening? Reputations worsen (in our model) when a focal individual exploits someone with a good reputation. We know that exploitation opportunities arise with probability ($1 - \rho$), but the focal individual will be the potential exploiter (not the potential victim) in only some of these (half of them, by assumption 8). Next their reputation will only worsen if their potential victim has a good reputation. Again by assumption 8, this will be (on average) the frequency of individuals with good reputations in the population (for now let’s call this variable $G$, later we will solve it’s value). Next, they need to be inclined to exploit their victim, which we have assumed happens according to their evolving disposition ($s$). Even if they chose not to exploit this well-reputed target ($1 - s$), there might still be some possibility ($\eta$) they do so by accident. Finally, even though the focal indi-
individual might want (and even try) to exploit someone, there might be some probability ($\phi$) that they fail entirely and accomplish nothing (including not worsening their reputation).

Do individuals who decide not to take a costly action when the opportunity arises ($1 - v$) experience a worsened reputation? This would imply that the community had a coordinated response to their inaction, and thus well-coordinated expectations about each other’s behaviour. This is the very assumption we are trying to avoid, so at first we want to model a world where their reputation does not change. As we will see, the resulting socio-ecology is one that favours an increasingly shared awareness of reputation-improving opportunities, and so we also want to see how our model changes as this shared awareness gradually increases. So, we include a parameter ($\zeta$) in our model that describes the probability that ‘not taking costly actions’ (or not ‘volunteering’, for brevity) actually worsens reputations.

Putting this all together we have:

\[
P_b = \frac{(1-\rho)}{2} G (s + (1 - s) \eta))(1 - \phi) + \zeta \rho (1 - \kappa)((1 - v) + v \epsilon) \\
\mathcal{G}(G) = \frac{P_r}{P_r + P_b} = \frac{2p(1-\epsilon)(1-\kappa)+\kappa)}{2p(1-\epsilon)(1-\zeta)(1-\kappa)+\kappa+\zeta(1-\kappa)+\epsilon(1-\kappa)+\eta(1-\rho)(1-\phi)}
\]

We express the stationary distribution as a function of $G$, the probability of interacting with a well reputed individual, whose value we do no know yet.

This general expression makes it easy for us to explore specific cases. For instance:

- a world before people start deliberately improving their reputations ($\kappa \to 1$),
- a world where failing to meet community expectations does not worsen reputations ($\zeta \to 0$), and
• a world where it does ($\zeta > 0$),

• domains in which exploitation is easy ($\phi \to 0$) or hard ($\phi \to \frac{1}{2}$), and so on.

Another approach to reputations

Rather than parametrising the relative rates of opportunities, another approach to modelling change in reputations is to assume that each individual confronts the same ordered sequence of interactions (for instance, see Panchanathan & Boyd, 2004). First they might have an opportunity to improve their reputation, followed by one or more opportunities to worsen it by exploiting others. This approach implicitly invokes assumption 8 by assuming that there is a constant probability (let’s call it $\omega$) that another opportunity to exploit arises before an opportunity to improve one’s reputation does. Consequently each subsequent exploitative opportunities’ consequences must be weighted by its ever smaller chances of occurring. The geometric sum of these diminishing probabilities converges to a single number ($\frac{1}{1-\omega}$, if the first exploitative event is certain, $\frac{\omega}{1-\omega}$ if it also occurs with probability $\omega$). The modeller can then derives the relative rate (and weights the consequences) of reputation improvement to exploitation. Specifically, $1 : \frac{1}{1-\omega}$ or $1 : \frac{\omega}{1-\omega}$.

We feel it’s simpler to merely parametrise this rate and find the stationary distribution. However the solution is identical in both cases. To translate between these two approaches, use this map: $\rho = \frac{\omega-1}{\omega}$ if the first exploitation opportunity is certain, or just $\rho = 1-\omega$ if it is not.

It is also worth noting that in this model I allow reputations to improve exogenously (i.e., with probability $\kappa$), but they only ever fall endogenously (i.e., when someone exploits someone else). This allows me to focus our model more tightly on the dynamics of exploitation and simplifies some of the math, at the cost of creating some potentially degenerate reputational equilibria (i.e., special cases which may not be true in general) at the boundaries of our system. Specifically, in the absence of accidental exploitation, individuals who refuse to exploit good guys (i.e., $s \to 0$) have perfect repu-
tations. I have explored models where reputations also fall at random, and their dynamics and qualitative implications do not differ from those of the simpler, more tractable model presented here.

3.1.4 Behavioural dynamics

Next, we need to say something about what the fitness consequences of a focal individual’s actual behaviour are. Let’s go piece by piece, defining the fitness consequences of improving one’s reputation ($W_1$), exploiting ($W_2$) and being exploited ($W_3$), before combining them ($W = W_1 + W_2 + W_3$). Again, we assume that the probability of encountering an opportunity to exploit someone with a good reputation is ($G$). Now we also need to be explicit about the probability that a randomly sampled member of the community will exploit the focal individual, given that the focal individual has a good reputation. In our model this is just the mean community disposition to exploit ($S$).

($W_1$): Given the opportunity to do so ($\rho$), an individual so disposed ($\nu$) may pay a cost ($-k$) to attempt to improve their reputation. We assume they pay this cost even if their attempt fails ($1 - \epsilon$).

($W_2$): Given an opportunity to exploit (which occurs with relative frequency $\frac{1 - \rho}{2}$) a focal individual could earn some takings ($t$) and inflict some damage ($d > t$) if they meet someone with a bad reputation ($1 - G$), or if they meet someone with a good reputation ($G$) and are disposed to exploit them ($s$) or do so accidentally ($\eta$), and don’t mess it up ($1 - \phi$).

($W_3$): The same is true of other individuals exploiting the focal individual, given the focal individual’s reputation. To reason about the average fitness consequences of these events, we need a way of referring to the average reputational state of a focal individual over their lifetime (let’s call it $g$ for now).
All together this yields:

\[ W_1 = (-k)\rho (1 - \kappa)v \]

\[ W_2 = (+t)^{\frac{1 - \rho}{2}} (1 - \phi)((1 - G) + G(s + (1 - s)\eta)) \]

\[ W_3 = (-d)^{\frac{1 - \rho}{2}} (1 - \phi)((1 - g) + g(S + (1 - S)\eta)) \]

\[ W(g, G) = W_1 + W_2 + W_3 \]

### 3.1.5 Combining reputations and behaviour

Now we have two dynamics models. One tells us how reputations change and the how often an individuals will have a good reputation \((G)\) as a function of their behavioural dispositions \((s, v)\), their likelihood of interacting with a well reputed peer \((G)\). We also have a model of the fitness consequences that shape how those behavioural dispositions evolve \((W)\), which is also depends on the likelihood of meeting a well-reputed community member. Our next step is to connect these models, by assuming that the chance of meeting someone with a good reputation \((G)\) is well approximated by our model of how often your community members have good reputations \((G)\).

#### Assumptions.

9. The long-run probability of encountering someone with a good reputation \((G)\) is well approximated by the mean of the stationary distributions of the reputations of the individuals in the population (i.e., \(G \sim \mathcal{F}(G)\)). This implies that reputational dynamics are much faster than the dynamics of behavioural dispositions.

Given this assumption, when the population is monomorphic (i.e., ev-
eryone has the same dispositions: $V$ and $S$), we can easily solve $G$ as:

\[ G = G(G) \]

\[ G = G(G) \]

Note that there are two quadratic solutions to this equation, but one is always outside the logical bounds of $G \in [0,1]$; the other is presented above in Equation 3.1.

How good is this approximation? We can get some traction on this question by simulating these dynamics. Included with this document is a script for the statistical modelling computer system $R$. It first creates a populations of $n$ RESIDENT individuals with dispositions $S$ and $V$, and reputations set randomly to either good (1) or bad (0). It then randomly selects an individual, randomly gives them either an opportunity to improve their reputation or exploit someone, randomly selects an exploitation target from the same population and plays out the consequences exactly as described above. It also creates a population of $n$ 'rare' INVADEERS (with dispositions $s$ and $v$), and carries out the same process for them, with the one exception that their exploitation targets are chosen from the RESIDENT population rather than their own. After $n$ iterations, we observe the mean frequency of individuals with good reputations in each population ($G$ and $g$ for RESIDENTS and INVADEERS respectively). We do this $t$ times, such that (in expectation) each individual has faced $t$ opportunities for reputation change. Using the attached script you can readily and rapidly repeat this for any combinations of parameters. These simulations show that $G$ and $g$ reliably and rapidly converge on $G(G)$ in any reasonably sized community who interact many times. A typical simulation run is presented in Figure 3.4.

Below we use a similar technique to find $G$ when the population is polymorphic (comprised of multiple discrete types at known frequencies).
Figure 3.4: Simulations showing the accuracy of analytic predictions about Resident and Invader equilibrium reputations. The theoretically predicted (solid lines) and simulated (dots) cumulative proportion of time that individuals have had a good reputation. Here the population contains 100 common Residents and the average of 100 different trajectories of rare Invaders who interacted with them. All errors and $\kappa$ are set to $\frac{1}{20}$, $\zeta$ set to 0 (i.e., NIR-V) and $\rho$ set to $\frac{2}{5}$. Residents traits are $S = .8, V = .1$, while Invaders have $s = .5, v = .2$. During each ‘turn’, as many individuals are randomly selected to experience a potentially reputation changing events as there are individuals. Consequently, in expectation, each individuals experienced as many events as there are turns.
3.1.6 Discrete strategy approach

Now that we have formally described the reputational and behavioural interactions, we explore two approaches to reasoning about their long-term outcomes. The first is the invasion analysis technique of evolutionary game theory. In this (mathematically aided) thought experiment, we imagine populations comprised of the three kinds of individuals described above: **Exploiter** ($s = 1, v = 0$), **Miser** ($s = 0, v = 0$) and **RepCoop** ($s = 0, v = 1$).

Since evolution started from uncooperative communities, we first imagine a population comprised entirely of **Exploiter** (the residents) and ask whether a rare cooperator (the invader) could thrive there. We next imagine a population entirely comprised of the two kinds of cooperators and ask whether an **Exploiter** could thrive there. While this approach is useful for gleaning simple and useful insights into these boundary cases, we will also want to ask how large a fraction of cooperators are need before they start to have an advantage, on average, over **Exploiter**.

To get started, let’s define the population frequencies of **RepCoop** ($x$), and **Exploiter** ($y$). This implies the frequency of **Miser** ($1 - x - y$). Next, let’s express the stationary distribution of the reputation and fitness of a randomly sampled individual practicing each strategy:
\[ G_{\text{RepCoop}}(G) = \mathcal{I}(G)\big|_{v=1,s=0} \]
\[ G_{\text{Miser}}(G) = \mathcal{I}(G)\big|_{v=0,s=0} \]
\[ G_{\text{Exploiter}}(G) = \mathcal{I}(G)\big|_{v=0,s=1} \]
\[ G_{\text{mixed}}(G) = xG_{\text{RepCoop}}(G) + yG_{\text{Exploiter}}(G) + (1-x-y)G_{\text{Miser}}(G) \]

\[ W_{\text{mixed}}(g) = W(g,G_{\text{mixed}}(G))\big|_{v=x,s=y} \]
\[ W_{\text{RepCoop}} = W_{\text{mixed}}(G_{\text{RepCoop}}(G))\big|_{v=1,s=0} \]
\[ W_{\text{Miser}} = W_{\text{mixed}}(G_{\text{Miser}}(G))\big|_{v=0,s=0} \]
\[ W_{\text{Exploiter}} = W_{\text{mixed}}(G_{\text{Exploiter}}(G))\big|_{v=0,s=1} \]

Again, by assuming that the stationary distribution of reputations is actually a good approximation to the long-run average probability of meeting someone with a good reputation, we can readily solve for \( G \).

\[ G = G_{\text{mixed}}(G) \]

Again, we can use a similar simulation (now with a mix if three strategies rather than an Invader and Resident) to check the accuracy of this approximation. Again, we find it very accurate. A sample simulation run is presented in Figure 3.5 and the included simulation code allows you to readily explore any parameters you please.
Figure 3.5: Simulations showing the accuracy of analytic predictions about equilibrium reputations of three discrete strategies. The theoretically predicted (solid lines) and simulated (dots) cumulative proportion of time that individuals of three discrete types (distinguished by color) have had a good reputation, and the population average (black). Here a population comprised of 33 individuals of each type was simulated, with all errors and $\kappa$ set to $\frac{1}{20}$, $\zeta$ set to 0 (i.e., NIR-V) and $\rho$ set to $\frac{2}{5}$. During each ‘turn’, as many individuals are randomly selected to experience a potentially reputation changing events as there are individuals. Consequently, in expectation, each individuals experienced as many events as there are turns. Here ALLD refers to EXPLOITER.
While we have been able to derive a precise expression for this quantity, it is long, complicated and does little to hone our insight on its own. Instead, we pursue two techniques for extracting simpler intuitions about this evolving system. First, we can numerically map out these expression to visually glean their implications (presented and discussed further below).

Second, we can gain useful insights by analytically exploring the conditions under which a member of a monomorphic population (the Resident) has higher fitness that a rare mutant of another type (the Invader). To do this we assume populations are large enough that we can assume Invaders never interact with their own kind without distorting our conclusions.

**Assumptions.**

10. Populations are large enough that we can accurately approximate dynamics by ignoring the impact of rare mutants on plentiful residents.

**Invasion Analysis: Exploiter**

Our model shares a common (and intuitively correct) feature of all models of the evolution cooperation. Cooperative strategies (i.e., Miser and RepCoop) cannot (at least deterministically) invade a population of unremitting defectors (Exploiter). Their kindness is costly but never rewarded (not even indirectly) when they are rare.

To see that this is the case in all the possible worlds we are modelling, we inspect the conditions for these two strategies to invade a population of Exploiter.

\[
W_{\text{Miser}} |_{x=0, y=1} > W_{\text{Exploiter}} |_{x=0, y=1}
\]

Which simplifies to the inequality

\[
\rho((1 - \zeta)\kappa + \zeta) > \sqrt{\rho^2((1 - \zeta)\kappa + \zeta)^2 + 2\kappa(1-\rho)\rho(1-\phi)}
\]
Since the RHS is at least as large as \( \sqrt{\rho^2((1 - \zeta)\kappa + \zeta)^2} \), with an additional term added, this inequality is strictly false. Thus, Miser cannot invade Exploiter. Notice however that this additional term is proportional to \( \kappa \). That is, if \( \kappa \) is zero, a special degenerate case results where Miser has the same fitness as Exploiter at the Exploiter monomorphic equilibrium. This is because without random reputational improvement Exploiter has a zero reputation at equilibrium and so Miser never forgoes an opportunity to exploit Exploiter. If Exploiter’s reputation is perturbed even slightly from zero, Miser sometimes considers them good and refrains from exploiting them, but their mercy goes unrewarded in an Exploiter population.

Next, consider RepCoop invading Exploiter:

\[
W_{\text{RepCoop}} \big|_{x=0,y=1} > W_{\text{Exploiter}} \big|_{x=0,y=1}
\]

We can simplify this inequality as

\[
\sqrt{\rho^2(\zeta^2 + (1 - \zeta)^2 \kappa^2) + 2 \kappa \rho ((1 - \zeta)\zeta \rho + (1 - \rho) (1 - \phi))} < \rho ((1 - \zeta) \kappa + \zeta)
\]

and again readily see that it is never true.

So, a population of Exploiter is stable against invading cooperators.

**Invasion Analysis: Miser and RepCoop**

Miser is able to invade RepCoop when

\[
\frac{d}{\kappa} < \frac{2 \rho (1 - \kappa (1 - \xi)) \left(1 - \frac{\lambda (\sigma)}{\eta(1 - \phi)(1 - \rho)}\right)}{(1 - \eta)(1 - \phi)(1 - \rho) \left(\frac{\omega (\sigma)}{\eta (1 - \phi)(1 - \rho)} - \frac{2 \rho(1 - \zeta)\kappa \zeta + \zeta}{\rho (1 - \epsilon)(1 - \kappa)}\right)}
\]

\[
a_2 = \sqrt{\rho^2(1 - \epsilon(1 - \zeta)(1 - \kappa))^2 + 2 \rho \eta (1 - \epsilon(1 - \kappa))(1 - \phi) (1 - \rho) - \rho (1 - \epsilon(1 - \zeta)(1 - \kappa))}
\]

Meanwhile RepCoop is able to invade Miser when
\[
\frac{d}{\kappa} > \frac{2\rho(1-\kappa(1-\zeta))\left(1-\frac{\zeta(\zeta-k\kappa)}{\kappa(1-\phi)}\right)}{(1-\rho)(1-\phi)\left(2\rho(1-\zeta)\kappa - \zeta - \zeta_k\kappa(1-\kappa)\right) + 2\rho(1-\zeta)\kappa - \zeta - \zeta_k\kappa(1-\kappa)}
\]

\[
a_3 = \sqrt{\rho \left(\zeta + \kappa - \zeta(1-\kappa)\right)^2 + 2\eta \kappa(-1+\rho)(-1+\phi)}
\]

\[
a_4 = \rho(\zeta(1-\kappa) + \kappa)
\]

These inequalities define four regimes of interactions between Miser and RepCoop strategies. Either (a & b) one and not the other monomorphic population is stable, (c) both are stable (in which case an internal unstable equilibrium must exist between them) or (d) both are unstable (in which case an internal, stable, polymorphic equilibrium exists, though this only occurs in a very small region of the parameter space). Figure 3.6 provides a visual intuition into when these four regimes occur in terms of parameters \(\rho\), \(\kappa\) and \(\zeta\).

To summarise, Miser has a relative advantage when opportunities for reputation improvement are common relative to opportunities for reputation loss (\(\rho\) is high), reputations often rise exogenously (\(\kappa\) is high) and communities do not yet expect reputation-improving acts (\(\zeta\) is low), and so Miser individuals suffer little reputation loss for failing to volunteer. The relationships of the payoffs can readily be seen on the LHS of the inequalities. Miser thrives when exploitation causes little damage (\(d\) is small) and reputation improvement is costly (\(k\) is large).
\( \zeta = 0 \) (NIR-V)  \( \zeta = \frac{1}{2} \)  \( \zeta = 1 \) (NIR-M)

**RepCoop** (blue) & **Miser** (red) resist
**Exploiter**

‘**RepCoop** resists **Miser**’ (blue) and ‘**Miser** resists **RepCoop**’ (red)

**RepCoop** stable (blue), **Miser** stable (red) and both unstable (yellow)
Figure 3.6 (preceding page): Invasability plots for the three discrete strategies in NIR. Shaded regions show whether monomorphic populations can resist invasion by rare mutants. The top row shows when RepCoop (blue) and Miser (red) resist Exploiter invaders. The middle row shows when these two strategies resist invasion against each other, and indicates regions of bistability (i.e., overlapping regions) and stable mixed populations (i.e., unshaded regions). By superimposing the upper two regions, the bottom row shows when RepCoop and Miser resist invasion by both others strategies, and when Exploiter invades both (yellow). In the small unshaded regions in the bottom row, strategies resist. In each sub-figure, the horizontal axis shows the relative frequency of opportunities for reputation improvement ($\rho$) and the vertical axis shows, when reputations improve, the probability that the improve exogenously for free ($\kappa$), rather than endogenously for cost ($k$). For these figure I assume that $t = \frac{1}{4}, d = k = 1$, and all errors ($\epsilon, \phi, \eta, \xi$) are $\frac{1}{20}$. 


Invasion Analysis: Exploiter invades

The condition for Exploiter to invade Miser is

\[
a_5 = \sqrt{\rho \left( (\zeta + \kappa - \zeta \kappa)^2 \rho + 2\eta \kappa(-1+\rho)(-1+\phi) \right)}
\]

\[
a_6 = \rho(\zeta(1-\kappa)+\kappa)
\]

\[
\frac{t}{d} > 2(1-\eta)\rho \frac{\eta \kappa(1-\rho)(1-\phi)-(\zeta(1-\kappa)+\kappa)(a_5-a_6)}{(a_5-a_6)(a_5-(1-2\eta)a_6)}
\]

The condition for Exploiter to invade RepCoop is

\[
a_7 = \sqrt{(1-\epsilon(1-\zeta)(1-\kappa))^2 \rho^2 + 2\eta(1-\epsilon(1-\kappa))(1-\rho)\rho(1-\phi)}
\]

\[
a_8 = (1-\epsilon(1-\zeta)(1-\kappa))\rho
\]

\[
\frac{d-\lambda}{k} < \frac{\eta}{\rho^2 \eta + 2(\zeta+k-\zeta \kappa)\rho} - 2\rho \frac{1-\kappa(1-\zeta)}{(1-\phi)(1-\rho)} \left( \frac{\lambda}{(1-\phi)(1-\rho)} + \frac{\eta}{a_7-a_8} \right)
\]

It is worth noting that as \( \eta \) goes to zero, the expression \( \frac{\eta}{a_7-a_8} \) approaches \( \frac{1-\epsilon(1-\zeta)(1-\kappa)}{(1-\epsilon(1-\kappa))(1-\rho)(1-\phi)} \). This allows us to readily derive the key simplifications in the main text.

Though this precise expression is fairly complex to puzzle over symbolically, a visual presentation makes its implications clear. The top row of Figure 3.6 shows the regions where RepCoop and Miser can resist invasion for parameters \( \rho \), \( \kappa \) and \( \zeta \). Both strategies do better when \( \rho \) is low, since reputations are more valuable and costs paid to maintain them are better recouped.

When not-volunteering carries no negative consequence (\( \zeta = 0 \)), Miser is quite robust. However as \( \zeta \) increases, Miser struggles if they cannot rely on reputations improving extrinsically (\( \kappa \) low), while RepCoop can hold its ground.

Alternatively, we can unpack these expressions by considering special,
theoretically interesting cases of the parameters. Here we examine the sta-

tability conditions for RepCoop in particular, since it is the more theoretically

interesting of the two.

**Per-scenario breakdown of RepCoop stability**

**NIR-P**

When the simplest, earliest reputations first emerged among our ances-
tors, we can assume that individuals were not aware of their own reputations
and rarely took deliberate action to improve them. Instead, reputations im-
proved exogenously with some probability (tracked here by \( \rho \)). For instance,
they may have just gotten better over time as others forgot the past, or in-
dividuals may have accidentally performed actions that improved them. To
investigate this situation we evaluate our inequality when reputations only
improve at random (\( \kappa \to 1 \)) and community don’t care if you fail to volunteer
(\( \zeta \to 0 \)). After some simplification, we get

\[
\frac{t}{d} < 1 - \frac{2\rho}{1 + \rho - \phi + \rho\phi}
\]

Which, interestingly, is (given the same case of the parameters)

\[
\frac{t}{d} < R_{\text{RepCoop}} - R_{\text{Exploiter}}
\]

\[
t < d(R_{\text{RepCoop}} - R_{\text{Exploiter}})
\]

Put simply, for EXPLOITER to thrive the advantages of exploitation need
to exceed to additional damage incurred by having a bad reputation.

**NIR-V**

This is the key case in the emergence of cooperation by NIR. NIR-P
shows us that a society can emerge, and be stable in the face of mutation
and natural selection, in which individuals gain a fitness advantage from
maintaining a high reputation. We now have in hand a plausible selection
pressure favouring mutations which cause individuals to take costly actions
that improve their reputation. But just how costly can these actions be,
and do these deliberate efforts at reputation improvement change NIR’s dynamics?

To investigate this case we relax our constraints on \( \kappa \), assuming only that \( \zeta = 0 \). Here we show the simplified case where \( \eta \to 0 \). The supplemental Mathematica file contains complete expressions.

The condition for cooperators to resist invasion can again be simplified in terms of the reputational equilibria (given this case of the parameters).

\[
\begin{align*}
  t + k \frac{\rho}{1-\rho} \frac{2(1-\kappa)}{(1-\phi)} &< d \left( 1 - \frac{2\kappa \rho}{2\kappa \rho + (1-\rho)(1-\phi)} \right) \\
  t + k \frac{\rho}{1-\rho} \frac{2(1-\kappa)}{(1-\phi)} &< d (G_{\text{RepCoop}} - G_{\text{Exploiter}}) \\
  \frac{t}{d} &< (G_{\text{RepCoop}} - G_{\text{Exploiter}}) - k \frac{\rho}{1-\rho} \frac{2(1-\kappa)}{(1-\phi)}
\end{align*}
\]

Now we can see precisely the degree of extra burden that the cost of deliberate reputation improvement places upon cooperators. The damage spared them by their good reputation \((G_{\text{RepCoop}} - G_{\text{Exploiter}})\) must not only exceed the benefits of exploitation they’re forfeiting \((\frac{t}{d})\) but also the costs they pay to maintain it \((k \frac{\rho}{1-\rho} \frac{2(1-\kappa)}{(1-\phi)})\).

If REPCoop individuals prosper, NIR-V can sustain both non-exploitation and some degree of arbitrary costly reputation improvement, as long as damage is great relative to takings and opportunities for reputation improvement are relatively scarce compared to opportunities for exploitation.

If reputations sometimes improve costlessly \((\kappa)\), Exploiter have an advantage. In the limiting case where reputations cannot improve without deliberate effort, our inequality simplifies to:

\[
\frac{d-t}{k} > \frac{\rho}{1-\rho} \frac{2}{1-\phi}
\]

**NIR-M**

The earlier models give us some reason to think that natural selection could have favoured costly cognitive adaptations for a) attending to one’s reputation, b) attending to what actions would raise it, and c) being moti-
vated to take those actions (even if they’re costly). Eventually, this could lead to communities of individuals who very deliberately and carefully try to improve others opinions of them.

Since in such a community, individuals would be carefully attending to their own opportunities to improve their reputation by contributing to others’ welfare. This could equip them with all the cognitive prerequisites they’d need to begin noticing others’ opportunities for reputation improvement by acting in their interests, and be disappointed if they didn’t. Such disappointment could be the foundation for coordinated reputational punishment of those who don’t meet community expectations. Since such spontaneous, widespread disapproval would actually motivate other-fitness-enhancing reputation-reparation, it would be self reinforcing, eventually driving up the probability that failure volunteer worsens reputations ($\zeta \to 1$).

Could such a society still sustain cooperation? Just how costly a level of volunteering could it sustain?

$$t \left( \frac{1 - \varepsilon(1 - \kappa)}{1 - \varepsilon(1 - \zeta)(1 - \kappa)} \right) + 2k(1 - \kappa) \frac{\rho}{(1 - \rho)(1 - \phi)}$$

$$< d \left( \frac{1 - \varepsilon(1 - \kappa)}{1 - \varepsilon(1 - \zeta)(1 - \kappa)} - \frac{2\varepsilon \rho + 2\zeta \rho(1 - \kappa)}{2\varepsilon \rho + 2\zeta \rho(1 - \kappa) + \frac{(1 - \varepsilon(1 - \kappa))(1 - \rho)(1 - \phi)}{1 - \varepsilon(1 - \zeta)(1 - \kappa)}} \right)$$

$$\frac{i}{d} < \frac{\mathcal{G}_{\text{RepCoop}} - \mathcal{G}_{\text{Exploiter}}}{\mathcal{G}_{\text{RepCoop}}} - k \frac{1}{\mathcal{G}_{\text{RepCoop}}} \frac{\rho}{(1 - \rho)} \frac{2(1 - \kappa)}{1 - \phi}$$

A simple way to understand what this inequality is telling us is to compare it to NIR-V’s condition. Now the reputational advantage that cooperators achieve is greater. That is, the term $(\mathcal{G}_{\text{RepCoop}} - \mathcal{G}_{\text{Exploiter}})$ is now weighted inversely by $(\mathcal{G}_{\text{RepCoop}})$, which is smaller than one, making the whole term larger. Meanwhile the costs they pay are more onerous too, similarly weighted by $\frac{1}{\mathcal{G}_{\text{RepCoop}}}$.

Overall, except in cases when volunteering costs are very low ($k$ is small relative to $\frac{i}{d}$), RepCoop will be less stable under NIR-M than NIR-V. If community demands begin to escalate, individuals who simply ignore them eventually come to thrive.

As we’ve seen above, these additional costs buy RepCoop an advantage
against MISER, and as we’ll see below, they also result in a higher level of equilibrium volunteering.

**Numerical Maps**

By analysing the stability conditions for monomorphic populations we’ve gained some insight into when different strategies can thrive, and seen that mixed populations will sometimes also contain an unstable equilibrium. These unstable mixed-equilibria define the basins of attraction of the monomorphic boundary equilibria, offering insight into how likely each strategy is to emerge. We’ve also seen that stable mixed population of REP/COOP and MISER is also sometimes possible.

To offer some traction of the locations of these internal equilibria, we have mapped out their locations in Figure 3.7 for several different parameter regions. Note that unlike our assumption-testing simulations above, this is not a simulation. It is merely a different way of interpreting analytic expressions. In general their dynamics are identical to the insights gleaned from our invasion analyses. Cooperative strategies thrive when exploitation is inefficient ($\frac{t}{d}$ is small), bad reputations stick ($\rho$ is low), REP/COOP thrives when the cost of volunteering are low ($k$ is small) and individuals are skilled at navigating reputations ($\zeta$ high).

The attached Mathematica file present an interactive barycentric plot, by means of which you can readily explore the dynamics of this three-strategy system across the full scope of parameter values. This file also includes is the full expression for $G_{Mixed}$ in all its massive glory, and all the code required to re-derive the full model.
\( \rho \in \frac{1}{10}, \frac{2}{10}, \ldots, \frac{9}{10} \)

\( t \in \frac{1}{10}, \frac{2}{10}, \ldots, \frac{9}{10} \)

\( k \in \frac{1}{2}, 1 \ldots \frac{9}{2} \)
**Figure 3.7 (preceding page):** Meta-barycentric plots of NIR boundary equilibria. First-order barycentric plots depict the evolution of the relative frequency of discrete NIR strategies. Corners of the triangle represent monomorphic populations of RepCoop (left), Miser (top) and Exploiter (right) and spaces between them represent mixed population in the corresponding proportion. Along their exterior edges barycentric plots may contain ‘boundary equilibria’—relative frequencies at which the two strategies at the abutting corners have equal fitness. This meta-plot shows how the location of those boundary equilibria changes as three key parameters change (outer rows, $r$, $t$ and $k$). When the row-parameter is at its minimum, the locations of boundary equilibria are depicted with a star (stable) or filled circle (unstable). Green stars show edges where fitness is always equal. Increasingly lighter arrows (stable: dashed, unstable: solid) show where the equilibria move to as the corresponding parameter changes. If an equilibrium does not shift substantially at the first increment in the parameter sequence (e.g. as $\rho$ changes from $\frac{1}{10}$ to $\frac{2}{10}$), the first parameter value that makes a substantial difference is printed beside the corresponding arrow (e.g., the stable internal equilibrium moves dramatically as $\rho$ goes from $\frac{5}{10}$ to $\frac{4}{10}$ in the top-right plot). Since each edge contains a single internal equilibrium, the stability of monomorphic populations at the corresponding corners can easily be inferred. Corners abutting edges with stable equilibria are themselves unstable, and vice-versa. Unless otherwise indicated by column or row labels, parameters are set to: $k = d = 1; t = \frac{1}{4}; \rho = \kappa = \frac{1}{10}; \varepsilon = \phi = \eta = \frac{1}{20}$. 
3.1.7 Evolving continuous traits interpretation

There is a second way we could reason about the long term evolution of these dispositions. Rather than assuming that change happens by a dramatic mutation (e.g., a population of Exploiter \((S = 1, V = 0)\) in which a single cooperator \((s = 0, v = 1)\) arises, we can assume it happens by small mutations which spread to fixation before another mutation arises. Using this approach, we assume the members of the resident population all have approximately the same dispositions \((0 \leq S, V \leq 1)\). These dispositions evolve gradually over time by small genetic mutations arising that change the disposition slightly. Such mutations typically either spread to fixation or vanish (depending on whether they provide a fitness advantage or disadvantage) before another mutation arises.

To formally reason about the long-run consequences of such a process, we ask what would happen to a mutant whose dispositions varied from the population average by some small amount \((s = S + \delta_s, v = V + \delta_v)\). How small an amount? Small enough that a linear approximation of the difference between the fitness of the resident and that of the invader gives an accurate approximation of the evolution of the system.

In evolving systems where small dispositional changes yield fitness differences that are approximately linear,\(^1\) these approximations can yield accurate inferences even if large mutations sometimes occur. For systems whose dynamics are highly chaotic or just very non-linear, this approximation only holds in the limit of very small mutations. Numerical inspection of NIR suggest that it’s fitness difference are typically monotonic and approximately linear, giving us confidence in the insights of this adaptive dynamics analysis (for a discussion of other limitations of the assumptions underlying this ‘adaptive dynamics’ approach, see Waxman & Gavrilets, 2005).

\(^1\)for instance, if having a greater disposition to do something improved your fitness by, on average \(x\) fitness units, then having a disposition doubly as great would yield close to \(2x\) units improvement
Assumptions.

11. Everyone’s similar enough, and
12. Mutations are small enough, and
13. Tend to spread to fixation before new mutations arise.
14. Mutations that successfully invade will typically spread to fixation.

Using these assumptions, we can solve for the probability that an average resident has a good reputation at any point in time ($\hat{G}$), and also a rare invader in that same population ($\hat{g}$).

$$
\hat{G} = \mathcal{G}(\hat{G})|_{s \rightarrow S,v \rightarrow V}
$$

$$
\hat{g} = \mathcal{G}(\hat{G})
$$

Using this we can readily express the average fitness of residents and rare invaders in any given population, and the difference between them as:

$$
W_{\text{Invader}} = W(\hat{g}, \hat{G})
$$

$$
W_{\text{Resident}} = W(\hat{G}, \hat{G})|_{s \rightarrow S,v \rightarrow V}
$$

$$
W_{\text{diff}} = W_{\text{Invader}} - W_{\text{Resident}}
$$

With this difference in hand we can readily infer, for any given resident
population (i.e., \(S, V\)), whether a rare mutant would have an advantage by examining the ‘Selection Gradients’:

\[
\vec{SG}(V, S) = \begin{pmatrix}
\frac{dW_{diff}}{ds} \\
\frac{dW_{diff}}{dv}
\end{pmatrix}
\bigg|_{s=S,v=V}
\]

Notice that we evaluate these derivatives at the point where the invader’s traits are identical to the resident’s. Effectively we are taking a first order Taylor series approximation of the evolutionary dynamics at the vicinity of the resident’s traits. That is, we are linear approximating the effect of small mutations in \(S\) and \(V\) on fitness.

These selection gradients are a powerful tool and let us readily map out the dynamics of the system for any given parameter combination. In the supplemental Mathematica file we provide interactive plots of these selection landscapes, where they can be interactively observed as parameters change.

We can also glean useful insights by analysing the selection gradients analytically. We have found that to make this analytically tractable we need to a) considering the limiting cases of NIRV (\(\zeta = 0\)) and NIRM (\(\zeta = 1\)) in turn and b) consider a simplified system where errors do not occur (i.e., \(\phi = \eta = \epsilon = 0\)). We have numerically compared this simplified approximation to dynamics with reasonable errors (\(\phi \sim \eta \sim \epsilon \sim \frac{1}{20}\)) and found that they generally describe a qualitatively identical system, where equilibrium locations and null clines are only slightly perturbed by errors. There is one exception. Assuming away accidental defection (\(\eta = 0\)) causes a special degenerate case when residents are entirely undisposed to exploit (\(S = 0\)). In this case nothing ever worsens residents’ reputations and so there is no reason to ever volunteer (\(\hat{V} \to 0\)). However even very small rates of accidental defection can lead to substantial rates of equilibrium volunteering.

To tackle this we proceed by analysing internal dynamics in the special case of no errors and to show that the internal equilibrium is unstable and dynamics will eventually drive dispositions to their boundaries. We then
consider each boundary in turn and explore the \((S = 0)\) with making these assumptions.

**Unstable Internal Equilibria**

The first step is to find this system’s equilibria, dispositions that do not lead to evolutionary change \((\hat{V}, \hat{S})\), for NIRV and NIRM respectively, by solving:

\[
\vec{S}\hat{G}(\hat{V}, \hat{S}) = (0,0)^T
\]

\[
\vec{S}\hat{G}(\hat{V}_{NIRV}, \hat{S}_{NIRV}) = \vec{S}\hat{G}(\hat{V}, \hat{S})|_{\zeta=0}
\]

\[
= \begin{pmatrix}
-\frac{d(2\kappa k^2(r)-(1-p)r^2)+2\kappa p(t+r)^2}{2(1-\kappa)p(k+t)(dk+t(k+t))} \\
\frac{dk}{dk+t(k+t)}
\end{pmatrix}
\]

\[
\vec{S}\hat{G}(\hat{V}_{NIRM}, \hat{S}_{NIRM}) = \vec{S}\hat{G}(\hat{V}, \hat{S})|_{\zeta=1}
\]

\[
= \begin{pmatrix}
-\frac{k}{1-\kappa} \\
\frac{(1-p)-(2\kappa p)}{d(1-p)}
\end{pmatrix}
\]

\[
\text{or}
\begin{pmatrix}
\frac{d(1-p)^2-2\kappa p(k+k+1-d)}{2(1-\kappa)k^2p(d+t)} \\
\frac{d(1-p)-2\kappa p}{(1-p)(d+t)}
\end{pmatrix}
\]

While NIRV contains a single equilibrium, NIRM contains two. However, notice that one equilibrium lies strictly outside the bounds of the system \((at - \frac{k}{1-\kappa})\), and so regardless of its stability, dynamics around this equilibrium will push the evolving variable to their boundaries rather than some internal point.

To make this inference about the other equilibria we can assess their ‘convergent stability’. That is, we ask whether selection will drive dispositions towards or away from these points. We do this by assessing the Jacobian matrix of these gradients. That is, if \(\vec{S}\hat{G}_V\) and \(\vec{S}\hat{G}_S\) are the components of the selection gradients that correspond to selection in the \(V\) and \(S\) dimensions.
respectively:

\[ J = \begin{pmatrix} \frac{\partial S_{GV}}{\partial V} & \frac{dS_{GV}}{dS} \\ \frac{\partial S_{GS}}{\partial V} & \frac{dS_{GS}}{dS} \end{pmatrix} \]

By evaluating the Jacobian at an equilibrium we express the rate of change in the fitness difference as resident populations approach this point in two-dimensional space.

The Jacobians evaluated at the NIRV and second NIRM equilibria are:

\[
J \bigg| \xi = 0, V = \bar{V}_{NIRV}, S = \bar{S}_{NIRV} = \begin{pmatrix} \frac{-t(dk+t(k+1))(2t(k+1)+d(2k+t))(p-1)}{2d(2k+t)^2} & \frac{(dk(2k+t)+t(k^2-t^2))(k-1)p}{d(k+1)(2k+t)} \\ \frac{-t(dk+t(k+1))(k-1)p}{d(2k+t)} & \frac{2k(k+t)(k+2t)(d(k+t+k))(k-1)^2p^2}{d^2(2k+t)(p-1)} \end{pmatrix}
\]

\[
J \bigg| \xi = 1, V = \bar{V}_{NIRM}, S = \bar{S}_{NIRM} = \begin{pmatrix} \frac{t^2(d+t)^2(p-1)^2(-dt+t+k)}{2k(-pt+t+2k)p)(d(-pt+t+k)-k)p) & \frac{-k(d+t)(k-1)(p-1)p(2d(-pt+t+k)-r^2(p-1))}{2(r(p-1)+2k)(k+p+d(t(p-1)-k))} \\ \frac{-k^2(d+t)^2(k-1)^2p^2}{d(-pt+t+2k)p)(d(-pt+t+k)-k)p) & \frac{k^3(d+t)(k-1)^2p^2((p-1)d+2k)}{d(-pt+t+2k)p)(d(-pt+t+k)-k)p) \end{pmatrix}
\]

For NIR’s equilibria to be convergently stable all the eigenvalues of this matrix must be negative (i.e., dynamics must drive the system back towards this point in all dimensions). This will only be true if the determinant of the Jacobian (i.e., \(J_{11}J_{22} - J_{12}J_{21}\)) is positive (see Routh-Hurwitz Conditions in Otto & Day, 2011, chapter 8). However, for both NIRM and NIRV, a little algebra shows that this is never true. That is, NIR’s internal equilibrium is never stable. Evolutionary dynamics will always drive populations away from this point and towards the boundaries (i.e., where \(S = 1, S = 0, V = 1\).
or $V = 0$) of this system. Our careful numerical investigations of many parameter combinations suggest that this is also always true for intermediate values of $\zeta$ and reasonable error rates.

To get a more intuitive sense of these dynamics, we provide interactive phase-space plots in the Mathematica supplement and static versions for theoretically interesting parameter combinations in Figure 3.8.

**Boundary Dynamics: The S-null-cline is always unstable.**

First a little algebra shows that, just as under a discrete strategy interpretation, a world of complete defection must always be stable. Specifically:

$$\tilde{SG}_S > 0$$

If no-one respects reputations, selection will never favour a shift away from this situation. Next we find the location of the system’s null-cline in the S-dimension. That is, the curve at which $\tilde{SG}_S = 0$. This expression tells us, for a given $V$, the value of $S$ at which selection gradients are flat. See the Mathematica supplement for the full expression and Figure 3.8 below for a visualisation of this curve.

We find a single S-dimension null-cline (see Mathematica supplement). Given this, and the fact that selection must push upwards at the $S = 1$ edge, we know that when this null-cline is within the bounds of the system (i.e., $0 < S < 1$) it must be always unstable, or dynamics would not continue to push the system against its upper $S$-bound. Equivalently, we know it must be stable when it is located at $S > 1$. 
NIR-P
\((\kappa = 1)\)

NIR-V
\((\zeta = 0)\)

NIR-M
\((\zeta = 1)\)

\[\rho = \frac{1}{10}\]

\[\rho = \frac{1}{2}\]
Figure 3.8 (preceding page): Selection gradients for three version of NIR (columns) in two different parameter regions (rows), depicting how selection would shape continuously evolving dispositions to exploit well-reputed peers ($S$, vertical dimension) and volunteer to improve your reputation ($V$, horizontal dimension). Orange lines depict the approximate (when errors are zero) null-cline in the $S$ dimension; dispositions to exploit above this line tend to get greater, dispositions below (in the shaded region) get smaller. The red dots indicate the precise location of stable $V$ equilibria along the $S = 0$ and $S = 1$ edges; dispositions to Volunteer tend to approach these points. Parameters are set to: $k = d = 1; t = \frac{1}{4}; \kappa = \frac{1}{10}; \varepsilon = \phi = \eta = \frac{1}{20}$. Here one can see that though NIR-M sustains higher rates of volunteering in cooperation-favouring parameter regions (e.g., top row), NIR-V continues to suppress exploitation even when costly-reputation improvement does not pay (e.g., bottom row).
Boundary Dynamics: The evolution of $V$.

Since we know the system has a single unstable null-cline running through it’s $S$-dimension, we know dynamics will eventually push it to $S = 0$ or $S = 1$. What will happen to $V$ on these edges?

When $S = 1$, the selection gradient for volunteering ($\vec{SG}_V$) is strictly negative, specifically it is $(1 - \kappa)(-k)\rho$. This makes sense. If reputation does nothing to stop others exploiting you, why would you pay to maintain it? Consequently when exploitation is rife, volunteering will be eliminated.

When $S = 0$, we can find the location of the $V$ equilibrium (call it $\hat{V}_{S=0}$) by solving for when the selection gradients are zero. While we do find precise analytic solutions for these points, they are again long and unwieldy expressions that offer little insight on their own. Instead, here we present the limiting cases of NIRV and NIRM, and below we show the consequences of this equilibrium on actual volunteering rates (i.e., weighted by $\rho$ and $1 - \kappa$).

Equilibrium selection

We’ve seen three different ways that early reputation-using societies–those in which individuals’ opinions of one another were somehow coordinated–may have structured their interactions: NIR-P, NIR-V and NIR-M. It is natural to ask, if these societies coexisted, what would be their consequence of their interactions? Which would thrive and which would be outcompeted? We have two pieces of information that could help us answer this question.

The first is the reputational equilibrium (i.e., stationary distribution) at the cooperator behavioural equilibrium. Those societies that have the least internal exploitation may have a competitive advantage. For societies at their cooperative equilibria ‘least exploitation’ means ‘fewest individuals with bad reputations’. Here, NIR-M is transparently outcompeted by NIR-V and NIR-P, since it includes an additional mechanism for worsening reputations. The relative standing of the two deliberate-improvement models to NIR-P depends on just how many more costly opportunities there are for reputation improvement than costless ones (i.e., $\kappa$), which is an empirical question.
There is another feature of NIR that can inform us about long run between-society interactions. It is not implausible that (at least some of) the reputation-boosting acts we modelled are actually instances of positive cooperation. That is, individuals bearing costs to improve the fitness of their peers.

What is the relative rate of volunteering (e.g., gleaning favour by prosocial giving) between NIR-M and NIR-V? Here we can use the equilibrium volunteering rates at the non-exploitation boundary $\hat{V}_{S=0}$ to gain some insight. This is presented in Figure 3.9. As opportunities for prosociality increase, the incentive to give declines in NIR-V as individuals less often find themselves with bad reputations, but continues to increase in NIR-M. However as volunteering costs (weighted by opportunities) become too onerous, NIR-M can suddenly collapse to it’s full-defection equilibrium, while NIR-V does not.

In general, this suggests that as long as exploitation is common and inefficient, NIR-M systems should sustain higher rates of volunteering than NIR-V systems, and so be better candidates for victory in between-group competition.

3.2 Additional supplemental materials

Additional supplemental materials are provided as computer files. They include:

- R-code for running reputation simulations and numerically approximating equilibrium locations in the discrete-strategy interpretation.

- A Mathematica file (proprietary symbolic algebra software), by means of which the main results can be readily rederived.
Figure 3.9: Relative volunteering rates for NIR-V and NIR-M. Relative rates of costly reputation-raising contributions to others’ welfare maintained by NIR-V (red lines) and NIR-M (blue lines) at their respective cooperative equilibria. Lines extend as long as the cooperative equilibrium is stable (i.e., $\vec{SG}_{S}|_{S=0} < 0$), when $\frac{k}{\bar{d}}$ is 2 (unbroken lines), 1 (dashed lines) and $\frac{1}{2}$ (dotted lines). Here $d = 1, t = \frac{1}{4}, \kappa = \frac{1}{10}$ and all errors $(\varepsilon, \phi, \eta)$ are $\frac{1}{20}$. Notice that if costs become too onerous, cooperation under NIR-M collapses while NIR-V continues to sustain cooperate as volunteering gradually declines.
Chapter 4

Surveying indirect reciprocity: how do people assign reputations?

Humans can be impressive cooperators. Our metropolises are crowded with anonymous passers-by whom we could swindle, waylay, defraud, pickpocket, intimidate or otherwise exploit. Yet, we don’t. Instead we hold the bus door for them, offer them directions, give them alms, and sometimes even risk our lives to save theirs. Of course, it hasn’t always been so, and some contemporary societies are still fraught with bandits, thieves, charlatans, rapists and corrupt officials. Recently considerable scholarly attention has been applied to explaining how human cooperation arose, targeting questions like: how is cooperation sustained, why do our patterns of cooperation differ from other species, and why do they vary so dramatically between societies? Why do some groups thrive in mutual aid and cooperation, while others collapse under the weight of distrust and mutual exploitation?

Any complete answer to these questions will likely feature contemporary cultural institutions. For instance: police forces, laws, courts, debt, taxation, intellectual property, democratic elections, labour unions and parking fines. However such institutions cannot be the root of the explanation, since they themselves depend on many cooperative interactions. Police officers,
judges and vote-counters usually stand to gain far more than the average person from corruption and deceit. If sufficiently many people ignored them, laws, debts, taxes and parking fines would be unenforceable. Our institutions seem to hold one another aloft. Together they free us from worrying about whether each stranger, shop keeper or new friend will betray us. But if each institutions depends on some or many individuals already being cooperatively inclined, how did whole flotilla become buoyant in the first place, and why have other animals not followed similar trajectories?

Reputations, and reputation-based cooperation, play a central explanatory role the three main types of explanations that interested thinkers have posited.

The first kind places causal primacy on our extraordinary intelligence (e.g., Pinker, 2010). First we became smart, then we figured out how to cooperate, conceived of and designed clever institutions, and so on.¹ However clever ancestral humans may have been, these individual-intelligence-based explanations must be coupled with population-level accounts of how early human groups first escaped uncooperative equilibria (situations in which every individual benefits most by being uncooperative). Intelligence alone cannot sidestep these population-level questions; after all, cleverer-still contemporary humans also struggle with such dilemmas. One simple gambit is to claim that language allowed our ancestors to talk and, especially, to gossip. Language and intelligence allowed them to coordinate community-wide response to uncooperative individuals, making cooperation an individually fitness-enhancing, smart choice. Since formal models of this process branched off from models of ‘direct’ reciprocation between pairs of individuals (e.g., Axelrod & Hamilton, 1981; Boyd & Lorberbaum, 1987; Doebeli & Hauert, 2005; Trivers, 1971; van Veelen et al., 2012, more generally, these are models of viable strategies for repeated interaction between pairs of in-

¹These intelligence-based arguments, though they are perhaps the most common folk-explanation, face many unsolved theoretical challenges. For instance, how could selection pressures produce extraordinary intelligence, bearing the concomitant costs of big brains and smaller guts (Aiello & Wheeler, 1995; Kotrschal et al., 2013), unless people were already cooperative enough that they freely shared the cultural knowledge that makes big brains worth having?
individuals), this population-level reputation-coordination process is usually called Indirect Reciprocity (IR).

A second category of explanations of human cooperation give casual primacy to cooperation-sustaining mechanisms continuous with other species, especially kin-selection and direct reciprocity. These are sometimes called ‘misfire accounts’, because they posit that we, effectively, mistake strangers for kin or long-term interaction partners. Our psychology, these accounts posit, is calibrated to an ancestral past where most our interaction partners were kin or stuck around a long time. They have yet recalibrate to the anonymous cities and transcontinental trade networks of the Holocene. These misfiring cooperative dispositions provided the foundation for more sophisticated cooperative institutions including, possibly, the deception-free sharing of valuable knowledge that allowed our ancestors (and us) to become so smart in the first place. However, kin-selection struggles to sustain cooperation as phylogenetic distance increases, and direct reciprocity is only effective in dyads or small groups (Chudek et al., 2013b). These explanations must also be paired with an account of how, even in a cooperatively misfiring species, large-scale cooperative institutions first arose. Again, reputation provides a parsimonious bridge from individual psychological dispositions to self-sustaining population-level dynamics.

A third kind of explanation shifts the focus from just the evolution of our psychology to the evolution of our institutions themselves. Such culture-gene coevolutionary accounts start by explaining the genetic adaptations that may have led humans to transmit cultural information in the first place; and to do so well enough that culture began accumulating and evolving across generations into technologies and institutions that no human could devise on their own. These accounts then model how our evolved psychology biases the evolution of our cultural knowledge, and how that in turn biases the genetic pressures on our psychology. They derive predictions about contemporary psychology, behaviour and institutions based on the long-term consequences of these two interacting evolutionary processes. Since reputations—which

---

2 Misfire arguments, though common among evolutionary psychologists, also face their share of theoretical and empirical challenges (Chudek et al., 2013b; Fehr & Henrich, 2003)
socially convey information about one’s peers—are themselves a kind of cultural information, the question of how reputations can sustain cooperative cultural knowledge sharing is particularly relevant to culture-gene coevolutionary theorists.

While the first two kinds of explanations merely assume that reputational information is freely and accurately transmitted among individuals, culture-gene coevolutionary theorists have a long history of confronting the cooperative dilemmas inherent in cultural transmission. They have proposed several mechanisms that could sustain generalised cooperation, including the quality of reputational information, among a cultural species; including: conformity (Henrich & Boyd, 2001), deference (Henrich & Gil-White, 2001), and credibility enhancing displays (Henrich, 2009).

However you prefer to explain the foundations of human cooperation, reputation is likely to play a central role. Next I review the main theoretical insights about the population processes by which reputation brings about cooperation (Indirect Reciprocity, IR). I then turn to testing these predictions against contemporary people’s actual intuitions. While I synthesise IR theory from a culture-gene coevolutionary perspective, the empirical tests are relevant to all perspectives on human cooperation.

**Reputation, in theory**

At first glance, reputations seem a simple route to cooperation. If you do bad things to others, your community will eventually find out and start doing bad things to you. In the long run, the reputational costs of exploitative or uncooperative acts outweigh the immediate payoffs and good guys prosper. But formal explorations of reputations’ population-level dynamics quickly become quite complicated. To help unfamiliar readers keep a handle on this complexity, I provide a glossary of important terms in Table 4.1.

First, there are many different ways that reputations could influence behaviour—different action-rules. That is, there are many different behavioural strategies people could employ, given their own reputation and that of the potential target of their actions. Imagine that your commu-
<table>
<thead>
<tr>
<th><strong>Cooperative Dilemma</strong></th>
<th>A situation where an Actor could chose between a behaviour that advantages them, and one that advantages one or more others more.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Prisoners’ dilemma</strong></td>
<td>A cooperative dilemmas with only two individuals involved.</td>
</tr>
<tr>
<td><strong>Positive Dilemma</strong></td>
<td>A cooperative dilemma where the active Action yields a cost for the actor and larger benefit for the Target(s), relative to doing nothing.</td>
</tr>
<tr>
<td><strong>Negative Dilemma</strong></td>
<td>A cooperative dilemma where the active Action yields a benefit for the actor and larger cost for the Target(s), relative to doing nothing.</td>
</tr>
<tr>
<td><strong>Actor</strong></td>
<td>The individual who makes a choice in the focal cooperative dilemma.</td>
</tr>
<tr>
<td><strong>Target</strong></td>
<td>The individual affected by the Actor’s action.</td>
</tr>
<tr>
<td><strong>Action</strong></td>
<td>What the actor actually does, cooperating or defecting.</td>
</tr>
<tr>
<td><strong>Cooperating</strong></td>
<td>Choosing the other-advantaging choice in a cooperative dilemma.</td>
</tr>
<tr>
<td><strong>Defecting</strong></td>
<td>Choosing the self-advantaging choice.</td>
</tr>
<tr>
<td><strong>Helping</strong></td>
<td>Cooperating in a positive dilemma (cf. <em>passively defecting</em>).</td>
</tr>
<tr>
<td><strong>Exploiting</strong></td>
<td>Defecting in a negative dilemma (cf. <em>passively cooperating</em>).</td>
</tr>
<tr>
<td><strong>Reputation</strong></td>
<td>A cognitive representation of an individual which influences how one treats them in a cooperative dilemma.</td>
</tr>
<tr>
<td><strong>Judge</strong></td>
<td>An individual assessing a cooperative dilemma, and revising their representation of the Actor’s reputation.</td>
</tr>
<tr>
<td><strong>Action-rule</strong></td>
<td>A strategy for how to Act in cooperative dilemmas, given the reputation’s involved.</td>
</tr>
<tr>
<td><strong>Assessment-rule</strong></td>
<td>A consistent system for judging the Actor in a cooperative dilemma, given their action (1st-order judgement), their Target’s reputation (2nd-order judgement) and their own previous reputation (3rd-order judgement).</td>
</tr>
</tbody>
</table>

*Table 4.1: Glossary*
nity is particularly gossip prone and reputations really matter. Should you always cooperate with others? Would you be better off only cooperating when your reputation is bad? Or perhaps only when the potential target of your act is good? The consequences of these choices depend on how your community judges these acts—their *assessment-rules*.

Imagine that one of your well-reputed peers had helped a poorly-reputed target (e.g., loaned them some money). Would you (and your community) lessen your opinion of them for consorting with scum, or glory them as a Good Samaritan? Say they had exploited a bad person (by stealing from a thief, for instance), would you think them as virtuous as Robin Hood or as a villain for their crime? It is not obvious *a priori* how individuals and communities should reputationally judge one-another’s behaviour, let alone how a socio-ecology populated by many different action-rules and assessment-rules would unfold.

Early formal IR theory considered a simple ‘first order’ assessment-rule called ‘Image Scoring’ (Boyd & Richerson, 1989; Nowak & Sigmund, 1998). Under *Scoring*, any cooperative act improves reputations and any defection (i.e., uncooperative act) worsens them, regardless the *Actor’s* (i.e., individual acting) and *Target’s* (i.e., individual acted upon) prior reputations. *Scoring’s* action-rule is to only cooperate with Targets whose reputation is sufficiently good.

However further investigations revealed that *Scoring* is not evolutionarily stable (i.e., it will be out-competed by other strategies) if people sometimes make mistakes (Leimar & Hammerstein, 2001; Panchanathan et al., 2003). That is, if sometimes individuals intend to cooperate but don’t manage to (‘I really did mean to return your shovel, and I really did misplace it’), a *Scoring*-based community would exploit one-another often enough that unremitting defectors (individuals who never cooperate with anyone) could prosper in their midst. This happens because *Scoring*-communities reputationally-punish anyone who defects even on unequivocal scoundrels. This fact inspired theoretical work on more sophisticated ‘Standing’ strategies. More recent work has described more sophisticated variants of *Scoring* (Berger, 2011; Rankin & Eggimann, 2009), or specific kinds of interactions.
Standing assessment-rules are 2nd-order, they are sensitive to Targets’ reputations. The earliest Standing models (e.g., Panchanathan & Boyd, 2004) assumed that cooperative acts always improve reputations, while only defecting on good people worsens them; defecting on bad people leaves reputations unchanged. Standing’s action-rule is just like Scoring’s. However, unlike Scoring, even if people occasionally defect when they intended to cooperate, Standing-individuals in a Standing-community will do better in the long run than unremitting defectors.

Ohtsuki and Iwasa (Ohtsuki & Iwasa, 2004) surveyed all 4064 possible combinations of action-rules and assessment-rules, when reputations are binary (individuals can be good or bad, nothing in between) and individuals can either cooperate or defect, no other reputation-relevant events occur. They found that eight of those stood out as particularly evolutionarily viable (they could not be invaded by other strategies, and maintained high levels of cooperation) and called them the ‘leading eight’. Like Standing, the leading eight like cooperation with good Targets (it improves reputations) and dislike defection on good Targets (it worsens reputations). What tells the leading eight apart is how they judge actions (cooperation or defection) on bad Targets.

The mathematical techniques available to IR-theoreticians offer clear insights into long-term consequences when entire groups use just one assessment-rule, even if there are also unremitting cooperators or defectors present, or if sometimes individuals make genuine mistakes. However the complexities that result from many different assessment-rules interacting with a single community quickly become analytically intractable (cf. Uchida & Sigmund, 2010, for a single exception).

Two other important developments in IR-theory bear mention. Once reputations exist, they can be linked to any arbitrary behaviour (not just

---

3 These strategies have been dubbed by an ad hoc assortment of, sometimes inconsistent, names. For brevity and clarity I will refer to them all as Standing-strategies, since they are all (at least) second-order IR reputational systems.
cooperative dilemmas) to prop up its evolutionary stability (Panchanathan & Boyd, 2004). Imagine for example a reputation-based community at a primary school, where individuals with particularly bad reputations (losers) are picked on and exploited by their peers, while individuals with particularly good reputations (cool kids) receive disproportionately preferential treatment. Being mean to cool kids worsens one’s reputation, while being mean to losers does not. These ingredients are enough, in theory, for the reputational system to persist at the school, even as generations of children who instantiate it come and go. Now imagine that one generation decides, by whatever dynamic processes such decisions emerge, that anyone who likes the television programme ‘Sesame Street’ is a loser. Given its reputational consequences, this distaste for Sesame Street could persist across many generations of children at the school, even though it is neither a cooperative dilemma itself, nor is it directly connected to reputationally influenced cooperative and exploitative acts in this community. Analogously, reputations could support the persistence in historical time of maladaptive cultural traits that are otherwise difficult to explain (such as footbinding or female infibulation). More importantly, while IR-itself only stabilises cooperation among dyads, this potential for linking reputations to arbitrary behaviour provides an avenue for IR to also prop up larger scale cooperative ventures, such as selfless sacrifice in war (e.g., Mathew & Boyd, 2011), or public maintenance of more sophisticate cooperation-sustaining institutions (e.g., Sigmund et al., 2010).

Second, recent work has drawn attention to an important difference between positive cooperation (cooperating by helping others, or defecting by doing nothing) and negative cooperation (defecting by exploiting others, or cooperating by doing nothing). For brevity, I call this the *valence* of cooperation. While early work implicitly treated these situations as identical (or, at least, does not explicitly draw a distinction between them), Chapter 3 highlights important differences. While it is fair to assume that explicit action (i.e., actively cooperating by giving a gift or exploiting by stealing) could carry reputational consequences, the same does not hold for passive inaction (e.g., cooperating by not stealing someone’s unguarded valuables,
or passively defecting by not warning them about an encroach danger you foresee). Such inaction is likely to go unnoticed (and so unpunished or unrewarded) by anyone other than the potential actor, unless a community has well coordinated expectations about what opportunities exist, when, and what the normative responses are to them. Since such coordination of social norms to non-events is unlikely early in the coemergence of cooperation and culture, Chudek & Henrich conclude that negative cooperation in particular may have played a key, early role predict that our psychology may be particularly attuned to noticing and responding to it.

**Reputation, empirically**

Most empirical tests of IR have been done in the context of experimental games. Participants typically compete in computer-mediated repeated, positively-valenced prisoners’ dilemmas (i.e., they can pay a cost to create a larger benefit for someone) with anonymous strangers for small amounts of money. In these settings, investigators give participants access to different kinds of potentially reputation-relevant information (such as their target’s history of cooperating or defecting with others) and measure what effect it has on their choices.

There are several reasons that these only studies provide limited insight into whether and how IR structures interactions in contemporary populations, let alone how it may have influenced our ancestors whether it left a signature on contemporary cognition.

**Incomplete Assessment** The majority of studies (e.g., Charness et al., 2011; Engelmann & Fischbacher, 2009; Seinen & Schram, 2006; Simpson & Willer, 2008; Sommerfeld et al., 2008, 2007; Wedekind, 2000; Wedekind & Braithwaite, 2002) only provide participants with 1st-order reputational information (the actions of the Actor), and so cannot infer whether participants use 2nd- or higher-order assessment-rules. A few studies have provided participants information on Targets’ reputations, usually as a list of cooperations or defections by each Target, sometimes at a cost (Ule et al., 2009). Though early work
found no evidence of 2nd-order *Standing*-like assessment, subsequent work (Bolton et al., 2005) found the opposite—that *Standing* assessment is common enough that it can substantially skew Actors’ outcomes. While these 2nd-order studies are valuable, they still typically only investigate one or two of the many possible 2nd-order assessment rules described by theory.

**Mono-cultural** These studies are mono-cultural and typically use the globally peculiar sample of Western university undergraduates (Henrich et al., 2010). These leaves them poorly positioned to draw inferences about how universal reputation-assessing intuitions are, and by extension, how IR may have influenced early human evolution.

**Exclusively positive dilemmas** Perhaps due to tradition, these studies frame games exclusively as positive dilemmas—participants can help each other make more money. Where negative interactions are considered (e.g., Gächter & Herrmann, 2009; Milinski & Rockenbach, 2012; Ule et al., 2009), they appear as costly punishment (where Actors pay a cost to disadvantage Targets) rather than as negative cooperative dilemmas. Existing studies are not in a position to judge whether dilemma valence matters for IR.

**Competitive, artificial context** People’s reputational intuitions may be culturally transmitted (e.g., by imitation of parents’ or peers’ actions and assessments). If so, they would take the form of context-specific behavioural rules and response, and may differ between, say, the workplace, the bar and the family home. Measurements such context-specific variation are valuable for evaluating cultural-evolution interpretations of IR theory. Existing studies put participants in an unusual, highly artificial context where they know they are expected to compete with strangers for small sums of money while observed by scientists. Such contexts could evoke their distinct behavioural strategies and normative responses, concealing patterns of IR that governs participants’ lives outside the lab.
**Anonymity** A substantial drawback of experimental games is a consequence of one of their greatest strengths. It is standard practice in experimental games to ensure participants’ anonymity. This helps assure investigators that experimental variables are influencing participants’ responses, rather than concerns about what others might think of them after the study. However IR theory presupposes a socioecology entirely unlike this. IR emerges among communities of individuals who know each other well, interact for large portions of their lives, are interested in accumulating gossip and detailed information about one-another’s behaviours in many domains, and have detailed cognitive representations of each other that inform their behavioural choices. Consequently, representations of the specific target and actor may be needed to evoke participants’ IR-relevant cognitions.

Writers, story-tellers and movie-makers are masters of evoking emotions, judgements and sympathies in their audiences. To create sentiments of righteous approval when a hero vanquishes a villain, or abandons them to their demise instead of helping, they first create a relationship between the audience and the villain (e.g., by showing the villain delightedly spraying pepper in the eyes of yelping puppies). The audience’s subsequent approval when the hero’s actions may depend implicitly on their own vitriolic cognitive representations of the villainous Target. Since IR is inherently a system of long-term interactions with well known individuals, merely reading a list of anonymous others’ defections, even a second order list, may be insufficient to evoke participants’ IR reputation assessment-rules.

One way to side-step these concerns is to measure reputation change in naturalistic settings, or better, natural experiments. However this approach can only evaluate the subset of assessments that correspond to what is happening ‘in the wild’. One straightforward solution is to systematically describe the full set of situations relevant to IR, and simply ask people how they would respond to them. The effectiveness of this approach depends on whether people can, and are willing to, accurately simulate and report their reputational assessments of a verbally described situation. Even if these
self-reports are imperfect, they provide a valuable complement to existing experimental-game-based tests of IR.

Of course, ideally we would like to strengthen our inferences by finding ways to conduct experimental games that overcome the limitations above. However before attempting this costly and difficult feat, it behoves us (and is our ethical due to the tax-payers who fund our investigations) to initially survey this landscape using the most cheap and direct method available to us: simply asking people. That is what I have undertaken here.

**Present study**

By asking people for their reputational judgements in precisely those situations described by IR-theory, I aim to address these questions:

**Scoring or Standing?** Are people’s reputational judgments informed by 2nd-order information (e.g., defecting on bad guys is ok), or exclusively 1st-order (e.g., defecting is always bad)?

**Distribution of reputational systems** How are people’s reputational judgments distributed across the space of all possible 1st-order and 2nd-order reputational systems?

**Does the valence matter?** Do reputational judgments differ between positive cooperative dilemmas (where someone can help others or do nothing) than they do in negative ones (where they can exploit others or do nothing)? How, specifically? Is there covariance with the valence of people’s responses (e.g., do people prefer negative responses (stealing) to defection in negative dilemmas)?

**Are there cross-cultural differences?** Do people from phylogenetically distinct cultures use different reputational systems?

**What are the covariates of reputational judgments?** Do people’s reputational judgments systematically covary with the source of their information (gossip or direct observation), that actual consequences for the target independent of the intended act (a good outcome, a bad
one, or ) and participants’ individual differences (including age, sex, religion, education and time taken to make decisions)?

**Genes, culture or individual learning?** Which of these three mechanisms were more likely to have produced the patterns of variability we observe in people’s IR strategies?

### 4.1 Methods

To infer each participant’s reputational intuitions, I solicited their judgments about two scenarios (available in the supplemental materials). In each scenario I asked participants to imagine arriving in a new situation and meeting four new people, either a new workplace or a new group of friends. First I introduced them to a ‘Target’ character, who was cued either ‘good’ or ‘bad’. I did this by describing their behaviour (cooperating or defecting) in three dilemmas: an n-person public goods dilemma, a pairwise asynchronous prisoner’s dilemma, and an ‘information dilemma’, where the Target possess information that they could benefit by keeping secret or they could share to the benefit of others (see Table 4.2).
<table>
<thead>
<tr>
<th>Setting</th>
<th>Type</th>
<th>Good Target</th>
<th>Bad Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workplace</td>
<td>Public Goods</td>
<td>She is often the one who cleans the small, shared kitchenette at work, even though she’s rarely, if ever, the one who creates a mess.</td>
<td>She often makes a mess in the small, shared kitchenette at work, but you’ve rarely, if ever, seen her cleaning it.</td>
</tr>
<tr>
<td></td>
<td>Dilemma</td>
<td>When she works on a project with another workmate, she always makes sure to share the credit if it goes well and willingly shares the blame if it goes badly.</td>
<td>When she works on a project with another workmate, she tries to take all the credit if it goes well, even though she is rarely willing to share in the blame if it goes badly.</td>
</tr>
<tr>
<td></td>
<td>Information</td>
<td>Whenever she meets a valuable client she puts them in contact with other relevant team members, even though she could get an advantage by being their exclusive point of contact.</td>
<td>Whenever she meets a valuable client, she tries to prevent them contacting other relevant team members, so she can gain an advantage by being their exclusive point of contact.</td>
</tr>
</tbody>
</table>

| Friends | Public Goods                | A few weeks ago she offered to buy tickets to an event for several of your new friends. Later you found out that the tickets had actually cost more than she had asked your friends to give her. She had lost money so that the tickets would be cheaper for her friends. | A few weeks ago she offered to buy tickets to an event for several of your new friends. Later you found out that the tickets had actually cost less than she had asked your friends to give her. She had made money by having her friends pay more for the tickets than they originally cost. |
|         | Dilemma                      | Whenever someone is moving, she is almost always there to help them pack and move their furniture to their new house, even though she has never moved herself. | She has moved several times and had other friends come and help her pack and move her furniture to her new house, even though she has always been too busy to help anyone else when they were moving. |
|         | Pairwise                     | She loves shopping. Whenever she hears about a great sale, she always tells all her friends about it too, even though doing so makes it more likely that the sale items will run out before she finds what she wants. | She loves shopping. When she hears about a great sale, she never tells all her friends about it until she has had a chance to buy everything she wants, even though doing so makes it more likely that the sale items will run out before her friends find what they want. |
|         | Prisoner’s                  |                                                                                                        |                                                                                                        |
|         | Dilemma                      |                                                                                                        |                                                                                                        |
|         | Information                  |                                                                                                        |                                                                                                        |
|         | Dilemma                      |                                                                                                        |                                                                                                        |

**Table 4.2:** The cues used to establish the Target’s reputation, classified by type and scenario. Here I refer to the Target, whose name which varied between conditions, as ‘she’.
Next, participants were introduced to three Actor characters. Each Actor faced an identical cooperative dilemma, either positively valenced (opportunity to help the Target) or negatively (opportunity to exploit the Target). One Actor (the Cooperator) helped (positive valence) or didn’t exploit (negative valence) the Target. Another (the Defector) did the reverse. The order in which participants read about the Cooperator and Defector was randomised for each participant, but always reversed in the second scenario they read. A third individual (the Neutral) always faced the dilemma after the Cooperator and Defector, but was interrupted by the Target before they could act.

Participants then were asked to make four judgements about each of these four characters (our dependent measures). Each judgement was made by moving a slider (initially at its mid-point) or along a horizontal line between two boundary labels. Slider position was encoded as an integer between zero and one-hundred, with higher numbers representing more favourable (‘better’) judgements, which was then rescaled to the unit interval to facilitate beta-regression (Cribari-Neto & Zeileis, 2010; Ferrari & Cribari-Neto, 2004; Smithson & Verkuilen, 2006).

Our dependent measures were:

- **Good** How good a person they believe each individual is. Boundary labels: ‘Good’ and ‘Bad’.
- **Like** How well they think they would like each individual. Boundary labels: ‘Like’ and ‘Dislike’.
- **Favour** How likely they would be to do a slightly inconvenient favour for the individual. Boundary labels: ‘Do Favour’ and ‘Don’t do Favour’.
- **Money** How likely they would be to return ten dollars that the individual had dropped. Boundary labels: ‘Return’ and ‘Keep’.

The first two question were intended to tap cognitive representations of reputation. First, by directly asking how ‘good’ each individual is. Second, since terms like ‘good’ may evoke normative answers that map imperfectly to
participants’ implicit sentiments, by asking about a vaguer sense of ‘liking’. The third and fourth questions were designed, to measure the behavioural consequences of reputation. The third solicited participant’s behaviour in the positive dilemma of doing someone a favour. Soliciting participant’s behaviour in negative dilemmas (i.e., opportunities to actively exploit someone) was challenging, since such behaviour is strongly proscribed in most contemporary societies. Based on pretesting, I settled on ‘not returning money’ as the closest approximation to the prototypical negative dilemma of ‘stealing’ which still generated at least some variability in people’s answers.

At the end of the study, I gave people a limitless text box and asked them to explain to us their rational for the choices they had made. These are available in their entirety in the supplemental materials.

I systematically varied the following aspects of this core design:

**Random between scenarios**

The following variables were randomly assigned to each participant in their first scenario and were different in the second scenario:

- **Names** The names of the Target and Actors (picked from one of two fixed sets: Tracy, Ashley, Betty, Cassie; and Stephanie, Judy, Katie, Lucy).
- **Setting** The setting in which the events took place (among friends or in the workplace).
- **DV Order** The order in which each of the four questions was presented.
- **First Actor** Whether the first Actor to face a dilemma is the Cooperator or Defector. The second Actor is always the other of this pair, while the third is always Neutral.

**Manipulated between participants**

The following variables were explicitly manipulated between conditions:
**Target Reputation Order** The reputation of the Target (either good or bad) in each scenario. All participants saw one scenario with a good Target and one with a bad Target. Which they saw first systematically varied between conditions.

**Source** The ostensive source of the information the participant is reading: their own experience, or gossip from a reliable source.

**Consequences** The final consequences for the target: none mentioned, bad (a social event fails) or good (a social event succeeds).

**Valence** The valence (positive or negative) of the dilemma faced by Actors.

In positive dilemmas, the Actors could pay a cost to provide a benefit to the Target. In negative dilemmas, Actors could extract a benefit by imposing a cost on the Target.

Here is an example scenario. In this instance, the Target (Stephanie) is cued as having a good reputation, the Source of the information is gossip and the Consequences are bad for Stephanie. The three Actors are Judy, Katie and Lucy. The difference between positively and negatively valenced variants of the dilemma are indicated with square brackets and bolded text. In the positive variant, Judy cooperates by doing something, in the negatively valenced version she defects by doing something. Katie does the opposite, but achieves it by inaction. Which of these (action or inaction, cooperation or defection) happened first was assigned at random and controlled for in regressions.

In both settings, positively and negatively valenced versions of the dilemma differed only in whether someone considered adding or taking pieces of paper from an unattended pile.

Imagine that you’ve recently started a new job and are getting to know your workmates. You’ve again met several people and learned a little about them.

One of the first people you met was Stephanie. You’ve noticed several things about Stephanie.
Stephanie is often the one who cleans the small, shared kitchenette at work, even though she’s rarely, if ever, the one who creates a mess.

When Stephanie works on a project with another workmate, she always makes sure to share the credit if it goes well and willingly shares the blame if it goes badly.

Whenever Stephanie meets a valuable client she puts them in contact with other relevant team members, even though she could get an advantage by being their exclusive point of contact.

Recently you witnessed some very interesting interactions involving Stephanie and three other individuals: Judy, Katie and Lucy. You heard about these events from a friend who managed to witness it all unseen. You trust this friend and are confident that this is all true.

One day at work Stephanie wanted to finish early because she was hosting a social function that evening that she needed time to organise. However Stephanie also had a very large stack of paperwork that she needed to finish before she could leave for the day. In fact, everyone in the office has to do identical paperwork from time to time, but Stephanie happened to have a lot due that day. Everyone could see that Stephanie was very anxious because if she didn’t have enough time to prepare for her social function it would likely be a failure. Even though each employee’s paperwork is indistinguishable, it is strict company policy that each employee must do their own, so Stephanie could not ask others for help.

In about the middle of the afternoon, Stephanie had to go to a meeting. She left her stack of paperwork on her desk unattended.

First, Judy came by and noticed the paperwork. She looked around and thought that no-one was watching. It was obvious that she could take some of Stephanie’s paperwork and do it herself / add some of her own paperwork to Stephanie’s pile and, if no-one saw her, then no-one would ever know. Judy stood and looked at the paperwork pile for quite a while, clearly considering her options. Eventually she decided to quickly take some of the paperwork from Stephanie’s large pile. Now she had more paperwork to do herself, but Stephanie had less. / quickly add some of her own paperwork into Stephanie’s large pile. Now she had less paperwork to do herself, but Stephanie had more.

Not long after Judy had left, Katie came by and also saw the pile.
She also looked around and didn’t see anyone watching and then stood and looked at it for a quite while, considering the same choice as Judy. Unlike Judy, Katie eventually decided to do nothing, she merely walked away.

Finally, after Katie had left, Lucy came by and also saw Stephanie’s unattended paperwork. She also looked around and then looked at it for a while, but before she could reach a decision, Stephanie came back. Lucy just ended up chatting with Stephanie briefly.

In the end, Stephanie did not finish the paperwork in time and her social event was a disaster.

**Operationalising reputation: dichotomous theory to continuous empiricism**

A major challenge of operationalising IR-theory is that formal models almost exclusively consider binary reputations (you’re either good or bad, there’s nothing in between), or sometimes discrete, small-integer reputations (Leimar & Hammerstein, 2001; Nowak & Sigmund, 1998). Real people seem to, prima facie, represent reputations continuously (you can be a little bit better reputed than me, a little less well than my sister), or at least readily offer continuous answers to reputational questions.

Theorising about binary reputations makes sense. Though I may represent your reputation continuously, when it comes to actually using your reputation to make a cooperative decision (e.g., to help you or walk away), I often have to make a binary choice. Either your reputation will be above a critical threshold, or not. By simplifying reputations as ‘above or below this threshold’, IR-models become more transparent and mathematically tractable, without sacrificing too much generality. In binary models, the work of representing continuous changes that do not cross the threshold is carried, implicitly, by the magnitude of various probabilistic scaling parameters, such as the relative frequencies with which events occur.

I translated models of dichotomous reputations into predictions about continuous reputational change by using the map in table 4.3. For instance, when a dichotomous theory predicts that an event would cause an individual to keep a good reputation but change their bad reputation to good, I
considered this equivalent to predicting that continuous reputations would ‘improve’ following the same event.

One particular challenging translation is posed by dichotomous assessment-rules that specify that, after an event, ‘good individuals become bad but bad individuals become good’. Two of the leading eight reputational systems specify such reputational ‘flipping’. However I have chosen to exclude this possibility from this investigation for both a theoretical and a practical reason.

Theoretically, flipping makes little sense for continuous reputations. There are two ways one could interpret it. First, it might imply that reputations always converge towards a neutral mid-point. In the long run this would, ceteris paribus, eliminate the reputational differences that sustain cooperation and cause the system to collapse. Alternatively they might over-shoot in the opposite direction, with good individuals becoming even more bad and bad individuals more good, causing ever more dramatic swings in continuous reputation. It is hard to imagine how this could be sustained in practice. Though both these alternatives could be ‘patched-up’ by additional assumptions (such as non-linear interaction probabilities), I know of no theoretical work that has attempted this. I suspect such systems are unlikely to play a large role in human cooperation, and in any case consider them beyond the scope of early empirical work.

Practically, ignoring such 3rd-order reputational flipping makes the empirical challenge of testing IR far more tractable. By restricting the information participants have to only 2nd-order (the Target’s reputation and Actor’s action, but not the Actor’s reputation), I reduce the complexity and length of scenarios they read, reduce their memory load and chances for confusion, and greatly rein in the combinatorial explosion of possible interpretations.

Thus, I assess only 2nd-order reputational systems, the ‘sensible six’

---

4So, for instance, by performing the same action twice (e.g., stealing your sandwich), I would be considered a horrible person and then be considered even better than I was before I’d done anything.

5Note that these are 2nd-order assessment-rules under my continuous translation (but not necessarily under a dichotomous interpretation). That is, an Actor’s previous reputation does not matter when assessing how their Action changes their reputation.
Dichotomous Theory Continuous Interpretation

| (G → G, B → G)  | Improvement (↑) |
| (G → G, B → B)  | No Change (≈)  |
| (G → B, B → B)  | Worsening (↓)  |
| (G → B, B → G)  | N/A            |

Table 4.3: A map from dichotomous reputations reputation changes (e.g., from good (G) to (→) bad (B)) to their continuous interpretations (e.g., continuous representations of the reputation will experience improvement).

of the ‘leading eight’. This restriction makes this project simpler, more tractable and likely more accurate, at the cost of neglecting the least plausible theoretical possibilities. My results could suggest that people are in fact using more complex strategies however, as you will see, they do not.

Continuous challenges: Boundaries and magnitudes

Moving to continuous interpretations of reputations entails confronting the limitations of measuring continuous quantities. I considered three alternatives. The first was magnitude-free measures, such as rank ordering each character on each question (liking, goodness, etc.). This has the advantage of offering unambiguous information about participants’ relative rating of each character, but offers no information about the magnitude of differences between them. For instance, a participant who judged a Defector on a bad Target as far better a Defector on a good Target but both worse than a Neutral or Cooperator, would be indistinguishable from a participant who disliked both defectors equally.

A second alternative is an unbounded continuous scale. For instance, participants could be asked to assign a number from negative to positive infinity describing their rating of each subject. Though this transparently captures rank-ordering and some index of magnitude, magnitude differences can be dramatically different between subjects and without extensive calibration for each subject, they are very difficult to compare.

The alternative I chose was a bounded continuous scale. Participants an-
sowied each question by moving a slider between two boundary labels. While this makes it easy to interpret magnitudes and compare participants, it introduces rank-ordering ambiguity at the boundaries. For instance, imagine an individual who sets the sliders for both Cooperator and Neutral maximally to ‘Return’ on the Money measure. Do they consider these two individual’s reputations as equal, or one as higher but both high enough to score maximally on this particular question? Consequently my analysis may over-estimate the frequency of no-change (≈) reputational systems.

Participants

Surveys were administered online, and participants were recruited via Amazon’s Mechanical Turk. This online sample that has been shown to give reasonably reliable answers to simple questions (Rand, 2012).

To detect possible cultural differences, I recruited participants from the two best-represented populations on Mechanical Turk: North Americans (35.4% male; mean (sd) age: 34.5(13.4)) and Indians (53.5% male; mean (sd) age: 31.1(10)). Both populations were presented with identical materials in English. 144 participants were recruited from each population. They were assigned to each condition in semi-random order: each condition was assigned once in random order, then each a second time, and so on.

Some participants were then excluded on the basis of three criteria. First, I logged all participants’ IP-addresses and any repeat addresses had all their instances excluded. Second, the bulk of participants took between eight and eighteen minutes to complete the study, however a small number managed it in under five minutes. I excluded these individuals as very likely to have been inattentive, since my pilot-subjects were not able to achieve the same when instructed to complete the survey as fast as they could. Third, participants were asked to type two specific, ordinally specified letters from the English alphabet; anyone failing this question was excluded. Excluded participants (41 Indians and 24 Americans) were replaced from their respective populations in a random order.

Participants were asked for the following demographic information: age
in years, sex, languages spoken, education level, self-reported ethnicity, and religious affiliation.

This study was initially designed purely for the American sample, and their data was gathered first. When it became clear that their judgements showed interesting patterns (detailed below), I extend the study to Indian participants. This was largely a convenience sample, since a large population of Indian participants are available on Amazon’s Mechanical Turk, the recruitment platform on which the American version had already been deployed. This is not an ideal cross-cultural comparison, since character names were all in English and the situations described were invented by English-speaking North American resident graduate student (myself). However, it does provide good preliminary insight into whether, approximately, reputational judgements are similar or strikingly different across cultures.

Analyses

Our design allows for two different kinds of analyses.

Continuous Analyses

First, I modeled participant’s judgement as beta-distributed (the natural distribution for bounded continuous measures), and constructed linear regression models to estimate the relationship of other variables to the mean ($\mu$, logit-linked) and precision ($\phi$, log-linked) of this distribution (Cribari-Neto & Zeileis, 2010; Ferrari & Cribari-Neto, 2004; Smithson & Verkuilen, 2006).

I was particularly interested in the effect of four main predictors and their interactions:

**Target-Good** A dichotomous indicator of whether the Target was cued a ‘good’ (vis-à-vis bad) in the scenario in which the judgement was made.

**Actor-Cooperator** Whether the individual being rated was the Cooperator (vis-à-vis defector).
Sample-Indian Whether the participant who made the rating was recruited from the Indian (vis-à-vis U.S.A.) sample.

Valence-Negative Whether the participant had read about a negatively (vis-à-vis positively) valenced dilemma.

I also controlled for a set of covariates in every model. Specifically: participants’ age, sex and how long it took them to complete the study, whether they saw the good Target at work or among friends, which scenario they saw first, which setting the particular data-point being regressed was judged in, the source of information (direct or gossip) and the ultimate consequences for the target (good, bad or not reported).

To calibrate each individual’s judgements in each scenario to their own baseline, I also included their rating of the Neutral character on the same measure as a covariate in μ-models.

To maximise statistical power, I analysed each judgement (not some aggregate of them) and modelled the within-participant covariance in answers by computing clustered robust confidence intervals.

Discrete Analyses

The continuous analysis attempts to infer population parameters (e.g., those of the beta-distribution that best describes participants’ responses). However this experimental design also allows me to, for each individual, to pick out precisely which of the 81 possible continuous, 2nd-order reputational systems their choices instantiated. We can ask whether each individual, for instance, considered a someone who defects who on a well reputed target as worse then someone who defects on a poorly reputed target.

To calibrate each individual’s judgement to their own within-scenario baseline, I first subtract each judgement from their rating of the Neutral character. This makes inferences robust to participants who were inclined to rate all characters better in one scenario (i.e., if they always favoured friends over workmates).

To discretely classifying a judgement difference as ‘improving’ (†), ‘worsening’ (¶) or ‘unchanged’ (≈) one must specify a minimum threshold for a
difference. Just how close must two judgements be using a 100-point mouse-based slider, before we assume the participant is trying to convey an identical judgment and attribute the difference to an unsteady hand, poor vision or laziness? While my research assistants could readily set two sliders exactly equivalent when asked to, I wanted to give participants a greater margin of error. Here I report results using a 2-point threshold. The supplemental materials also present results for 0-, 2- and 10-point thresholds (they are clearly labelled in the directory: /Chapter 4/discrete_strategy_tables/).

4.2 Results

These data are rich and amenable to many interesting analyses. Here I report those I found most pertinent to answering the questions outlined above. Table 4.5 summarises the key likelihood-maximising parameter estimates that underlie my conclusions. Full quantitative details are available in the supplemental materials (in the file /Chapter 4/regression_tables.pdf).

Did dependent measures capture reputational judgements?

Table 4.4 presents the correlations between my four dependent measures. I designed Linking and Good to measure the same underlying cognitive representation of reputation, and indeed they were highly correlated. They also correlated highly with judgments about doing the characters a Favour, but all three measures were more distinct from judgments about whether to return some lost Money to a character. This final measure was exceptional in that almost 45% of the 2304 judgments I recorded (one for each of four characters, in two scenarios, for two samples of 144 participants) were set to at their maximum boundary (participants would certainly return the money), as opposed to merely approximately 15% on other measures. I suspect this reflects compliance to social norms about how people should behave in these situations. Though this restricted variability dampens the sensitivity of this measure, it makes the 55% of deviations particularly interesting. If taking the money involves violating a normative proscription, it is even more plausibly an index of participant’s behavioural dispositions in
Table 4.4: Correlations between our four measures of participants’ reputational judgments

<table>
<thead>
<tr>
<th>Liking</th>
<th>Good</th>
<th>Favour</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good</td>
<td>0.86</td>
<td></td>
</tr>
<tr>
<td>Favour</td>
<td>0.82</td>
<td>0.8</td>
</tr>
<tr>
<td>Money</td>
<td>0.38</td>
<td>0.4</td>
</tr>
</tbody>
</table>

negative dilemmas (where a deliberate action is required to extract a benefit at others’ expense).

I am confident that these measures index participants’ representation of characters’ reputation.
Figure 4.1 (preceding page): American Participants’ absolute judgements of Targets and relative judgements of Actors. Judgements are divided by dilemma Valence (rows) and dependent measure (columns). Horizontal lines represent participants’ mean absolute judgements (left axis) and their 95% confidence intervals (obliquely shaded zones, estimate from beta-models) of Neutrals (red lines), good Targets (green, leftmost lines) and bad Targets (black, rightmost lines). Symbols represent the mean (pinpointed by a red dot) of the difference (right axis) in each individual participant’s judgements of Cooperators (angelic faces, 😇) and Defectors (crossed swords, ⚔) relative to their own Neutral judgement in the same scenario. Since symbols represent the difference between two beta-distributed variables, their 95% confidence intervals (transparently shaded zones) are asymptotically normal approximations.
Figure 4.2: Indian Participants’ absolute judgements of Targets and relative judgements of Actors. The same visual representation as Figure 4.1, for Indian participants.
<table>
<thead>
<tr>
<th>Sample (→)</th>
<th>Measure (↓)</th>
<th>U.S.A.</th>
<th>India</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model (↓)</td>
<td>Valence (→)</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Manipulation (TargetG)</td>
<td>Liking</td>
<td>4.14(0.29)**</td>
<td>3.57(0.30)**</td>
</tr>
<tr>
<td></td>
<td>Good</td>
<td>3.49(0.29)**</td>
<td>3.44(0.29)**</td>
</tr>
<tr>
<td></td>
<td>Favour</td>
<td>3.72(0.28)**</td>
<td>3.55(0.35)**</td>
</tr>
<tr>
<td></td>
<td>Money</td>
<td>1.85(0.45)**</td>
<td>3.24(0.43)**</td>
</tr>
<tr>
<td>1st Order (ActorC)</td>
<td>Liking</td>
<td>0.02(0.19)</td>
<td>2.45(0.22)**</td>
</tr>
<tr>
<td></td>
<td>Good</td>
<td>0.23(0.17)</td>
<td>2.51(0.19)**</td>
</tr>
<tr>
<td></td>
<td>Favour</td>
<td>0.26(0.21)</td>
<td>2.30(0.19)**</td>
</tr>
<tr>
<td></td>
<td>Money</td>
<td>−0.24(0.23)</td>
<td>1.38(0.19)**</td>
</tr>
<tr>
<td>2nd Order Cooperators (TargetG)</td>
<td>Liking</td>
<td>0.78(0.27)**</td>
<td>−0.10(0.18)</td>
</tr>
<tr>
<td></td>
<td>Good</td>
<td>0.43(0.22)^</td>
<td>−0.10(0.13)</td>
</tr>
<tr>
<td></td>
<td>Favour</td>
<td>0.52(0.26)^</td>
<td>−0.12(0.15)</td>
</tr>
<tr>
<td></td>
<td>Money</td>
<td>−0.23(0.32)</td>
<td>0.10(0.12)</td>
</tr>
<tr>
<td>2nd Order Defectors (TargetG)</td>
<td>Liking</td>
<td>−0.33(0.14)^</td>
<td>−0.62(0.18)**</td>
</tr>
<tr>
<td></td>
<td>Good</td>
<td>−0.27(0.12)^</td>
<td>−0.47(0.16)**</td>
</tr>
<tr>
<td></td>
<td>Favour</td>
<td>−0.51(0.18)^</td>
<td>−0.41(0.18)^</td>
</tr>
<tr>
<td></td>
<td>Money</td>
<td>0.09(0.26)</td>
<td>−0.42(0.16)^</td>
</tr>
<tr>
<td>2nd Order Difference (ActorC * TargetG)</td>
<td>Liking</td>
<td>0.77(0.42)^</td>
<td>0.63(0.33)^</td>
</tr>
<tr>
<td></td>
<td>Good</td>
<td>0.67(0.31)^</td>
<td>0.45(0.25)^</td>
</tr>
<tr>
<td></td>
<td>Favour</td>
<td>1.08(0.41)^</td>
<td>0.37(0.24)</td>
</tr>
<tr>
<td></td>
<td>Money</td>
<td>−0.12(0.42)</td>
<td>0.56(0.17)^</td>
</tr>
</tbody>
</table>
Table 4.5 (preceding page): **: $p < .01$, *: $p < .05$, ~: $p < .1$

Summary of key results for Chapter 4.
Coefficients and (standard errors) describing the relationship likelihood-maximising relationship between participant’s beta-distributed judgments and ‘whether the actor being judged is a Cooperator (vis-à-vis Defector)’ ($ActorC$), ‘whether the Target of their action has a Good (vis-à-vis Bad) reputation’ ($TargetG$), and their interaction, controlling for all manipulated variables (see methods), and participants’ age, sex and how long they took to finish. Full models with all covariate parameter estimates and precision estimates are available in the supplemental materials.

Divided by sample (U.S.A. or India, columns) and dilemma-valence (positive [+], or negative [−]) and four measures of participant’s reputational judgments (inner rows). Outer rows describe five analyses, showing the relationships between: Manipulation, Target judgements and our manipulation of the target’s reputation; 1st Order, Actor judgements and whether the actor cooperated or defected; 2nd Order Cooperator/Defector, judgements of each type of Actor, and whether their Target was cued as Good; 2nd Order Difference, the magnitude of the difference between judgments of Cooperators and Defectors and whether the Target was cued as Good.

Notice that (1) our manipulation appears effective; (2) 1st-order reciprocity is only evident in negative dilemmas; (3) we only see evidence for 2nd-order reciprocity in our US sample, though our Indian sample trends in the same direction; and (4) among our US sample 2nd-order differences are most evident in ‘behavioural responses in kind’: the MONEY measure responds to negative dilemmas while the FAVOUR measure responds to positive dilemmas.
Was the Target’s reputation manipulated successfully?

Across all measures, in all samples and valences, judgments of Targets were judged worse in ‘Bad scenarios’ (one’s where the manipulation attempted to worsen the Target’s reputation) than in Good scenarios. This is evident when regressing Target judgments on the reputation manipulation in each scenario (TARGET-GOOD), controlling for all covariates (see Supplemental Tables for a full exposition, and the first row of Table 4.5 for a quantitative summary). I also saw an interesting cultural interaction: the difference between Target judgements in Good and Bad scenarios was smaller among my Indian (mean judgements: .82 and .53) than among my U.S.A. sample (.94 and .39). While Target judgements still significantly covaried with the manipulation in India, it seems to have been less effective there.

Figures 4.1 and 4.2 make these patterns visually apparent by showing mean Target ratings in Good (green horizontal bar) and Bad scenarios (black horizontal bar) in both samples.

Did participants dislike defection, irrespective of Target reputation (i.e., first-order IR)?

In other words, were Cooperators rated better than Defectors? Yes, but much more so in negatively-valenced than positively-valenced dilemmas.

I regressed participants’ judgements on whether the Actor they were judging was a Cooperator or Defector (ACTOR-COOPERATOR), interacted with participants’ culture (SAMPLE-INDIAN) and dilemma valence (VALENCE-NEGATIVE), controlling for all manipulated variables (see methods), and participants’ age, sex and how long they took to finish. These analyses revealed a three-way interaction between these predictors, across all dependent measures. Consistently, Cooperators were judged better than Defectors (independent of Target reputation) in negative dilemmas, while effects in positive dilemmas merely trended non-significantly in the same direction (see Table 4.5, 1st Order row). This negativity bias was also consistently weaker (though still statistically significant) among my Indian sample. These results are depicted in figures 4.1 and 4.2 by the relative heights of Cooperator
symbols (✓) to Defector symbols (✗).

Were participants’ judgments contingent on the Target’s reputation (i.e., second-order IR)?

The strictest translation of dichotomous theories into continuous reality yields two predictions. Under Scoring, the magnitude of reputation change should be independent of the Target’s reputation. Under Standing, defecting on bad Targets should not reduce reputations at all, while defecting on good targets should. My data are inconsistent with both these strict predictions.

The ‘2nd order’ rows in Table 4.5, and the relative heights of green-shaded (good Target actions) and black-shaded (bad Target actions) symbols in Figures 4.1 and 4.2 show that, in contrast to the strict Scoring prediction, participants’ judgments did sometimes depend on Targets’ reputations.

The distance between the black-shaded region around defection symbols (✗), representing a 95% confidence interval around the mean judgment of defections against bad Targets) and the red line and its shaded 95% confidence interval (representing judgments of the Neutral character) shows that, in contrast to the strict Standing prediction, participants did condemn those who defected on Bad targets.

A more relaxed interpretation of Standing is that reputation change will be moderated continuously by a Target’s reputation. That is, defecting on bad Targets may still entail reputation loss, but less than defecting on less good Targets. My data supports this relaxed interpretation among my American sample. Among my Indian sample this was only a trend.

There are two ways to assess 2nd-order IR. First, we can ask how judgments of Defectors were related to the reputation of their target (TARGET-GOOD, Table 4.5, 2nd Order Defectors row). Across all measures, judgments were influenced by an interaction between valence and culture. Among Americans, across all measures and both valences (except MONEY in positive dilemmas) those who defected against Good Targets were judged worse than those who defected against Bad Targets. Indian participants tended to trend in the same direction, but only achieved statistical significance on the LIKING measure in negative dilemmas.
However when we ask the same question about Cooperators (Table 4.5, 2nd Order Cooperators row), we see no evidence that those who Cooperated with Good targets were judged better than those who cooperated with Bad targets in negative dilemmas. In positive dilemmas, only among Americans were those who Cooperated with good Targets preferred to those who cooperated with bad Targets, and even then only on the cognitive (GOOD and LIKING) and positive behavioural (FAVOUR) measures, not for negative behaviour (MONEY).

We can ask another 2nd-order question: ‘is the magnitude to which cooperators were preferred to defectors (i.e., the 1st-order effect) greater when the Target is Good?’ To answer this, I examined the interaction between target reputation (TARGET-GOOD) and Actor identity (ACTOR-COOPERATOR; Table 4.5, 2nd Order Difference row). Again my data reveal robust interactions between valence and sample across all measures. Americans showed substantial effects on the positive behavioural measure (FAVOUR) when judging actors in positive dilemmas, and reciprocally, a strong response on the negative behavioural measure (MONEY) when judging actors in negative dilemmas. In both cases, the difference between Cooperator and Defector ratings was greater when the Target was good. My cognitive measures (GOOD and LIKING), trended in the same directions as these effects, approaching statistical significance independently and achieving it if pooled.

Among my Indian sample, positive dilemmas trended similarly to Americans’ (achieving significance only for GOOD ratings), while negative dilemmas showed no evidence of a relationship, except for LIKING, which registered a significant effect of similar magnitude to Americans’.

To summarise, Americans showed the clearest evidence of 2nd-order indirect reciprocity, especially when responding behaviourally ‘in kind’. That is, participants’ 2nd-order responses were strongest when judging whether they would do a favour for someone they’d seen interacting in a positive dilemma (taking action to provide costly help, or refusing to do so), or when judging whether to keep someone’s misplaced money who had just interacted in a negative dilemma (taking action to gainfully exploit someone, or refusing to). As in other cases, my Indian sample tended to show ame-
riorated trends in the same direction, though effects here were particularly weak and inconclusive.

**Covariates and precision**

The supplementary regression tables specify detailed estimates of the relationships of the covariates to the models reported above. To gain more succinct insight into their overall impact, I regressed participant’s judgements on to these covariates across all measures.

The order in which scenarios were presented did not substantially affect participants’ judgements. There was no evidence of rating differences between the two kinds of scenarios (friends or work), though Indian participants’ tended to give higher ratings overall if they read about the ‘good Target’ in the context of work scenario.

Older participants tended to give higher ratings in all circumstance, especially in positive dilemmas. Males, on the other hand, tended to give lower ratings in positive dilemmas.

Those participants who took the most time to answers also gave more variable answers, and among American participants, tended to give higher ratings. Participants gave less variable ratings (and these ratings tended to be higher, but not statistically significantly) when the ultimate consequences for the target were mentioned, regardless of whether those consequences were good or bad. The fact that both ‘good’ and ‘bad consequences’ caused a shift in the same direction relative to ‘no consequences’ suggests that the mere reminder that the Actors’ actions had tangible consequences for the Target influenced their judgements.

There was no evidence that the ostensive source of the information (directly witnessed or heard from a reliable source) had any appreciable influence on judgements. Though, of course, the source in all cases was my online study and participants were merely being asked to imagine the scenarios, so this lack of effect should not be over-interpreted.
**Discrete Analysis**

To discretely test for successful manipulation of the Target’s reputation, I subtracted each participant’s Target judgement from their judgement of the Neutral character in the same scenario. I then categorised each participant according to whether, relative to Neutrals, they judged Targets who’d been cued as good better than those cued as bad ($G > B$), the reverse ($G < B$) or approximately equally ($G \approx B$) at various thresholds. Here I report results for a 2-point threshold (i.e., judgements within 2-points of each other are considered equal), other breakdowns are presented in the supplemental materials. I used a $\chi^2$ goodness of fit test to assess whether the number of participants in each category differed from expectation when randomly sampling from truly indiscriminate participants. I also used a simple non-parametric binomial test to assess whether the proportion of individuals favouring the Good target was higher than those favouring the Bad target. Participants’ judgements unequivocally responded to the manipulation. These results are presented in Table 4.6.

I performed parallel analyses for Cooperators (Table 4.7) and Defectors (Table 4.8). Unsurprisingly this reveals similar patterns to those gleaned from my continuous analyses. Those who cooperate with Good Targets are more often rated as better than those who cooperate with Bad Targets, especially in positive dilemmas. Those who defect on Bad Targets are rated better than those who defect on Good Targets, especially in negative dilemmas. These patterns are clearest among my American sample while Indian responses are noisier.
<table>
<thead>
<tr>
<th></th>
<th>Negative Dilemma</th>
<th></th>
<th>Positive Dilemma</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Liking</td>
<td>Good</td>
<td>Favour</td>
<td>Money</td>
</tr>
<tr>
<td><strong>U.S.A.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$G &gt; B$</td>
<td>67</td>
<td>67</td>
<td>67</td>
<td>29</td>
</tr>
<tr>
<td>$G \approx B$</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>41</td>
</tr>
<tr>
<td>$G &lt; B$</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>$Pr(\chi^2)$ $\approx$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$Pr(Binom)$ $\approx$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>India</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$G &gt; B$</td>
<td>51</td>
<td>54</td>
<td>53</td>
<td>34</td>
</tr>
<tr>
<td>$G \approx B$</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>23</td>
</tr>
<tr>
<td>$G &lt; B$</td>
<td>16</td>
<td>14</td>
<td>18</td>
<td>15</td>
</tr>
<tr>
<td>$Pr(\chi^2)$ $\approx$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.02</td>
</tr>
<tr>
<td>$Pr(Binom)$ $\approx$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
</tr>
</tbody>
</table>

**Table 4.6:** Frequencies of discretely categorised judgements of Targets. The raw number of individuals who rated Targets in Target-Good scenarios, relative to those in Target-Bad scenarios as better ($G > B$), worse ($G < B$) or approximately equally, at threshold 2, on each of our four measures, broken down by dilemma valence and sample. For each measure we include the p-value from a $\chi^2$ test of whether the proportions of individuals in each of the three categories is equal, and a binomial test of whether the proportion who judged Targets better when the Target was good differed from those judging them bettered when the Target was bad.
<table>
<thead>
<tr>
<th></th>
<th>Negative Dilemma</th>
<th></th>
<th>Positive Dilemma</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Liking Good Favour Money</td>
<td></td>
<td>Liking Good Favour Money</td>
<td></td>
</tr>
<tr>
<td>U.S.A.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( G &gt; B )</td>
<td>19</td>
<td>19</td>
<td>23</td>
<td>12</td>
</tr>
<tr>
<td>( G \approx B )</td>
<td>21</td>
<td>22</td>
<td>23</td>
<td>49</td>
</tr>
<tr>
<td>( G &lt; B )</td>
<td>32</td>
<td>31</td>
<td>26</td>
<td>11</td>
</tr>
<tr>
<td>( Pr(\chi^2) \approx )</td>
<td>0.13</td>
<td>0.2</td>
<td>0.88</td>
<td>0</td>
</tr>
<tr>
<td>( Pr(Binom) \approx )</td>
<td>0.09</td>
<td>0.12</td>
<td>0.78</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Liking Good Favour Money</td>
<td></td>
<td>Liking Good Favour Money</td>
<td></td>
</tr>
<tr>
<td>India</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( G &gt; B )</td>
<td>28</td>
<td>28</td>
<td>31</td>
<td>24</td>
</tr>
<tr>
<td>( G \approx B )</td>
<td>12</td>
<td>11</td>
<td>8</td>
<td>25</td>
</tr>
<tr>
<td>( G &lt; B )</td>
<td>32</td>
<td>33</td>
<td>33</td>
<td>23</td>
</tr>
<tr>
<td>( Pr(\chi^2) \approx )</td>
<td>0.01</td>
<td>0</td>
<td>0</td>
<td>0.96</td>
</tr>
<tr>
<td>( Pr(Binom) \approx )</td>
<td>0.7</td>
<td>0.61</td>
<td>0.9</td>
<td>1</td>
</tr>
</tbody>
</table>

**Table 4.7:** Frequencies of discretely categorised judgements of Cooperators. The raw number of individuals who rated Cooperators in Target-Good scenarios, relative to those in Target-Bad scenarios as better (\( G > B \)), worse (\( G < B \)) or approximately equally, at threshold 2, on each of our four measures, broken down by dilemma valence and sample. For each measure we include the p-value from a \( \chi^2 \) test of whether the proportions of individuals in each of the three categories is equal, and a binomial test of whether the proportion who judged Cooperators better when the Target was good differed from those judging them better when the Target was bad.
Table 4.8: Frequencies of discretely categorised judgements of Defectors. The raw number of individuals who rated Defectors in Target-Good scenarios, relative to those in Target-Bad scenarios as better ($G > B$), worse ($G < B$) or approximately equally, at threshold 2, on each of our four measures, broken down by dilemma valence and sample. For each measure we include the p-value from a $\chi^2$ test of whether the proportions of individuals in each of the three categories is equal, and a binomial test of whether the proportion who judged Defectors better when the Target was good differed from those judging them betted when the Target was bad.

<table>
<thead>
<tr>
<th></th>
<th>Liking</th>
<th>Good</th>
<th>Favour</th>
<th>Money</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>U.S.A.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$G &gt; B$</td>
<td>26</td>
<td>24</td>
<td>23</td>
<td>14</td>
</tr>
<tr>
<td>$G \approx B$</td>
<td>9</td>
<td>8</td>
<td>13</td>
<td>43</td>
</tr>
<tr>
<td>$G &lt; B$</td>
<td>37</td>
<td>40</td>
<td>36</td>
<td>15</td>
</tr>
<tr>
<td>$Pr(\chi^2) \approx$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$Pr(Binom) \approx$</td>
<td>0.21</td>
<td>0.06</td>
<td>0.12</td>
<td>1</td>
</tr>
<tr>
<td><strong>India</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$G &gt; B$</td>
<td>19</td>
<td>28</td>
<td>34</td>
<td>21</td>
</tr>
<tr>
<td>$G \approx B$</td>
<td>13</td>
<td>10</td>
<td>10</td>
<td>25</td>
</tr>
<tr>
<td>$G &lt; B$</td>
<td>40</td>
<td>34</td>
<td>28</td>
<td>26</td>
</tr>
<tr>
<td>$Pr(\chi^2) \approx$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.75</td>
</tr>
<tr>
<td>$Pr(Binom) \approx$</td>
<td>0.01</td>
<td>0.53</td>
<td>0.53</td>
<td>0.56</td>
</tr>
<tr>
<td><strong>Positive Dilemma</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$G &gt; B$</td>
<td>18</td>
<td>18</td>
<td>21</td>
<td>6</td>
</tr>
<tr>
<td>$G \approx B$</td>
<td>19</td>
<td>22</td>
<td>23</td>
<td>52</td>
</tr>
<tr>
<td>$G &lt; B$</td>
<td>35</td>
<td>32</td>
<td>28</td>
<td>14</td>
</tr>
<tr>
<td>$Pr(\chi^2) \approx$</td>
<td>0.02</td>
<td>0.11</td>
<td>0.58</td>
<td>0</td>
</tr>
<tr>
<td>$Pr(Binom) \approx$</td>
<td>0.03</td>
<td>0.06</td>
<td>0.39</td>
<td>0.12</td>
</tr>
<tr>
<td>$G &gt; B$</td>
<td>25</td>
<td>24</td>
<td>26</td>
<td>17</td>
</tr>
<tr>
<td>$G \approx B$</td>
<td>11</td>
<td>25</td>
<td>19</td>
<td>31</td>
</tr>
<tr>
<td>$G &lt; B$</td>
<td>36</td>
<td>23</td>
<td>27</td>
<td>24</td>
</tr>
<tr>
<td>$Pr(\chi^2) \approx$</td>
<td>0</td>
<td>0.96</td>
<td>0.45</td>
<td>0.13</td>
</tr>
<tr>
<td>$Pr(Binom) \approx$</td>
<td>0.2</td>
<td>1</td>
<td>1</td>
<td>0.35</td>
</tr>
</tbody>
</table>
It bears mention that since these tests are non-parametric, since these categories are necessarily approximate and since they were constructed by discarding all information about judgment magnitude, these discrete analyses are dramatically underpowered relative to the continuous analyses reported above. Though the continuous analyses are superior for assessing theoretically central questions about the environment a potential defector (or cooperator) faces, discrete analyses open up the possibility of examining the distribution of individual IR strategies among those judging behaviour, and provide a cleaner match to existing theory.

Our design allows us to classify each individual into exactly one of the 81 ($3^{2+2}$) potential 2nd-order reciprocity reputational systems. That is, they could have made three possible judgements (reputations improving, ↑; worsening, ↓; or not changing, ~) about two kinds of actions (Cooperate or Defect) on two kinds of Targets (Good or Bad). The supplemental materials include large tables that presents the proportion of individuals in each of these category, overall and split by sample and valence.

For simpler insights, these proportions can be averaged along their margins. The supplemental materials include tables that show how cooperators alone (and defectors alone) are judged, depending on their Target’s reputation and how those who interact with good Targets are judged, depending on whether they cooperated or defected. Here I present the analysis most pertinent to distinguishing Standing from Scoring, and for distinguishing among the ‘sensible six’ strictly 2nd-order leading eight Standing strategies: how actions on a Bad Target are judged. Table 4.9 shows the proportion of people (at a 2-point threshold) who judged Cooperating and Defecting on Bad Targets each of the three possible ways (rows and columns), on each of my four measures (four filled circles and proportions in each cell).

First, as discussed above, the Money measure was unique in that judging everyone equally was by far the dominant response. Second there is a great deal of heterogeneity in people’s reported reputational reactions. Though participants favoured some reputational systems on average, no single strategy was instantiated by the majority and almost the full gambit of possible responses received some representation. Third, the plurality of
participants (around 15%) instantiated a strict 1st-order Scoring strategy, while only around 10% were spread among the strictly defined Standing strategies (see full breakdown in supplemental figures). The marginal distributions of judgments about interactions with Bad targets (Table 4.9) are particularly informative. While in positive dilemmas Scoring is about equally represented as the two cooperator-favouring Standing strategies (Table 4.9, top row), in negative dilemmas it is clearly the dominant reputational system. However, this seems to be an epiphenomenon of an even stronger signal. In positive dilemmas (where cooperating involves deliberate action), most participants judge Cooperators as better than Neutral individuals and vary substantially in their opinion of Defectors, while the opposite is true in Negative dilemmas. Indeed, in negative dilemmas though the majority of participants rated Defectors as worse reputed that Neutrals, a substantial fraction of individuals rated Cooperators as equal to or worse than Neutrals—assessment-rules that theory has dismissed as evolutionarily unsustainable.

Overall, besides providing the first quantitative estimates of the distribution of reputational systems in modern populations, my discrete analysis highlights the considerable heterogeneity in reputational judgements, and the substantial impact that dilemma valence has on individuals’ judgements. It also underscores a key pattern from my continuous analysis: people most strongly and consistently respond to defections in negative dilemmas.
### Positive Dilemmas

<table>
<thead>
<tr>
<th>Cooperate</th>
<th>Defect</th>
<th>Cooperate</th>
<th>Defect</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\uparrow)</td>
<td>19% 11%</td>
<td>(\uparrow)</td>
<td>15% 12%</td>
</tr>
<tr>
<td>(\approx)</td>
<td>13% 10%</td>
<td>(\approx)</td>
<td>13% 9%</td>
</tr>
<tr>
<td>(\downarrow)</td>
<td>7% 4%</td>
<td>(\downarrow)</td>
<td>3% 8%</td>
</tr>
<tr>
<td>(\approx)</td>
<td>1% 1%</td>
<td>(\approx)</td>
<td>7% 1%</td>
</tr>
<tr>
<td>(\downarrow)</td>
<td>12% 15%</td>
<td>(\downarrow)</td>
<td>7% 9%</td>
</tr>
<tr>
<td>(\uparrow)</td>
<td>10% 6%</td>
<td>(\uparrow)</td>
<td>9% 6%</td>
</tr>
</tbody>
</table>

### Negative Dilemmas

<table>
<thead>
<tr>
<th>Cooperate</th>
<th>Defect</th>
<th>Cooperate</th>
<th>Defect</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\uparrow)</td>
<td>8% 4%</td>
<td>(\uparrow)</td>
<td>44% 48%</td>
</tr>
<tr>
<td>(\approx)</td>
<td>8% 4%</td>
<td>(\approx)</td>
<td>39% 22%</td>
</tr>
<tr>
<td>(\downarrow)</td>
<td>2% 2%</td>
<td>(\downarrow)</td>
<td>17% 16%</td>
</tr>
<tr>
<td>(\approx)</td>
<td>3% 3%</td>
<td>(\approx)</td>
<td>18% 10%</td>
</tr>
<tr>
<td>(\downarrow)</td>
<td>12% 15%</td>
<td>(\downarrow)</td>
<td>13% 15%</td>
</tr>
<tr>
<td>(\uparrow)</td>
<td>10% 6%</td>
<td>(\uparrow)</td>
<td>15% 11%</td>
</tr>
</tbody>
</table>
Table 4.9 (preceding page): A discrete partition of participants’ reputational judgements, aggregated by how Cooperators and Defectors are judged when they act upon Bad Targets. Judgements of individuals who either cooperated with (rows) or defected on (columns) badly reputed targets are sorted into those who considered them better (↑) than Neutral individuals (who were in an identical situation, but did not have an opportunity to act), worse (↓) or approximately equal (≈, within 2% of the width of the mouse-based slider participants used). The proportion of individuals in each category is listed and shown as the proportion of the circle’s area filled, for our four measures: whether you like the actor (top left circle in each cell), deem them ‘good’ (top right), would return their misplaced money (bottom left) or would do them a slightly inconvenient favour (bottom right). Theoretically predicted strategies are coloured: members of the ‘leading eight’ 2nd-order Standing strategies are shaded orange, while 1st-order Image Scoring is in green. Participants are divided by outer columns into those who read about positive and negative dilemmas.
4.3 Discussion

To the best of my knowledge, this is the first attempt to systematically map people’s reputational intuitions onto the space of possibilities outlined by Indirect Reciprocity (IR) theory by simply asking them. It has several limitations. First and foremost, I did not directly witness either the transmission of reputations nor people’s reputation-based reactions to others’ real behaviour. Instead, I merely asked people to report how they would respond to a variety of theoretically relevant situations. In so far as people are able to simulate hypothetical situations and have accurate introspective access to their reactions in those situations, this method should yield accurate insights and lay the way for more complex behavioural studies.

However, I do not believe we should put too much faith in introspection, nor do I advocate for the exclusive use of survey methods. Rather, this survey is a cheap and informative first pass at measuring people’s patterns of reputational intuitions where cognitive intuitions of specific individuals have been cued. Given the results here, we have good reason to attempt experimental investigations which somehow preserve the actual anonymity of our participants, while still cueing the person-specific cognitive mechanisms that underlie our reputational judgements.

Second, our materials were deployed over the internet. This allowed us to tap a larger, more diverse, cross-cultural sample and draw inferences about a broader population. However this came at the cost of sacrificing some control over our study population. I do not know who my participants were, except by trusting the information they reported (see Rand, 2012, for a justification of such trust). I also cannot be sure their responses were genuine and they were not just thoughtlessly ‘clicking through’ to earn small amounts of money. Though I used several measures to counteract such behaviour (IP tracking, timing, dummy questions), it is likely that inattentive or apathetic participants added noise to our data. However such mindless clicking should have varied randomly with respect to difference in our scenarios and conditions, so our statistical techniques should still be able to pick out the signal of genuinely-answering participants. If anything, the
noisiness of sample should give us more confidence in the patterns that were clearly detected. Furthermore, many participants’ qualitative accounts of their decisions show evidence of careful consideration of the scenarios and genuine judgement of the characters.

Third, to keep our design tractable and because for the reasons outlined above they are less plausible, I omitted the possibility of assessing 3rd-order reputational judgements (i.e., I cued Target’s previous reputations, but provided no information on Actors’ previous reputations).

Finally, this first foray used scenarios set in a workplace and among friends. It is quite plausible that the distribution of IR systems varies across domains—people may have different rules for business and family, for instance—and I am so far only able to assess a small component of this variability.

Despite these limitations, our investigation has yielded several useful insights.

Diversity of strategies
Theoretical models of IR typically describe systems of interacting individuals that are attracted to monomorphic equilibria—that is, they predict that everyone will eventually use the same assessment- and action-rules, or just assume that they always do. In stark contrast, participants in our survey instantiated a large variety of reputational systems. Even on our MONEY measure, which seemed strongly biased towards judging everyone equally, no single assessment-rules ever described the majority of participants’ judgements.

Though some of this heterogeneity is attributable to the noisiness of our sample and of people’s simulated judgements about imagined scenarios, there seems to be more going on. When one system did hold the plurality, it tended to be first-order Image Scoring, followed by members of the ‘leading eight’ Standing strategies. Rather than dominance by one clear equilibrium resident, the landscape of contemporary IR seems to be complex ecology of many different interacting strategies, assessment- and action-rules. There
is also some evidence that strategies vary between cultures (see the full
catalogue of per-culture discrete strategy breakdowns in the supplemental
materials, in the directory: /Chapter 4/discrete_strategy_tables/), but as
discussed above, my convenience-based Indian sample does not allow me to
confidently draw this conclusion.

It is also possible that, with the advent of more sophisticated institutions
like formal laws, police-forces, written contracts and so on, IR no longer plays
a major role in contemporary societies. That is, theoretical model of IR may
be correct, but these systems may currently be far out of equilibrium because
extrinsic institutions are exerting far stronger pressures on behaviour. If
this is correct, then the proper domain of IR is small-scale historical and
contemporary societies. A testable prediction follows from this logic: the
distribution of IR judgements should be far tighter in small scale societies
with fewer formal cooperation-sustaining institutions.

Dilemma valence matters

IR theories have almost exclusively considered the dilemma of helping some-
one or doing nothing (cf. Chapter three). Many theorists assume that these
same models also represent negative dilemmas, after all on paper the relative
payoffs are the same and one payoff matrix can be transformed into the other
by merely subtracting a constant. However in the real world these dilem-
mas are very different, as are people’s reputational responses to them. They
vary in whether cooperating (or defecting) involves deliberately, detectably
taking an action or merely ignoring an opportunity. Actually acting makes
it plain to any observers that an opportunity to act existed (whereas it may
otherwise slip by unnoticed), that the actor was aware of the opportunity
(which could otherwise be denied) and what choice the actor made.

Our participants’ strongest reputational responses were always to ‘active’
choices. That is, to cooperation in positive dilemmas and defection in neg-
ative dilemmas. However of these, defection in negative dilemmas triggered
by far the strongest reputational consequences, across both samples and set-
tings, of any variable in our design. This was the strongest, most consistent
signal in our data. These patterns are consistent with the negativity-biases (Baumeister et al., 2001; Cacioppo & Berntson, 1994; Rozin & Royzman, 2001) and omission-biases (Baron & Ritov, 2004; Cushman et al., 2006; DeScioli et al., 2011; Ritov & Baron, 1999; Spranca et al., 1991) that have been regularly observed in other domains of human behaviour. As far as I know, this study constitutes the first direct evidence that such biases also pervade second-order indirect reciprocity.

Furthermore among our American sample, who were more responsive to our manipulation in general, behavioural responses tended to match the valence of Actors’ actions. That is, when actors deliberately, actively defected (i.e., in negative dilemmas) participant’s responded by not returning some money they had misplaced (an active deviation from the social norm evident in our data), while if they active cooperated (i.e., in positive dilemmas) participant’s were more willing to go our of their way to do them a favour (a positive dilemma for participants).

Overall, our data suggest the IR theory should be more attentive to dilemma valence, and focus on defection in negative dilemmas in particular. This supports the key assumptions that set Negative Indirect Reciprocity (NIR, Chapter 3) apart from other IR models.

**Genes, culture or individual learning?**

IR theory is often used to explain how humans first started cooperating, providing an explanatory foundation for subsequent evolution of human brains and institutions. Under such accounts, if IR played an important enough role in human evolution, it may have left its signature on our IR-relevant intuitions and judgements. Specifically, we should see consistent reputational intuitions among people everywhere.

Our data only registered such a signal in one case: reputations fall when people actively defect in negative dilemmas. This happened consistently and strongly across both our samples. If humans share reliably developing, genetically encoded IR-relevant intuitions, our data suggest that they concern defection in negative dilemmas.
An alternative is that IR evolves culturally. That is, each individual’s intuitions emerge developmentally but are shaped by the distribution of others’ behavioural and reputational judgements in their society. Society-wide equilibria can emerge from these dynamic interactions, and persist over many generations. The interactions of societies at different equilibria could even lead to a form of cultural group selection. If this were so, we should see highly convergent judgements within groups and large differences between them. However, we did not. We saw a great deal of variability within-cultures and Indians’ responses were typically attenuated analogues of Americans’ responses. One exception is that while my American sample showed clear evidence of valenced reactions ‘in kind’, my Indian sample did not.

It is worth noting that our ability to draw cross-cultural inferences is limited since our Indian participants read materials in English, describing the interactions of English-named characters. While this first survey of the most conveniently accessible samples is useful, a robust answer to whether IR-intuitions vary culturally will require surveying a more phylogenetically distinct sample of cultures (America and India both inherited culture from the British), especially smaller-scale societies with few formal institutions.

Overall, the heterogeneous ecology of different reputational judgements I observed is most consistent with a) people developing their reputational intuitions by individual learning and b) doing so in a world were dynamic equilibria do not strongly attract them to any one system.

Overall conclusion
IR is potentially a very powerful tool for understanding both the origins and evolution of human behaviour and institutions. However the theoretical picture of IR is both complex and, on its own, inconclusive. Here I have attempted to wed this theory with data on the reputational judgements of individuals from two contemporary populations. I believe a productive next step would be to ask similar questions of individuals in small scale societies.
4.4 Supplemental

Additional supplemental materials are provided as computer files. They include:

- A list of all qualitative explanations that participants gave of their choices.

- Discrete partitions of participants into the IR-strategy types, for various threshold levels.

- Full regression tables for all analyses.

- The data file used to generate scenarios, from which any variant of the scenarios can (with a little cross-referencing) be reconstructed.
Chapter 5

It is better to profiteer on the guilty: is moral condemnation sensitive to reputation?

Moral psychologists continue to demonstrate that people’s moral intuitions can differ dramatically from what classic moral theories predict (e.g., Graham et al., 2009; Greene et al., 2009; Knobe, 2010). The last two decades have also seen rapid growth in the sophistication and scope of evolutionary models of human cooperation. These models explore how natural selection could produce a species with moral intuitions in the first place. For selection to have favoured contemporary individuals with pro-social moral intuitions, ancestral socio-ecological dynamics must have somehow protected cooperatively-disposed individuals from selfish ones. Formal evolutionary models specify how these mechanisms could work and imply predictions about how they would shape our psychology (recent examples include: Boyd et al., 2010; Panchanathan & Boyd, 2004; Sigmund et al., 2010, and chapter 3 of this dissertation).

These models have moved beyond simple kinship- or reciprocity-based
cooperation (Axelrod & Hamilton, 1981; Boyd & Lorberbaum, 1987; Hamilton, 1964; Smith, 1964; Trivers, 1971), which have broad applicability across species but struggle to explain many features of human cooperation in particular (Chudek et al., 2013b). Newer models probe the culture-gene coevolution of cooperative, cultural species like ours by emphasizing the role of socially and culturally transmitted reputations, norms and institutions (for a review, see Chudek & Henrich, 2011).

These two scholarly traditions have seen unfortunately little explicit overlap despite great potential for cross-fertilisation (Mesoudi, 2007; Mesoudi et al., 2006). There is a straightforward logical bridge. If humans evolved their moral intuitions in the selective environments described by these evolutionary models, *ceteris paribus*, moral psychologists ought to observe contemporary moral judgements and choices adapted to those environments.

Three good things can come of testing evolutionary models of cooperation with the tools of experimental moral psychology. First, psychological evidence can help reconstruct the picture of our evolutionary history, benefiting scholars across disciplines. Second, moral psychologist can ground their ultimate-level explanations in formal evolutionary theory, avoiding the highly speculative, untestable evolutionary story-telling that is too often invoked to justify moral intuitions. Third, formal evolutionary theory can direct our attention to moral phenomena we might not have otherwise considered.

Here I test a prediction about moral reasoning that follows from recent models of Negative Indirect Reciprocity (NIR, Chapter 3). These models describe a sequence of evolutionary developments that transformed early cognitive adaptations for forming friendships and coalitions (behaviour common to primates (Higham & Maestripieri, 2010; Langergraber et al., 2007; Perry & Manson, 2008; Seyfarth et al., 2012; Silk, 2002; Watts, 2002)) into a decentralised mechanism for enforcing individual adherence to arbitrary community-wide behavioural norms. To do so they assume our ancestors’ moral evaluations followed characteristic patterns, which ought to persist today.

To briefly summarise the logic of NIR, imagine that our ancestors regu-
larly encountered opportunities to exploit one another—to achieve some gain at one another’s expense—for instance by stealing food, dominating mating opportunities or taking advantage of another’s illness. Imagine also that they began socially-coordinating their individual opinions of one another into community-wide reputations. This might have been a consequence of early cognitive adaptations for generalised cultural learning (Boyd & Rich-erson, 2005; Mesoudi et al., 2006; Richerson & Boyd, 2004). Imagine that these individuals were less likely to exploit someone they liked (i.e., peers with a good reputation, their friends) than those they didn’t. Finally, imagine that individuals tended to dislike those who exploited their friends.

Since exploitation opportunities are often invisible unless they are acted upon (i.e., you would not know that I could have stolen your food cache, unless I do), an individual’s reputation can suffer if they exploit someone, but if they cooperative by inaction, it is likely to go unnoticed. This creates selective pressure for individuals to notice and act upon opportunities to raise their peers’ opinions of them (i.e., which, ultimately, causes others to exploit them less). NIR demonstrates how this pressure can eventually lead to a norm-psychology (Chudek & Henrich, 2011)—cognitive adaptations for being aware of, adhering to, and reacting negatively to the violation of social norms—which can act as the foundation for more sophisticated cooperation-sustaining mechanisms (e.g., Boyd et al., 2010; Panchanathan & Boyd, 2004; Sigmund et al., 2010).

NIR models suggest that humans are particularly likely to have genetically-adapted, reputation-based moral intuitions in negative cooperative dilemmas (opportunities to exploit), where defections are easily observed and consequences escalate when someone is exploited repeatedly. Meanwhile, in the socio-ecology described by NIR, positive cooperation (opportunities to help) are more likely to be sustained by norms and institutions, since they require communities to coordinate on shared expectations about when and how individuals should act. So, for example, NIR anticipates that across cultures and history, people’s reactions to someone stealing someone else’s resources ought to be consistent. Specifically, the thief’s reputation should fall, though in inverse proportion to the victim’s reputation. On the other
hand, NIR anticipates substantial, norm-driven cultural variability in people’s reputaitonal judgements of someone who donates resources to someone else.

Some recent empirical work (Engelmann & Fischbacher, 2009; Milinski et al., 2001, but cf. Bolton et al., 2005 and chapter 4) has suggested that modern humans assess reputations using first-order ‘scoring’ strategies (they consider any defection bad; Brandt & Sigmund, 2005; Nowak & Sigmund, 1998), rather than the theoretically more plausible second-order ‘standing’ strategies (they do not care if someone defects on a ‘bad guy’; Brandt et al., 2004; Engelmann & Fischbacher, 2009; Panchanathan et al., 2003). However this work has exclusively tested cooperation in positive (helping) dilemmas. If NIR correctly describes ancestral social dynamics, then modern humans ought to be more disposed to second-order reputational intuitions in negative dilemmas. That is, they should dislike people who exploit others, but not mind as much when the victims are ‘bad guys’.

A recent empirical discovery by moral psychologists provides a ripe opportunity to test this prediction. Inbar, Pizarro, & Cushman (2012) documented what I will call ‘the lucky profiteer effect’: people morally condemn those who profit from others loss, even if it is clear they did not contribute to causing it. These authors were motivated to challenge the prevailing assumption that individuals are considered blameworthy only if they caused harm or intended to. They presented their participants with vignettes describing an individual (let’s call him the profiteer) who makes a profit (e.g., on the stock market) from the misfortune of others (e.g., a devastating earthquake striking a third world country). Their participants rated both the moral blameworthiness of the profiteer’s action and the turpitude of his character. Across three studies participants disapproved of both the profiteer and his actions, even though he clearly did not cause the earthquake (in a fourth study, they observed similar patterns when the profit or loss of a company were at stake instead). Inbar et al. showed that this effect persists even when the profiteer merely positions themselves to benefit from a misfortune that does not ultimately occur, but disappears if the profiteer had other reasons for his actions (e.g., insuring his own assets in the third
world country). They conclude that people morally condemn the profiteer because they infer his wicked desires.

This may be just the pattern of moral condemnation assumed by NIR. NIR describes a socio-ecology where our ancestors condemned one another’s unjustified exploitation, however it is agnostic to how they identified such exploitation. Causal relationships are notoriously hard to identify, even for scientists armed with statistics. It is plausible that our ancestors recognized exploitation by merely noticing the coincidence of one individuals’ gain with another’s loss, without needing a clear causal link.

The mere coincidence of one individuals’ gain with others’ loss is still the driver of social discord, reputational damage and violent retribution in many societies (e.g., witchcraft accusations and the ‘evil eye’; Adinkrah, 2004; Elwedy, 1895; Evans-Pritchard, 1976). While NIR does not predict this disregard for causality, it does make clear predictions about the pattern of judgements and reputations that ensue after causes have been disregarded. Specifically, it predicts that the lucky profiteer effect should be ameliorated when profiteer’s victim has a bad reputation.

This chapter tests this prediction.

5.1 Methods, study one

I recruited 105 students from around the University of British Columbia campus, aged 18 to 25 years (mean = 19.6, SD = 1.7). To avoid confounds in comparing results to Inbar et al.’s, I drew a similar sample by including only adults who had grown up in North America and spoke English as their main language. Participants were randomly sorted in to one of two conditions that differed only in whether they read a scenario about criminals or refugees. The scenarios read (variants in square brackets):

Floret is a tiny, previously unpopulated island in the pacific that has recently become a haven for [refugees/criminals] fleeing from [persecution/justice] in their home countries. In fact, the island is now entirely peopled by [refugees/criminals]. Andrew, a stock-market trader, invested in special “disaster bonds” that
would only make a profit if a disaster hit Floret. A few months later a terrible hurricane hit Floret; many of the [refugees/criminals] were killed and the rest were left homeless and starving. Andrew made a large profit on his bonds and is now quite wealthy.

For brevity, I’ll refer to the two kinds of characters in these scenarios as ‘residents’ of Floret, and the ‘profiteer’ who, by luck alone, makes a profit from their suffering.

Each participant received a small candy bar for reading their scenario and rating their agreement with the following three statements (variable names in parenthesis).

(Good) I think Andrew is probably a good person.

(Approve) I disapprove of Andrew’s investment.

(Friends) I could probably be friends with Andrew.

I will refer to these questions by the variable names above. Each question was followed by a 10cm horizontal line with vertical dashes every 2.5cm, labelled ‘Strongly Disagree’ at the extreme left, ‘Strongly Agree’ at the extreme right and ‘Neither’ at the centre. Using a ruler, coders converted participants’ marks on this line to a 100 point metric. This was scaled to the open unit interval using the method described in (Smithson & Verkuilen, 2006).

The second question (Approve) was reverse coded so that higher values always imply a more positive evaluation of Andrew and his actions.

Since participants’ answered on a bounded scale, I considered model these answers as a beta-distributed random variable (Cribari-Neto & Zeileis, 2010; Ferrari & Cribari-Neto, 2004; Smithson & Verkuilen, 2006). This provided better fits, but qualitatively identical findings, as the standard technique of assuming an underlying normal distribution\(^1\).

\(^1\)That is, beta-models always had lower deviance than normal-models—they fit the data better. Deviance is the only relevant part of AIC, BIC and many other metrics of model comparison when both models are fit by estimating the same number of parameters. The normal and beta distributions both have two parameters. However the beta is supported on the bounded unit interval, while the normal, in theory, is boundless.
However, since my analyses were simple (i.e., no complex interactions like Chapter 4), involved only inferences about differences in central tendency not variability, and since the means of the variables I measured tended to be near the center of the interval, I found that assuming a normal distribution provided a very good approximation. To illustrate this, I have provided beta-regression confidence intervals beside normal-regression intervals in figures 5.1, 5.2 and 5.3. Below I provide the normal-regression parameter estimates, since they yield the same qualitatively conclusions but a) are likely to be more familiar to most readers, b) are easier to interpret (no link functions, beta-regression parameters are typically logit- and log-linked) and c) do not require an explicit regression model of variability (though do implicitly assume one), making it easier to communicate insights into differences between means.

5.2 Results, study one

5.2.1 Do these questions index the same underlying construct?
To answer this question we assessed the correlations between participants’ answers.

Participants who thought the profiteer was not a good person were more likely to disapprove of his actions \( r = 0.5 \), and thought themselves less likely to be friends with him \( r = 0.61 \). Participants who disapproved of the profiteer’s actions were less likely to be friends with him \( r = 0.47 \). Given these intermediate correlations, it is unclear whether these variables are all indexing a single underlying psychological disposition towards the profiteer or distinct but related judgements. As such, we report results for each variable and for an aggregated constructed by taking their arithmetic mean within participants.
5.2.2 Was condemnation of the lucky profiteer sensitive to the victims’ reputation?

Participants who saw the refugee scenario condemned the profiteer and his actions, and did so more than participants who saw the criminal scenario, who did not differ from a neutral response.

Including participants age and sex as covariates in regression models never produced even marginally significant predictors ($p > .5$), never improved model fit ($p > .9$) and did not qualitatively change the direction or significance levels of other model parameters.

The mean response of participants who saw the refugee scenario did differ significantly differed from ‘Neither agreement nor disagreement’ for each question and their aggregate. However, participants who saw the criminal condition did not differ significantly from responding ‘Neither’. Moreover, experimental condition was a significant predictor of the mean of participants’ ratings on all three measures and their aggregate (see Table 5.1 and figure 5.1).

Unlike Inbar et al., we did not include a contrast condition in study 1 where in which the profiteer benefited from the disaster not occurring. However, to the extent that the lucky profiteer effect requires participants to judged an agent more strongly than merely ‘neither agreeing nor disagreeing’, we only saw this effect when victims had good (refugees) but not bad (criminals) reputations.

5.3 Methods, study two

Study one suggests that the lucky profiteer effect is be ameliorated when the victim is poorly reputed. However, it has several limitations. First, I merely assumed that criminals were worse reputed than refugees, but did not measure it. Second, I used a WEIRD sample (Henrich et al., 2010): Western undergraduate students. Third, while I tested the reputational implications of NIR, I did not test its other key prediction: that people should have strong intuitions about negative cooperation in particular. That is, Andrew only had an opportunity to gain from others’ loss (a negative cooperative
Figure 5.1: Likelihood maximising estimates of the means of participants’ ratings (bar heights), across three measures (columns), across two conditions (bar colors) with 95% confidence internals. Beta-regression estimates of these same parameters are shown in grey; their mean is a circle. Ratings represent the proportion of the distance between ‘Strongly Disagree’ and ‘Strongly Agree’ that participants made their mark, scaled to [-0.5, 0.5] to correspond to regression models.
Table 5.1: Likelihood-maximising linear regression parameters of models describing participants’ normally-distributed ratings of the lucky profiteer, on our three measures and their aggregate (columns). Rows show the linear relationships of the listed predictors to these ratings. The dependent variable is the proportion of the distance along the rating line at which participant’s made their mark, scaled to [-0.5, 0.5] to facilitate statistical comparisons with the neutral mid-point.

<table>
<thead>
<tr>
<th></th>
<th>Good</th>
<th>Approve</th>
<th>Friend</th>
<th>Aggregate</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.01 (0.04)</td>
<td>-0.00 (0.04)</td>
<td>-0.03 (0.04)</td>
<td>-0.01 (0.03)</td>
</tr>
<tr>
<td>Residents: Refugees</td>
<td>-0.11 (0.05)*</td>
<td>-0.11 (0.06)^</td>
<td>-0.10 (0.05)*</td>
<td>-0.11 (0.04)*</td>
</tr>
<tr>
<td>Male?</td>
<td>0.03 (0.05)</td>
<td>-0.02 (0.06)</td>
<td>0.03 (0.06)</td>
<td>0.02 (0.05)</td>
</tr>
<tr>
<td>Age (years, centered)</td>
<td>0.00 (0.01)</td>
<td>0.00 (0.02)</td>
<td>-0.01 (0.01)</td>
<td>-0.00 (0.01)</td>
</tr>
</tbody>
</table>

**p < .01, * p < .05

dilemma), not to suffer a loss so that others could gain (a positive cooperative dilemma). This second study is designed to address these limitations.

I recruited 180 North Americans (48% male; aged 18 to 74, mean 36 years, std.dev. 13.9 years) via an online service (Rand, 2012, Amazon Mechanical Turk). I tracked both participants’ IP-addresses and Amazon Mechanical Turk ‘workerIDs’ to help ensure each was unique. I included a single question to test whether participants’ were paying attention and excluded anyone who failed it (11 individuals). I also asked participants to write a short paragraph explaining their decisions. All included participants wrote coherent paragraphs which agreed with their ratings.

Participants again read a paragraph about the island of Floret and ‘Alex’ a stock market trader (the profiteer), with several new systematic differences.

The residents of Floret were now one of:

(Criminals) War criminals fleeing justice during a brutal civil war in a nearby nation. These former warlords caused great suffering to innocent civilians in their efforts to accrue power and wealth for themselves.

(Refugees) Refugees fleeing persecution during a brutal civil war in a
nearby nation. These refugees experienced great suffering when they were caught in the crossfire of local warlords’ fight for wealth and power.

(Volunteers) Volunteer humanitarian aid workers, organising relief in a nearby nation during a brutal civil war. These volunteers take great personal risks to ease the suffering of innocent civilians, caught in the crossfire of local warlords’ fight for wealth and power.

Volunteers (who provide costly help to others, relative to Refugees) were added to investigate whether the reputational effects in study one are sensitive to any change in reputation, or are merely a response to very low reputations (Criminals). In study two I asked participants to judge both the residents of Floret and Alex, to measure whether this reputational manipulation was effective.

The profiteer now made one of two kinds of investments (emphasis did not appear in the originals):

(Miracle; negative cooperation) Alex deliberately and carefully constructed a complex portfolio of stocks, options and insurance policies which would only yield a large profit in the unlikely circumstance that something wonderful happened on Floret, something which brought wealth and prosperity to Floret’s residents. Otherwise, if something wonderful did not happen on Floret, the investment would yield a small loss.

(Disaster; positive cooperation) Alex deliberately and carefully constructed a complex portfolio of stocks, options and insurance policies which would only yield a large loss in the unlikely circumstance that something terrible happened on Floret, something which brought suffering and ruin to Floret’s residents. Otherwise, if something terrible did not happen on Floret, the investment would yield a small profit.

The addition of a corresponding opportunity for ‘lucky’ positive cooperation let me evaluate the sensitivity of the lucky profiteer effect to the valence of cooperation.
Finally, I now also varied the consequences for Floret between three scenarios: Intention Only, Lucky and Causal. I have further separated these by investment type and floret resident type. I have highlighted subtle differences in italics. These italics were not in the originals.

(Intention Only, Disaster) In the end nothing terrible happened on Floret, life went on as normal for Floret’s residents. So, as expected, Alex’s investment did not result in a large profit, instead it yielded a small loss.

(Intention Only, Miracle) In the end nothing wonderful happened on Floret, life went on as normal for Floret’s residents. So, as expected, Alex’s investment did not result in a large loss, instead it yielded a small profit.

(Lucky, Disaster) A few months later something terrible did happen on Floret that changed its residents’ lives. A terrible hurricane hit Floret. Many of the [criminals/refugees/volunteers] were killed and the rest were left homeless and starving.

(Lucky, Miracle) A few months later something wonderful did happen on Floret that changed its residents’ lives. Gold was miraculously found on Floret. Using this new found wealth, many of the

- criminals were able to bribe their ways out of prosecution and began undertaking even more ambitious crimes.
- refugees were able to successfully start new lives in peaceful countries.
- volunteers were able to successfully carry out their humanitarian projects and ease the suffering of others.

(Causal, Disaster) [(Lucky), followed by:] It also turns out that, by a stroke of luck, Alex found out about this impending hurricane shortly before it occurred, but long after Alex’s investment was set in stone. Alex may have been able to warn the people of Floret so they could
evacuate. This may have prevented Floret’s residents from suffering the disaster and stopped Alex from making a profit; but Alex chose to do nothing.

(Causal, Miracle) [(Lucky), followed by:] It also turns out that, by a stroke of luck, Alex found out about this impending gold discovery shortly before it occurred, but long after Alex’s investment was set in stone. Alex may have been able to inform various governments that could stake a claim on the gold. This may have prevented Floret’s residents from benefiting from the discovery and stopped Alex from making a loss; but Alex chose to do nothing.

NIR predicts, and previous work in other domains has found (Baron & Ritov, 2004; Cushman et al., 2006; Spranca et al., 1991), that Alex’s actions should have greater reputational consequences if they involve commission (i.e., making a phone call), rather than omission (i.e., not making a call even though he could have). To avoid conflating this effect with the mere addition of causality, Alex’s choice in ‘Causal’ scenarios is frame as an omission.

In this study I measured reputation judgements using the same measures described in Chapter 4. This facilitates more direct comparison between the two studies, and allows us to assess a proxy for participants’ own behaviour in both positive (FAVOUR measure) and negative (MONEY measure)dilemmas.

As a reminder, these measures are:

**Good** How good a person they believe each individual is. Boundary labels: ‘Good’ and ‘Bad’.

**Like** How well they think they would like each individual. Boundary labels: ‘Like’ and ‘Dislike’.

**Favour** How likely they would be to do a slightly inconvenient favour for the individual. Boundary labels: ‘Do Favour’ and ‘Don’t do Favour’.

**Money** How likely they would be to return ten dollars that the individual had dropped. Boundary labels: ‘Return’ and ‘Keep’.
5.4 Results, study two

5.4.1 Do these questions index the same underlying construct?

Table 5.2 presents the correlations among our dependent variables. Since our two measures of participants’ cognitive representations of reputation (GOOD and LIKING) were highly correlated (.83 for profiteer ratings, .87 for Floret resident ratings), we simplified analyses by taking their average as an index of participants’ cognitive representations of an individuals’ Reputation.

5.4.2 Was the manipulation effective?

In the case of criminals, yes. Participants consistently gave criminals lower ratings than refugees and volunteers across all measures (see Figure 5.2 and Table 5.3).

In the case of volunteers, less so. Though participants’ Reputation ratings of volunteers were significantly higher than their ratings of refugees this effect was about six times smaller than the difference between refugees and criminals, and was not reproduced by other measures.

Our scenarios seem to have more effectively caused participants to en-
Figure 5.2: Means (bar heights) of participants’ ratings of Floret residents (coloured columns), divided by the two kinds of cooperative dilemmas implied by the profiteer’s investment (negative: Disaster, and positive: Miracle), and the three kinds of consequences (outer rows), with 95% confidence internals inferred from normal (black) and beta (grey, means shown by circles) distributions. Lower bars imply that participants had a lower opinion of the residents being judged.
5.4.3 Did participants condemn the profiteer less for profiting on the suffering of the wicked?

Yes.

We can see evidence of the simple lucky profiteer effect in our participants’ tendency to move the slider from its initial mid-point (coded as 0.5) towards negative boundary label (coded as 0). The mean REPUTATION rating of a profiteer who profited from a disaster was 0.32 with a standard error of 0.03 (for full model details, see Table 5.4. It is unlikely that these data were sampled from a normal distribution with a mean of 0.5 ($p < .001$).

However, this effect was contingent on the person being judged (see Figure 5.3). Participants judged the profiteers’ REPUTATION more positively when those who suffered for their profit were criminals rather than refugees.

Table 5.3: Participants’ ratings of Floret residents in Study 2. Likelihood maximising parameters for regression models of participants’ normally distributed ratings of Floret’s residents (dummy variable reference level: refugees), across our three measures (columns). The dependent variable corresponds to the position of participants’ sliders between the boundaries, scaled to $[-0.5,0.5]$ to facilitate comparisons with the slider’s initial midpoint position.

<table>
<thead>
<tr>
<th></th>
<th>Reputation</th>
<th>Favour</th>
<th>Money</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.26 (0.04)**</td>
<td>0.40 (0.05)**</td>
<td>0.52 (0.05)**</td>
</tr>
<tr>
<td>Age (years, centered)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)*</td>
</tr>
<tr>
<td>Male?</td>
<td>0.01 (0.03)</td>
<td>0.00 (0.04)</td>
<td>0.00 (0.04)</td>
</tr>
<tr>
<td>Residents: Criminals</td>
<td>-0.50 (0.04)**</td>
<td>-0.47 (0.04)**</td>
<td>-0.27 (0.05)**</td>
</tr>
<tr>
<td>Residents: Volunteers</td>
<td>0.08 (0.04)*</td>
<td>0.05 (0.05)</td>
<td>-0.03 (0.05)</td>
</tr>
<tr>
<td>Scenario: Causal</td>
<td>0.01 (0.04)</td>
<td>-0.03 (0.05)</td>
<td>-0.11 (0.05)*</td>
</tr>
<tr>
<td>Scenario: Intention Only</td>
<td>-0.03 (0.04)</td>
<td>0.00 (0.05)</td>
<td>-0.10 (0.05)^</td>
</tr>
<tr>
<td>Investment Type: Miracle</td>
<td>0.01 (0.03)</td>
<td>-0.12 (0.04)**</td>
<td>-0.08 (0.04)^</td>
</tr>
</tbody>
</table>

**: $p < .01$, *: $p < .05$
Table 5.4 shows the likelihood maximising regression parameters that describe this relationships. Whether the profit was blind luck (Lucky scenario, $\beta = .23, p = 0.04$) or causal inaction (Causal scenario, $\beta = .22, p = 0.01$), a profiteer benefiting from harm to criminals was judged better than one benefiting from harm to refugees. However this reputation-contingent difference did not obtain when profiteer made their investment but no disaster occurred (Intention Only scenario, $\beta = .01, p = 0.93$).

While mere intention does not produce a second-order difference in ratings (i.e., between criminals and refugees/volunteers), the first order profiteer effect still persists. Though the mean of participant’s ratings it is not quite statistically significantly different from the neutral mid-point (coded 0 in Figure 5.3 and Table 5.4), it also does not significantly differ from their ratings in the Lucky scenario (refugees: $p = .33$; volunteers: $p = 0.41$). This second test is also the strategy that Inbar et al.’s used to arrive at their convergent conclusion: that the profiteer effect persist even when the disaster does not occur.

Participant’s were also less willing to pocket the lost money of a profiteer who had exploited criminals in Lucky scenarios (i.e., they were more willing to defect in a negative dilemma on someone who had (luckily, indirectly) benefited from the suffering of ‘good guys’), however their willingness to do them a favour (a positive cooperative dilemma) did not covary with resident reputation in these negatively-valenced ‘disaster investment’ dilemma scenarios.

### 5.4.4 Did participants reward the trader for taking a loss in miracle scenarios?

That is, did they judge Alex the stock market Trader more positively when Alex made investments that would yield a large loss when others fortuitously gained? Did their judgements depend on who these others were, and whether Alex causally influenced these events?

The only clear effect in miracle conditions is that traders who suffered for the good of criminals were slightly less-well liked than those who suffered for volunteers or refugees (see Figure 5.3 and Table 5.5). In these cases, the
Table 5.4: Participants’ ratings of the lucky profiteer in disaster scenarios. Likelihood maximising parameter estimates for regression models of participants’ ratings of the Trader in ‘disaster’ scenarios (a positively valenced dilemma), for the three measures (outer rows) and three levels of causation (columns). The dependent variable corresponds to the position of participants’ sliders between the boundaries, scaled to [-0.5,0.5] to facilitate comparisons with the slider’s initial midpoint position.

<table>
<thead>
<tr>
<th></th>
<th>Intention Only</th>
<th>Lucky</th>
<th>Causal</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-0.16 (0.10)^</td>
<td>-0.20 (0.08)^</td>
<td>-0.42 (0.07)^**</td>
</tr>
<tr>
<td>Age (years, centered)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>-0.00 (0.00)</td>
</tr>
<tr>
<td>Male?</td>
<td>0.10 (0.09)</td>
<td>0.07 (0.09)</td>
<td>0.08 (0.07)</td>
</tr>
<tr>
<td>Residents: Criminals</td>
<td>0.01 (0.12)</td>
<td>0.23 (0.11)^*</td>
<td>0.22 (0.08)^**</td>
</tr>
<tr>
<td>Residents: Volunteers</td>
<td>-0.00 (0.11)</td>
<td>-0.04 (0.11)</td>
<td>-0.02 (0.08)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Intention Only</th>
<th>Lucky</th>
<th>Causal</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-0.09 (0.12)</td>
<td>-0.06 (0.13)</td>
<td>-0.09 (0.13)</td>
</tr>
<tr>
<td>Age (years, centered)</td>
<td>0.00 (0.00)</td>
<td>0.01 (0.01)</td>
<td>0.00 (0.00)</td>
</tr>
<tr>
<td>Male?</td>
<td>0.18 (0.12)</td>
<td>0.17 (0.14)</td>
<td>0.01 (0.13)</td>
</tr>
<tr>
<td>Residents: Criminals</td>
<td>-0.04 (0.15)</td>
<td>0.13 (0.17)</td>
<td>-0.01 (0.16)</td>
</tr>
<tr>
<td>Residents: Volunteers</td>
<td>-0.07 (0.14)</td>
<td>-0.10 (0.18)</td>
<td>-0.17 (0.16)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Intention Only</th>
<th>Lucky</th>
<th>Causal</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.09 (0.14)</td>
<td>0.02 (0.12)</td>
<td>0.23 (0.15)</td>
</tr>
<tr>
<td>Age (years, centered)</td>
<td>0.00 (0.00)</td>
<td>0.01 (0.01)^</td>
<td>0.01 (0.01)</td>
</tr>
<tr>
<td>Male?</td>
<td>0.26 (0.14)^</td>
<td>-0.09 (0.14)</td>
<td>-0.14 (0.16)</td>
</tr>
<tr>
<td>Residents: Criminals</td>
<td>-0.01 (0.17)</td>
<td>0.40 (0.17)^*</td>
<td>0.11 (0.19)</td>
</tr>
<tr>
<td>Residents: Volunteers</td>
<td>0.03 (0.17)</td>
<td>-0.19 (0.17)</td>
<td>-0.04 (0.18)</td>
</tr>
</tbody>
</table>

The mean of participants’ ratings for in refugee conditions were significantly more positive than the slider’s initial half-way point, but the not in others. This was particularly evident on the FAVOUR question, a measure of participants’ positively-valenced dilemma responses to an individual who they had seen cooperating in a positive dilemma.

Furthermore, what average differences did exist were driven by judgements in the ‘Causal’ condition. That is, we saw almost no evidence that accidentally incurring costs to help someone incurs reputational consequences they way that accidentally profiting from their loss does.
Figure 5.3: Means (bar heights) of participants’ ratings of the profiteer (coloured columns), divided by the two kinds of cooperative dilemmas implied by the profiteer’s investment (negative: Disaster, and positive: Miracle), and the three kinds of consequences (outer rows), with 95% confidence intervals inferred from normal (black) and beta (grey, means shown by circles) distributions. Lower bars imply that participants had a lower opinion of the profiteer.
Table 5.5: Participants’ ratings of the lucky profiteer in miracle scenarios. Likelihood maximising parameter estimates for regression models of participants’ ratings of the Trader in ‘miracle’ scenarios (a positively valenced dilemma), for the three measures (outer rows) and three levels of causation (columns). The dependent variable corresponds to the position of participants’ sliders between the boundaries, scaled to [-0.5,0.5] to facilitate comparisons with the slider’s initial midpoint position.

5.5 Discussion

In two studies we detected a previously unobserved nuance in Inbar et al.’s Lucky Profiteer effect. Our participants only condemned those who profited on others’ misfortune, even if only by luck, when the victims were well-reputed (refugees fleeing persecution, or volunteers in a war-torn country), not ill-reputed (criminals fleeing persecution). This is the pattern of moral intuitions that NIR models assume shaped our ancestral socio-ecology.

I also observed a clear negativity-bias (Baumeister et al., 2001; Cacioppo
& Berntson, 1994; Rozin & Royzman, 2001). The Lucky Profiteer effect was only clearly present in negative cooperative dilemmas (where the profiteer makes a profit from another’s large loss), not positive ones (where they suffer so others may profit). Also, describing of an island of people who profited from other’s misery (war criminals) caused participants’ esteem to fall far more than describing people who made great sacrifices to help others (volunteer aid workers) caused them to rise. These unambiguous negativity biases in reputational evaluation match the ancestral patterns of reputational evaluation assumed by NIR.

In addition, I observed that though causal influence is not essential for people to condemn the coincidence of one person’s profit from another’s loss, adding an opportunity for causal influence greatly exacerbates the effect.

Finally, as in chapter 4, these data suggests that people tend to respond ‘in kind’ to valenced dilemmas. That is, our positively-valenced FAVOUR showed stronger reactions to positive dilemmas and our negatively-valenced MONEY measure responded more strongly in negative dilemma scenarios. This supports NIR’s novel emphasis on the valence of cooperative dilemmas, both now and historically.

From an ultimate explanatory perspective, NIR is agnostic to the proximate logic of why our moral intuitions follow these pattern. It may be, as Inbar et al. suggest, that we infer the profiteer’s wicked desires. Though this account does not prima facie explain why these effects should depend on the victim’s reputation, or why this should cease to be the case if no consequences follow. Or perhaps it is connected to feelings of fictive-affiliation or friendship with well-reputed victims but not ill-reputed ones, or it may simply be a raw, irrational, genetically-encoded moral intuition. Ultimate, evolutionary models like NIR are concerned with teasing out the long-run, dynamic evolutionary consequences of behaviour, whatever the proximate motivations that cause it.

While our study supports NIR as a model of the social dynamics that shaped our ancestors’ moral intuition, no single feature of contemporary psychology can, in isolation, prove an ultimate, evolutionary hypothesis. Any human behaviour we measure today is shaped by genetically evolved
dispositions, individual learning and development and, crucially, our rich corpus of culturally transmitted knowledge. There will always be many proximate explanations that fit contemporary psychological facts better than a distal hypothesis, whose predictions are necessarily less precise (though the two levels of explanation often support one another).

Rather, formally specified evolutionary theories ought to generate a broad range of predictions across many levels of explanation, including psychological phenomena, sociological and cross-cultural distributions of behaviour, patterns of historical change, and so on. The relevance of an ultimate theory to understanding our moral intuitions is arbitrated by the empirical fit of the breadth of its explanatory scope—it’s ability to connect a broad range of otherwise seemingly disconnected findings from across psychology and other disciplines—rather than its specificity for any particular phenomenon. Testing ultimate theories is necessarily an interdisciplinary effort, to which this study is a contribution.

Here we take a first step by testing one central cognitive prediction of NIR in one contemporary social group. Conclusively testing whether the lucky profiteer effect is consistent with NIR’s assumptions requires a far more ambitious project; for instance, systematic cross-cultural studies (e.g., Henrich et al., 2005) or comparisons of cultures whose phylogenetic relationships are known (e.g., Currie et al., 2010; Mace & Jordan, 2011).

In the meantime, we can ask whether it is even plausible that our moral intuitions were shaped by ancestral social dynamics at all? We believe it is. Our genome bears the signature of recent and ongoing positive selection (Laland et al., 2010; Nielsen et al., 2007) and some digestive adaptations (e.g., lactase persistence: Ingram et al., 2009; alcohol metabolism: Peng et al., 2010) have emerged and spread through human populations since the advent of agriculture (i.e., on time-scales of thousands of years). It is plausible that cognitive adaptations that make our moral intuitions sensitive to reputation have done the same, so long as we can be confident that reputations carried potent fitness consequences for similar, or longer, timescales.

While reputations do not directly leave archaeological footprints, there is some indirect evidence of their presence. Anatomically modern humans—
whose skulls closely resemble ours, and who were at least physiologically capable of gossip—began rapidly migrating across the world’s diverse ecologies by 50 thousand years ago (Endicott et al., 2009; Hudjashov et al., 2007). Our less cognitively-modern hominids cousins achieved a similar feat over a million and a half years earlier. Convergent evidence from multiple far-flung sites makes a convincing case that hominids had mastered the controlled use of fire between 200-300 thousand years ago (e.g., Karkanas et al., 2007; Roebroeks & Villa, 2011), and plausible but more controversial case that fire use may have been present up to almost 2 million years ago (e.g., Clark & Harris, 1985; Gowlett et al., 1981; Wrangham, 2009). Together these suggest that up to two million years ago genus homo may have coordinated novel forms of cooperation (e.g., cooperative hunting) in new ecologies and transmitted complex cultural information (e.g., fire-use techniques, foraging and hunting techniques, ecological-lore, etc.). Transmission of complex culture is itself a cooperative dilemma, since dependence on culturally learned knowledge requires that individuals trust others, making them susceptible to exploitative deception. Since reputations are one of the most plausible mechanisms our early ancestors had for rapidly generating and sustaining cooperation in arbitrary domains, they have likely shaped our intuitions for tens of thousands of years (since our recent, rapid migration from Africa), and plausibly for hundreds of thousands (since our ancestors were anatomically modern) or millions (since the first migrations from Africa, the advent of stone tools and fire use); long enough to leave their mark on our genome and our cognition.

If, as our study suggests, further investigation concluded that NIR dynamics did shape human moral intuitions—such that we are readily inclined to exploit those with bad reputations, and to condemn others for doing so—we would be in possession of a powerful explanatory tool. It could, for instance, help make sense of many innocuous every day phenomena, such as why many of us only feel excited when an action movie hero kills many ‘bad guys’, but gnash our teeth when a villain in the same movie threatens a single innocent individual. More practically, this evolved moral intuition could help explain and intervene in the recurrent emergence of bullying in
schoolyards across the world; where some low status children are systematically degraded and exploited by some of their peers, while others look on with little or no moral outrage or inclination to intervene. More potently still, an NIR account of the Innocent Profiteer effect could offer us traction on the widespread historical and contemporary (Adinkrah, 2004) incidence of witchcraft accusations; where the coincidence of some people’s loss with other, low-status individuals’ gain (or lack of similar loss) prompts violent, sometimes deadly, retaliation by their societies.

As ever more comparative (e.g., Kitayama & Uskul, 2011; Nisbett, 2004) and phylogenetic (e.g., Currie et al., 2010; Mace & Jordan, 2011) research demonstrates that deep difference in human reasoning and behaviour are transmitted culturally, the onus increasingly falls on psychologists to move beyond extrapolating from convenient but unrepresentative western undergraduate samples (Henrich et al., 2010), and begin detailing the specific developmental, cultural and evolutionary trajectories that shape psychological effects. This especially true for moral psychology, where the universality and origins of moral intuitions are a focus of inquiry (e.g., Mikhail, 2007). Extrapolating and testing the predictions of formal evolutionary models is an important and underutilized tool for probing the ultimate roots of modern moral intuitions. We hope that this simple study can provide a model for further integration of formal evolutionary theory and moral psychology.
Chapter 6

Conclusions

Typical graduate dissertations in social psychology focus on one specific empirical phenomenon, such as prejudice, helping or religious belief. Social psychologists probe the intricacies of such phenomena with clever, sometimes wonderfully creative empirical techniques. Theoretical descriptions of social psychological phenomena usually evolve alongside the discovery of new evidence about them. The theoretical terms in which these descriptions are couched mix together concepts from other social-psychological investigations, folk-psychology and culture-specific ideas about the functional entities that inhabit minds. Often the assumptions made by such theories are implicitly hidden in the chaotic intellectual history of these terms and concepts. Gradually, collectively, these assumptions and concepts are refined as researchers (who are motivated to publish papers) preferentially borrow and recombine the concepts that have previously been most successful in generating exciting empirical effects.

There is a lot to recommend this approach. For one thing it is highly generative. Social psychologists have unveiled an impressive terrain of previously unsuspected human peculiarity (for the distribution of effect sizes they have discovered, see Richard et al., 2003). When the microscope was first discovered, it was a terrific idea for countless scientists to peer through it at whatever struck their fancy and describe the most exciting things they saw. This early excited phase of exploration provides the raw fodder from
which more general theories are built. Our recently discovered statistical and experimental methods are like a macroscope, lending us vision of another world not visible with the naked eye. It is a world inhabited by subtle patterns of human behaviour and judgement only apparent on aggregate, in the mean differences between individuals; individuals experimentally influenced just so.

I took a different approach.

Rather than trying to explain a particular contemporary empirical phenomenon, this dissertation is a contribution to the top-down challenge of retracing the biological origins of our social psychology. Rather than recombining the stew of often only partially mutually-consistent contemporary psychological concepts into a good description of one aspect of our social psychology and working up towards a more general understanding, this dissertation embarks from a purely theoretical challenge: explaining how our ancestors’ biological realities could have given rise to a social psychology in the first place.

To rigorously confront this challenge, I have restricted myself to only invoking concepts, terms and ideas that can be clearly, formally defined from the stand-point of evolutionary biology. I only allowed myself to use ideas that I would make sense to an abstract, non-human, external observer who had so far only managed make sense of chimpanzees.

I looked for combinations of these concepts which could generate the kind of explosion of cooperative, complex, cultural peculiarity that characterises humans today. I considered only combinations that could be stated with formal, mathematical rigour and were consistent with the principles of evolution by natural selection. My efforts interface with those of scholars across the sciences, who are already embarked on this enterprise.

In chapter 2, I reviewed an emerging picture of the evolutionary changes that launched our species on its unique trajectory. At the center of this picture is the idea that ‘culture’—our ability to transmit complex, encoded information across generations—made possible exciting new forms of sociality. It let us cooperate in ways that no other species has previously accomplished. These accounts are plagued by a puzzle: the cooperative dilemma.
of culture, the evil teacher problem. An evolving species cannot accumulate a shared corpus of a sophisticated cultural knowledge until some very general, very robust mechanism ensures they cooperate. Why? Because culture itself is a valuable public good that you can only acquire from other minds, and so is eminently exploitable.

In chapter 3, I proposed a novel solution to the cooperative dilemma of culture: Negative Indirect Reciprocity. I took one of the best existing candidates for a uniquely-human solution to the dilemma (reputation) and honed it until it assumed as little pre-existing culture as possible. This new model tells a precise story about one part of humans’ transition to modernity which, if accurate, can help explain why more sophisticated behaviours, cognitive adaptations and institutions exist. The model hinges on several key assumptions about our ancestors’ now unobservable psychology.

In chapter 3, I went on to argue that robust biases in contemporary human cognition—negativity- and omission-biases—provide support for these assumptions. However NIR’s assumptions are also more specific. NIR assumes that ancestral patterns of indirect reciprocity (in particular, 2nd-order reputational judgements) were biased in specific ways. To the best of my knowledge, prior to this dissertation no direct evidence existed of negativity- and omission-biases in 2nd-order reputational judgements.

NIR’s assumptions can be tested using the techniques of social psychology. If our ancestor’s socio-ecology was biased in the ways assumed by NIR, then our contemporary reputational judgements ought to be similarly biased. The subsequent chapters carried out these tests.

From the vantage of NIR, there are a real and important differences between positive cooperation—paying an absolute cost to beget an absolute benefit for others—and negative cooperation—paying a relative cost (by forfeiting the benefits of exploitation) to bring about a relative benefit (not having to suffer a loss). This stands in stark contrast the perspective implied by the mainstay of current empirical and theoretical inquiry into cooperation, which typically only considers positive cooperation and punishment (paying a cost to cause a cost, a very different phenomenon).

Our current scholarly paradigm implicitly (and even explicitly in conver-
sations I’ve had with influential scholars) assumes that positive and negative cooperation are equivalent—that insights about one are mere corollaries of investigations of the other. The evidence presented in this dissertation speaks against this assumption. Chapters 4 and 5 documented studies where a small manipulation of dilemma valence—the difference between adding and taking pieces of paper from an unattended pile in chapter 4, between investing indirectly in a disaster or a miracle in chapter 5—generated vastly different moral and reputational reactions from participants. NIR (chapter 3) suggests that the valence of cooperation matters in theory, chapters 4 and 6 show that it matters in practice too.

An NIR perspective on our evolutionary history assumes that negative dilemmas influence reputations more than positive ones. This was evidenced in contemporary cognition by both chapters 4 and 5. In both cases manipulations designed to worsen an individual’s reputation by making it clear to participants that they exploit others were more effective than attempts to improve their reputation by making it clear they help others. That is, negative dilemmas seem to have a stronger effect on reputations than positive ones. This effect is consistent with the ‘negativity-bias’ that pervades many domains of human cognition. The generality of this bias suggests it may be an ancient genetic adaptation to cognition, rather than a recent cultural adaptation in reputation assessment. This generality strengthens our inference from our contemporary observations to ancestral populations, building our confidence that our ancestors reputational evaluations were negativity-biased too.

In both chapters 4 and 5, second-order indirect reciprocity—that is, the degree to which the reputational consequences of an action depend on the reputation of the action’s target—was more evident (indeed, exclusively evident in Chapter 5) in the context of negative cooperative dilemmas. NIR makes more precise predictions still about second order indirect reciprocity in the context negative cooperative dilemmas. It predicts that ‘exploiting good guys is bad’. We saw this in both studies. Individuals who exploited a well-reputed individual were themselves judged more negatively both within (chapter 4) and between subjects (chapter 5). Additionally,
and critically for its dynamics to be plausible, NIR predicts that ‘exploiting bad guys isn’t as bad as exploiting good guys’. Again, this is just what we observed in chapters 4 and 5. This 2nd-order reputational effect is entailed by neither negativity- nor omission-biases, it is a prediction uniquely entailed by NIR.

All up, the studies in chapters 4 and 5 support the claim that our contemporary reputation-relevant intuitions were shaped by the ancestral socioecology described in chapter 3.

That said, these findings should not be overstated. They are a necessary first step in testing NIR. They provide initial support and motivation for further investigation, rather than constituting conclusive evidence on their own. This early evidence provides justification for undertaking costlier but potentially more influentially potent work, such as attempts to directly observe NIR dynamics in contemporary small scale societies.

In chapter 4, in addition to testing these predictions, I also attempted to expand our understanding of indirect reciprocity by full surveying the space of possible second-order continuous reputation assessment rules. Previous work on this topic has typically employed the method of experimental economic games. This powerful tool allows precise control of participants’ motivations, but pays for it with ecological validity; participants interact with anonymous representations on a computer screen rather than other real, concrete people. This previous work has drawn inconsistent conclusions about second-order indirect reciprocity (caring about the target of someone’s actions, when you judge them). Some studies observe it, others don’t. Some researchers have suggested that second-order indirect reciprocity is too difficult for humans.

I suspected that people might just struggle with abstract, anonymous representations of reputation. Our reciprocity-intuitions might require a specific cognitive representation of the people involved in a situation, rich with emotions and associations. So, I made the opposite trade-off. I sacrificed anonymity and had participants judge very specific, though fictitious, characters involved in interactions that they might well encounter in real life.
I found clear, cross-cultural evidence of first order indirect reciprocity, especially in negative dilemmas. I also found reasonable evidence of second order indirect reciprocity, especially among Americans. However, when I discretely categorised each participant’s judgements into one of the 81 possible assessment rules, the most common was first-order scoring assessment—where a target’s reputation doesn’t matter. This suggests a puzzle for indirect reciprocity theorists, since our data also suggests that some people use second-order standing assessment, which models suggest is more evolutionarily viable.

Chapter 5 is particularly exciting because it presents a novel test of NIR. Driven by an interest in how the law accords with our moral intuitions (specifically the question of whether causation is necessary for guilt) Noel Inbar and his colleagues documented a peculiar psychological phenomenon. People condemn those who profit while others lose, even when there is no causal connection between them. I call this the ‘lucky profiteer effect’. Inbar’s investigations of this phenomenon made no explicit connection to NIR, nor indirect reciprocity nor reputation more broadly. Nor was NIR developed to explain this phenomenon, nor with it in mind. NIR’s ability to predict the lucky profiteer effect is therefore a genuinely novel test of the hypothesis.

Though NIR superficially fits the lucky profiteer effect—both involve negative judgements of someone who profits while others suffer—it entails an additional prediction. The effect should be ameliorated if the target of loss has a bad reputation. This additional prediction was not foreseen by the discovers of the lucky profiteer effect nor, as far as I know, is it implied or required by the moral-psychological theories on which their explanations are based. I tested the prediction and found that it held.

The patterns of variability I observed in people’s judgements of the lucky profiteer are consistent with humans’ very robust negativity- and omission-biases, supporting NIR’s inference from contemporary to ancestral cognition. However the 2nd-order reputational effect—the amelioration of these patterns when victims are badly reputed—is uniquely predicted by NIR.

The most impactful social psychological discoveries usually excite and
surprise us because they are counter-intuitive. This refinement of the lucky profiteer effect is just the opposite. It seems entirely natural to us that people would condemn those who profit on the suffering of others, but less so if their profiteering on the misfortune of ‘bad guys’. This lack of novelty is just what we would expect if an NIR socioecology had shaped our intuitions. What makes this observation exciting is not its novelty, but the fact that the theory that predicted it was derived \( a \ priori \) from theoretical considerations of early human evolution, rather than being an empirically-refined description of people’s moral behaviour.

### 6.1 Near future directions

This dissertation documents early theoretical and empirical steps towards understanding the ancestralsocio-ecological dynamics that shaped our social psychology. Much remains to be done.

The theoretical model in chapter 3 demonstrates the plausibility of a socioecology where selfish, individually adaptive, promiscuous social learning gives rise to reputations. These in turn foster a socioecology where selfish, individually adaptive exploitation actually promotes prosociality, forming the substrate of more complex forms of information sharing and collective action. However chapter 3 is just one utterance in an ongoing conversation about the circumstances that set our species in motion. It answers some questions (“could indirect reciprocity arise in a species without culturally well-defined social roles and responsibilities?”), but raises others. These include

- Could selection hone strategies to prey on the socioecology of indirect reciprocity, as it could for direct reciprocity (e.g., Boyd & Lorberbaum, 1987; van Veelen et al., 2012)?

- How does migration between groups change NIR dynamics? This is a particularly pressing question since reasonable amounts of migration could have been a prerequisite for the accumulation of complex culture (Powell et al., 2009).
• What other mechanisms could solve the cooperative dilemma of culture? Plausible candidates include cognitive mechanisms for reducing credulity (e.g., Henrich, 2009), and learning biases that create strong asymmetries in social influence (e.g., Henrich & Gil-White, 2001) such that the most influential individuals benefit by having their cooperative acts widely imitated. Do these mechanisms synergise with or disrupt NIR?

On the empirical front, chapter 4 suggests that NIR fits our reputational intuitions and chapter 5 suggests it can explain surprising novel phenomena. However both studies have limitations.

Though I strove to carefully balanced the positively and negatively valenced scenarios in chapter 4—they differed only in whether an individual added or took pieces of paper from a pile—they may still have tapped very different, culturally-established norms. It is possible that these particular norms dictated our participants’ very different responses rather than the valence of the cooperative dilemma itself.

A clear answer will not emerge until these methods are replicated with many different scenarios, each carefully designed to compare positively- and negatively-valenced variants of its cooperative dilemma.

It is also possible that the patterns observed here are the consequence of convergent cultural evolution. For instance, patterns in participants’ judgements could have been artefacts of culturally evolved nuances in how people apply moral labels such as ‘good person’ (e.g., Gidron et al., 1993; Skowronski & Carlston, 1987). The convergent evidence from my more behavioural questions (about doing people favours and returning their money) help ease this concern, but the strongest evidence would require cross-cultural behavioural studies.

I merely took the first steps towards to providing a cross-cultural grounding for these effects. I observed that Americans and Indians both showed similar trends, though the Americans’ were substantially greater in magnitude. A strong genetic interpretation of NIR predicts that these reputational intuitions should be common to all humans. We cannot confidently test this
claim by examining any two cultures, however distinct; we would need to ask similar questions of many different, culturally remote individuals.

The results I document in chapter 4 suggest that people do make 2nd-order reputational judgments, but that these might not be easily tapped by the abstract, anonymous, impersonal methods of standard experimental economic games. This opens the door for creative researchers to design new non-verbal and behavioural studies of peoples’ reputational judgments that do not require anonymity. That is, the challenge is now to test indirect reciprocity, including negative indirect reciprocity, by experimentally manipulating what people’s real reputational reactions are to a situation, rather than merely asking them to imagine and report what they would do.

The profiteer effect examined in chapter 5 is also a first step. While it is exciting that NIR accurately predicts novel phenomena, the real test of a distal theory is whether it fits many different phenomena across many different contexts and domains. While future work could certainly further document the details of the lucky profiteer effect, the best test of NIR will be whether it meets the challenge of new, very different novel tests.

6.2 Distant future directions

I open this work by raising the possibility that there might exist some central explanatory principles that provide a deep, satisfying answer the question ‘what am I?’ I hoped my work would contribute to the long road to their discovery. But how could insights into ancestral social dynamics possibly do that?

Imagine that, by formally exploring the theoretical possibilities and testing their implications for contemporary cognition, we did eventually arrive at a correct and mostly complete description of the socio-ecological dynamics that drove our ancestors along the evolutionary road to us. Why would knowing about ancient history help us understand ourselves today?

Minds are information processing machines. They use sensory information to reconstruct an external reality. The functional and mechanistic details of how they do this are being carefully and skilfully reconstructed by
cognitive and developmental psychologists. However to achieving this end we may also require an objectively correct description of the external world that our minds are grappling with. In particular, of the socio-ecological dynamics that result from minds sharing information.

The core insight of culture-gene coevolutionary theories is that our cognitive reconstructions of the outside world fundamentally changed when our ancestors began honestly transmitting complex cultural information. The reconstructive process became a collaborative endeavour, one that is much more powerful and generative, and builds a far more accurate picture than any of us could accomplish alone. Understanding the details of this process is central to understanding humans. It involves understanding the core dilemmas and social dynamics that cultural learning generates, and the ways our minds have responded and adapted to them.

Today the realities our minds reconstruct are fantastically complicated. They are peopled by entities such as ‘software’, ‘silent partners’, ‘symbiosis’, ‘sacrilege’, ‘self-esteem’ and the ‘sanguine humor’. These representations really do change how we act, interact, think, form relationships, build and sustain institutions, and so on. Much of who we are depends on how we acquire these ideas and employ them. I believe that if our models of human cognition are to ever come close to grappling with this complexity, they must intersect with accurate models of the cultural information landscape on which it occurs.

The interface between ‘models of the origin and evolution of culture’ and ‘models of psychology and cognition’ is exciting. It suggests that sociology could interweave tightly with social psychology, allowing us to build ecological models that simultaneously understand the complex interplay of individual minds and emergent social institutions and phenomena, rather than assuming that one is constant while we examine the other. It holds the prospect of new approaches to history which explicitly incorporate our growing understanding of both the universal and culturally malleable ways that minds interact with evolving cultural institutions and concepts. It suggests the possibility of a personality psychology that could describe and predict an ‘ecology of personalities’ for a given contemporary or even histori-
ical group, including their long-run frequencies, developmental trajectories and interactions between them, rather than just post-hoc cataloguing personality differences and their correlates. It is, in short, a promising space of ideas in which we might find an accurate model of human minds and the complex ecological interactions between them. It is a space in which we might answer ‘who am I?’

The bottom-up work of accurately modeling of our cognitive processes receives a great deal of funding, attention and researcher effort. Top-down inquiry into the origins and nature of the cultural socioecology our minds inhabit is comparatively scarce. I hope this dissertation contributes to this top-down effort in three ways. First, by highlighting the cooperative dilemma of culture. This will hopefully restrict the space of unproductive, implausible speculation and help focus inquiry into theoretically potent puzzles. Second, by drawing attention the importance of dilemma valence—an influential dimension of cooperation and one that might be pivotal to solving the evil teacher problem, but that has been largely ignored thus far. Third, by drawing attention to one particular socio-ecological dynamic, Negative Indirect Reciprocity, which may have been and may continue to be a driver of the cooperative foundation of our social psychology.
Bibliography


172


