Shadow of the Leviathan: the role of dominance in the evolution of costly punishment

Submitted by David Stuart Gordon to the University of Exeter

as a thesis for the degree of

Doctor of Philosophy in Psychology

In August 2014

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of degree by this or another university

Signature…….

Acknowledgements

It doesn't feel like it has been four years since I moved to Exeter to begin my PhD. Since that time I have been fortunate to be surrounded by so many fantastic people, and without their support it is unlikely I would have been able to complete this thesis.

Firstly, I wish to extend my deepest thanks to my supervisors, Stephen Lea and Joah Madden. I will always be grateful for the fact that they were willing to take me on as a student after complications with my initial arrangement. More importantly, I wish to thank them for their guidance and expertise as we developed the themes that have resulted in this thesis, and also for allowing me the space to pursue my own ideas and interests within my research. I would especially like to thank Stephen for his consistent advice, support and constructive feedback, regardless of how many times he received a draft of the same chapter, and for being a source of endless interesting facts and anecdotes. I would also like to thank Joah for providing a distinctive animal behaviour perspective on the research, and for encouraging me to write in a concise and efficient manner that is not necessarily becoming of psychology.

This thesis could not have been completed without the support of my fellow PhD students. I would like to offer a special thanks to Christina Meier, for her unwavering friendship all these years, for the trips we have taken together, and with whom I have sampled the Sunday roast of just about every restaurant and pub in the Exeter area. I would also like to offer my special thanks to Katharine Steentjes and Caroline Farmer for being the best housemates anyone could ask for. Over the years they have been a source of laughter, advice and strength, and well as wine, through the joys and disappointments of life. You both made a student house truly feel like a home.

I also offer my thanks to Megan Birney for the ranting opportunities, political discussions and mutual support through “pencil in the eye” moments, to Marco Rego for showing me the value of a lunch away from my desk, to Dale Weston for his interesting, yet incorrect, opinions of films and television shows, and to Alex Butler for being an excellent sounding board and for being single-handedly responsible for my coffee addiction. I would also like to thank the people I have met outside of the university environment since I arrived in Exeter, especially my fellow Coffee Collective veterans Chris Prosser and Chris Butterworth, and the members of my weekly quiz team; James Lee, Mike Webster, Ben Glassman and Sebastian Wilcox, the latter for our many quiz victories and all for reminding me that there is a world outside the ivory tower of academia.

I would like to thank Alex Reid and Jessica Mallach for providing much-needed changes in scenery over the years, and would also like to thank my Manchester friends in general, with special mention to my oldest friend (and keeping with the theme of this thesis, ally) Tommy Kwong, for always welcoming back their exiled friend in the South. I would like to offer a special thanks to Rebecca Jones, whose (usually dinosaur themed) gifts always arrived at the perfect moment to brighten my day.

Last, but by no means least, I would like to thank my parents and twin brother, Stephen. Without your love, support and encouragement I could not have been able to embark on this PhD, nor would I be the person I am today.

Abstract

Costly ‘altruistic’ punishment, where an individual intervenes to punish someone for behaving unfairly towards another or for violating a social norm, seems to be vital for large-scale cooperation. However, due to the costs involved, the evolution of this behaviour has remained a puzzle. The thesis initially describes why punishment is costly and explains why current theories do not sufficiently explain its evolution in the context of these costs. The thesis then offers a solution to this puzzle in the form of a dominance-based theory of the evolution of punishment. The theoretical underpinnings of this theory are discussed in reference to the previous literature, specifically how a dominant position provides sufficient heterogeneity in the cost and benefits of punishment to allow the behaviour to evolve at the individual-level of selection.

Across 10 studies, the thesis empirically investigates the role dominance is theorised to play in costly punishment behaviour. First, the judgements observers make about punishers are investigated. It is demonstrated that punishers are perceived as dominant but, unlike individuals who engage in other aggressive behaviours, punishers are also well liked. While successful punishers are judged to be of the highest rank in a social group, the wider social judgements of punishers are dependent on the attempt at punishment only; successful and unsuccessful punishers are seen as equally dominant and well liked, suggesting that the willingness to attempt punishment can honestly signal both dominance and ones pro-sociality. However, additional studies show that observers a) perceive subordinate punishers will face a great deal of retaliation, b) show surprise when subordinates attempt to punish, and c) expect that dominants will punish and be successful, whereas subordinates are expected to never punish. Thus, while there are reputational benefits from punishment, only dominant individuals can actually access them.

Second, the effect of a dominant position on punishment behaviour is investigated. Two studies sought to simulate the greater access to resources that dominants enjoy, and demonstrate that individuals who receive more resources from group-level cooperation will punish free-riding more frequently and more severely than those who receive less resources. Moreover, individuals who are in a stable dominant position, i.e. who can continually benefit to a greater degree than others from group cooperation, punish even more frequently and severely than when individuals receive additional resources alone. The results show that individuals only punish when it is cheap for them to do so and when investment in the public good (by punishing) can produce higher future returns for them. A dominant position provides the opportunity for both of these. Further studies demonstrate that individuals at the centre of a social network, an example of a 'real life' informal dominant position, are more sensitive to unfairness when making punishment decisions compared to those at the periphery of a group. However, when punishment decisions are public, and there are no economic incentives to punish, individuals behave in a similar manner regardless of social position.

Taken together, the results of the empirical studies support the proposed dominance-theory of costly punishment. The theoretical implications of the dominance-theory of punishment are discussed in reference to both the proximate occurrence of punishment and its evolutionary origins in dominance and dominant behaviours. The practical implications of this theory will also be discussed, specifically in regard to when and why individuals will act in defence of the public good. While further investigation is necessary, a dominance-theory of punishment explains both results of this thesis and the findings of the wider literature, and as such provides a coherent and compelling explanation for the evolution of costly punishment and its associated emotions.

Contents

Acknowledgements 3

Abstract 5

Contents 7

List of Tables 17

List of Figure 19

Declaration 23

1 Chapter 1: literature review 25

1.1 Cost of punishment 28

1.1.1 Effectiveness of punishment 29

1.1.2 Resources and the net-cost of punishment 32

1.1.3 Retaliation 34

1.1.4 Second-order free-riding 36

1.2 Current theoretical explanations 37

1.2.1 Indirect Reciprocity 37

1.2.2 Costly Signalling 39

1.2.3 Spite 43

1.2.4 Strong Reciprocity 48

1.3 Summary of current theories 54

2 Chapter 2: an alternative theory – Dominance 57

2.1 Defining dominance 58

2.2 Why dominance? 62

2.2.1 Dominance and punishment in non-humans 65

2.2.2 Dominance, inequality aversion and costly punishment 69

2.3 Dominance and the costs of punishment 73

2.3.1 Effectiveness and retaliation 73

2.3.2 Resources and the net cost of punishment 78

2.3.3 Direct benefits and second-order free-riding 79

2.3.4 Indirect Benefits 82

2.4 A note on sex differences 84

2.5 Testing the relationship between dominance and costly punishment 87

2.5.1 Chapter 3: Measuring punishment behaviour 89

2.5.2 Chapter 4: Perceptions of punishers 89

2.5.3 Chapter 5: Dominance rank and observer perceptions of punishers 90
2.5.4 Chapter 6: Dominance and the behaviour of punishers 90
2.5.5 Chapter 7: Dominance and behaviour - naturally occurring dominance 90
2.5.6 Chapter 8: General discussion 91
3 Chapter 3: measuring punishment behaviour 93
3.1 General introduction 93
3.2 Study 1: is punishment behaviour sensitive to reputational gains? 95
3.2.1 Reputation 95
3.2.2 Type of reputation 96
3.2.3 Personality and punishment 97
3.2.4 The current study 98
3.3 Method 99
3.3.1 Participants 99
3.3.2 Materials and procedure 99
3.3.3 Group stability and Audience to punishment 99
3.3.4 Personality Measure: The Dirty Dozen 101
3.3.5 Manipulation check and demographic questions 101
3.4 Results 1: behaviour of participants 102
3.4.1 Punishment 102
3.4.2 Group Stability 102
3.4.3 Audience 103
3.4.4 Group Stability and Audience 104
3.4.5 The Dark Triad 104
3.5 Results 2: perception of a punishing other 104
3.5.1 Punishment and Outrage 104
3.5.2 Audience, Group Stability and the Dark Triad 105
3.6 Discussion 105
3.6.1 Group Stability 105
3.6.2 Audience 108
3.6.3 Attitude to punishers 110
3.6.4 Dark Triad 112
3.6.5 Limitations 113
3.6.6 Conclusion 114

3.7 Study 2: actions speak louder than words: the response to deceptive and non-deceptive signalling of punishment behaviour 115
3.7.1 Honest signalling 116
3.7.2 Conventional signalling and retaliation 117
3.7.3 Personality and punishment 118
3.7.4 The Current Study 119
3.8 Method 120
3.8.1 Participants 120
3.8.2 Materials and procedure 120
3.8.3 Signalling and punishment scenario 120
3.8.4 The Trait Dominance-Submissiveness Scale (TDS) 122
3.8.5 The Dirty Dozen 122
3.8.6 Comprehension and manipulation check questions 123
3.9 Results 123
3.9.1 Reaction to Charlie 123
3.9.2 Attitude to Alex 124
3.9.3 Emotional response to Alex (The Ekman emotions) 125
3.9.4 Punishment of Alex 126
3.9.5 Dominance 128
3.10 Discussion 129
3.10.1 Signal of punishment and participant behaviour 130
3.10.2 Response to honest or deceptive signalling 130
3.10.3 Punishment of a deceptive signaller 132
3.10.4 Trait-Dominance 133
3.10.5 Limitations 134
3.10.6 Conclusion 135
3.11 General discussion 136
3.11.1 People like punishers 137
3.11.2 Dominance equals more likely to punish 138
3.11.3 Abandoning the Dark Triad 139
3.11.4 General conclusion 139
4 Chapter 4: perceptions of costly punishers 141
4.1 General Introduction 141
4.1.1 Reputation and costly punishment 142

4.2 Study 3: both loved and feared: costly punishment is perceived differently from other agonistic behaviour 144
4.3 Method 145
4.3.1 Participants 145
4.3.2 Materials and procedure 145
4.3.3 Experimental Scenario 146
4.3.4 Likability and dominance questions 146
4.4 Results 147
4.4.1 Likeability 147
4.4.2 Perceived dominance (male participants only) 148
4.5 Discussion 148
4.6 Study 4: perceptions of a third-party are affected by their attempt at punishment and not its success 150
4.7 Method 151
4.7.1 Participants and materials 151
4.7.2 Experimental Scenario 151
4.7.3 Likability and dominance 152
4.7.4 Manipulation checks and demographic questions 152
4.8 Results 152
4.8.1 Relative dominance rank of the third party 153
4.8.2 Perceived dominance of the third party 154
4.8.3 Likability of the third party 154
4.8.4 Judgements of the third party and the threat posed by the aggressor 154
4.9 Discussion 154
4.10 General Discussion 157
4.10.1 Signalling dominance 157
4.10.2 Likeability 159
4.10.3 Indirect benefits of costly punishment 161
4.10.4 General conclusion 163
5 Chapter 5: dominance rank and observer perceptions of costly punishers 165
5.1 General introduction 165
5.1.1 Dominance and costly punishment 167
5.1.2 Dominance and the cost of punishment 168
5.1.3 The current studies 170
5.1.4 Operationalising dominance 171

5.2 Study 5: can only dominant individuals enforce a credible threat of punishment? 171
5.3 Method 172
5.3.1 Participants & Materials 172
5.3.2 Experimental Scenarios 173
5.3.3 Social perception questions 173
5.3.4 Manipulation checks and demographic questions 174
5.4 Results 174
5.4.1 Credible threat of punishment 174
5.4.2 Perceived dominance and likability 175
5.4.3 Success, likeability and retaliation 176
5.5 Discussion 176
5.6 Study 6: Dominance rank, outcome, and observer perceptions of costly punishers 178
5.7 Method 179
5.7.1 Participants 179
5.7.2 Materials and procedure 179
5.7.3 Experimental vignettes 180
5.7.4 Social perception questions 180
5.7.5 Manipulation checks and demographic questions 181
5.8 Results 181
5.8.1 Outcome 181
5.8.2 Likability 182
5.8.3 Dominance 183
5.8.4 Retaliation 184
5.9 Discussion 186
5.10 General Discussion 188
5.10.1 Dominance and the origins of costly punishment 189
5.10.2 Intervention or punishment? 190
5.10.3 General conclusion 192
6 Chapter 6: dominance and the behaviour of costly punishers 193
6.1 General introduction 193
6.1.1 Dominance and the cost of punishment 194
6.1.2 Dominance and the direct benefit of punishment 195
6.2 Study 7: additional benefit from group cooperation increases punishment behaviour 198
6.3 Method 200

6.3.1 Participants 200
6.3.2 Experimental design 200
6.3.3 Procedure 202
6.3.4 Statistical Analysis 202
6.4 Results 203
6.4.1 Punishment severity 203
6.4.2 Punishment frequency 205
6.4.3 Contributions (phases 1 & 2) 205
6.4.4 Phase 2 Contributions 208
6.4.5 Phase 1 contribution data 210
6.5 Discussion 210
6.5.1 Punishment 210
6.5.2 Cooperation 214
6.5.3 Conclusion 217
6.6 Study 8: Private gain and the public good - monopolisation of group resources by punishers' increases spending on punishment 218
6.7 Method 220
6.7.1 Participants 220
6.7.2 Experimental design 220
6.7.3 Benefit mechanisms 221
6.7.4 Stability 222
6.7.5 Procedure 223
6.7.6 Statistical analysis 223
6.8 Results 224
6.8.1 Punishment severity 224
6.8.2 Punishment frequency 227
6.8.3 Contributions 229
6.8.4 Punisher behaviour 230
6.9 Results 2: group level effects 232
6.9.1 Group efficiency 233
6.9.2 Group punisher earnings 233
6.9.3 Mean non-punisher earnings 236
6.10 Discussion 236
6.10.1 Punishment 236

6.10.2 Cooperation 239
6.10.3 Group efficiency and group-member earnings 240
6.10.4 Conclusion 242
6.11 General discussion 242
6.11.1 Dominance and costly punishment 242
6.11.2 A single (ineffective) punisher 244
6.11.3 Beyond dominance: a case of Noblesse Oblige? 246
6.11.4 General conclusion 249
7 Chapter 7: dominance and behaviour 2 - naturally occurring dominance 251
7.1 General introduction 251
7.1.1 Cooperation and dominance 251
7.1.2 Punishment and social status 253
7.2 Study 9: cooperation and punishment in an informal social network 254
7.3 Method 256
7.3.1 Participants and research context 256
7.3.2 Design 257
7.3.3 Procedure 257
7.3.4 Matching & Payment Procedure 258
7.3.5 Generating the social network 259
7.3.6 The Trait Dominance-Submissiveness Scale (TDS) 260
7.3.7 Statistical analysis 260
7.4 Results 261
7.4.1 Cooperation and social status 261
7.4.2 Punishment and social status 263
7.5 Discussion 266
7.5.1 Cooperation and social status 266
7.5.2 Punishment and network position 269
7.5.3 Conclusion 270
7.6 Study 10: reputation, cooperation and punishment in an informal social network 271
7.7 Method 273
7.7.1 Participants and research context 273
7.7.2 Design 273
7.7.3 Procedure 273
7.7.4 Matching & Payment Procedure 275

7.7.5 Generating the social network 275
7.7.6 Statistical analysis 276
7.8 Results 276
7.8.1 Cooperation and social position 276
7.8.2 Punishment and social position 276
7.9 Discussion 277
7.9.1 Cooperation and social status 277
7.9.2 Punishment and social status 279
7.9.3 Conclusion 281
7.10 General discussion 281
7.10.1 The selfish punisher: dominance and the motivation to punish 282
7.10.2 Measuring dominance 284
7.10.3 Future directions 285
7.10.4 General conclusion 286
8 General Discussion 287
8.1 Research question 287
8.1.1 Dominance and proximate punishment behaviour 287
8.1.2 Dominance and the origins of costly punishment 290
8.2 Practical implications: costly punishment, dominance and leadership 294
8.3 Future directions 296
8.3.1 Dominance, punishment and retaliation 297
8.3.2 Punishment and usefulness: a test of prestige 298
8.3.3 Dominance, punishment and ally retention 299
8.3.4 Punishment and social network position 300
8.3.5 Perceptions and expectations of dominance and punishment 301
8.3.6 Punishment and dominance in a virtual environment 302
8.4 Shadow of the Leviathan: dominance and the evolution of costly punishment 304
9 References 307
10 Appendix: vignettes and instructions 325
10.1 Appendix A: chapter 3 325
10.1.1 Study 1 scenario 325
10.1.2 Study 2 scenario 326
10.2 Appendix B: chapter 4 327
10.2.1 Study 3 scenario 327

10.2.2 Study 4 scenario 331
10.3 Appendix C: chapter 5 332
10.3.1 Study 5 scenario 332
10.3.2 Study 6 scenario 333
10.4 Appendix D: chapter 6 334
10.4.1 Study 7 participant instructions and comprehension questions 334
10.4.2 Study 8: participant instructions and comprehension questions 338
10.5 Appendix E 342
10.5.1 Social network questionnaire – Study 9 & 10 used identical questionnaires 342
10.5.2 Study 9 participant instructions 343
10.5.3 Study 10 participant instructions 346

List of Tables

Table 3.1: the relationship between punishment and the emotional response to a defection..... 103

Table 3.2: summary of most common participant responses to Alex's behaviour..... 127

Table 3.3: model summary 129

Table 6.1: mean contributions and spending on punishment across conditions. 204

Table 6.2: model summaries..... 207

Table 6.3: mean of group-level contributions and spending on punishment across conditions..... 226

Table 6.4: model summaries..... 228

Table 6.5: group level analysis - model summaries..... 235

Table 7.1: the relationship between trait dominance and social network position 261

Table 7.2: the relationship between social network position and Proposer offers under different punishment effectiveness. 262

Table 7.3: correlation between social position and Proposer offer under effective or ineffective punishment..... 264

Table 7.4: the relationship between network position and third party punishment behaviour under effective and ineffective punishment conditions..... 265

List of Figure

Figure 3.1: distribution of the percentage participants wished the defector's marks to be reduced. ..103

Figure 3.2: percentage of participants who punished above (grey) or below (blank) the median amount of punishment.....103

Figure 3.3: relationship between the severity of the punishment demanded by participants and their attitude to a punisher.....105

Figure 3.4: relationship between participant's outrage at a defection and their attitude to a punisher. 105

Figure 3.5: how supported participants felt in response to the signal of a conspecific.....124

Figure 3.6: attitude to Alex when he did (blank) or did not (grey) assist in punishing Charlie in relation to the former's initial signal.....124

Figure 3.7: emotional reaction to Alex whether he punished (blank) or did not punish (grey). Bars = 1 standard error.....125

Figure 3.8: participant's surprise that Alex did (blank) or did not (grey) punish in relation to his initial signal.....126

Figure 4.1: likeability of John depending on how formidable he was described as being. Bars = 1 Standard Error.....148

Figure 4.2: likability of John across different antagonistic encounters. Bars = 1 Standard Error. 148

Figure 4.3: perceived dominance of John across different antagonistic encounters. Bars = 1 Standard Error.....149

Figure 4.4: proportion of participants who, across conditions, ranked the Third Party, the Aggressor and the Victim as the most dominant character (Black bars), gave the character the middle rank (grey bars) and as the least dominant character (white bars). 153

Figure 4.5: proportion of participants who perceived the Third Party (white) or the Aggressor (grey) to be the most dominant character in each condition.....153

Figure 4.6: judgement of likeability (White) and dominance (grey) of the Third Party depending on the Third Party's response to an act of aggression. Bars = 1 Standard Error.....155

Figure 4.7: participants' judgements of the Third party between the Successful Punishment (white) and Increased Threat (grey) conditions. Bars= 1 Standard Error. 155
Figure 5.1: proportion of participants who believed the intervention by an a) dominant or b) subordinate punisher would be successful (grey) or unsuccessful (white). 174
Figure 5.2: participants' reaction to the intervention for a dominant (white) or subordinate (grey) Third Party. Bars = 1 Standard Error. 175
Figure 5.3: participants' perception a Third Party's likability and dominance when they engaged in Aggressive (white) or Non-aggressive (grey) punishment. Bars = 1 Standard Error. 175
Figure 5.4: predicted outcome of third party punishment depending on a) the rank of the third party or b) the rank of the aggressor (White = successful intervention, grey =unsuccessful intervention, black= no intervention). 182
Figure 5.5: predicted outcome of third party punishment depending on the rank of the third party and the aggressor (White = successful intervention, grey =unsuccessful intervention, black= no intervention). 182
Figure 5.6: the perceived dominance of the third party depending on the rank of the third party and the aggressor (dominant = white, subordinate = grey). Bars = 1 Standard Error. 183
Figure 5.7: the perceived risk of retaliation against a successful or unsuccessful intervention depending on the rank of the third party and the aggressor (dominant aggressor = white, subordinate aggressor = grey). Bars = 1 Standard Error. 185
Figure 6.1: punishment severity across punisher bonus conditions. Bars = 1 Standard Error. 205
Figure 6.2: percentage of punishment opportunities taken by punishers. Grey=punishment occurred, blank=no punishment occurred. 205
Figure 6.3: mean contributions by participants to the public pot in Phase 1 (blank) and Phase 2 (grey). Bars = 1 Standard Error. 206
Figure 6.4: mean contributions over time in Phase 1 (without punishment - dashed lines) and Phase 2 (with punishment - solid lines). 208

Figure 6.5: contributions over time in Phase 1 (no punishment) and Phase 2 (with punishment) by bonus conditions: diamond=0% Bonus, squares=10% Bonus, triangles = 25% bonus, crosses = 50% bonus. 209

Figure 6.6: severity of punishment across the different types of benefit available to punishers, when groups were random (blank) or fixed (grey). Bars = 1 Standard Error. 225

Figure 6.7: percentage of punishment opportunities taken by punishers across benefits conditions when groups were random or fixed. Grey =punishment occurred, Blank = punishment did not occur. 229

Figure 6.8: mean contributions by punishers across benefit conditions, when groups were Random (blank) or fixed (grey). Bars = 1 Standard Error. 230

Figure 6.9: contributions by punishers over time when groups were random (a) or fixed (b): Circles = no-benefit condition, crosses = bonus condition, triangles = monopoly condition. 231

Figure 6.10: difference between punisher contributions and mean of other group-members between benefit conditions. A negative value suggests the punisher contributed more than the mean of non-punishers. Bars = 1 Standard Error. 232

Figure 6.11: group-level effects when groups were random (blank) or fixed (grey), where a) shows overall group efficiency, b) shows earnings by those selected as punishers, and c) shows earnings by non-punishers. Bars = 1 Standard Error. 237

Figure 7.1: relationship between 'Socialise on Trip' InDegree network position and Proposer offers when punishment was effective (Circles, dashed line) or ineffective (Crosses, solid line). 263

Figure 7.2: relationship between 'Socialise at home' InDegree network position and Proposer offers when punishment was effective (Circles, dashed line) or ineffective (Crosses, solid line). 263

Figure 7.3: relationship between regression slope for punishment spending in response to Proposer offers and influence InDegree network position. 264

Declaration

The research reported in this thesis was carried out at the University of Exeter between October 2010 and August 2014 and was supervised by Prof. Stephen E. G. Lea and Dr Joah R. Madden.

This dissertation has not been submitted, in whole or in part, for any other degree, diploma or qualification at any university. Studies 3, 4, and 5 have been written in an attenuated form as a single scientific manuscript that has been submitted to a scientific journal. Study 8 has also been written as a scientific manuscript and submitted to a scientific journal. I have designed all studies (in collaboration with my co-authors: Madden and Lea) and collected the data for all studies. I wrote the first and final drafts of all manuscripts and prepared the figures and tables. My co-authors have edited the manuscripts.

I would like to thank my intern, Adam Dunt, for his help with the data collection for Study 5, and Dr Jess Isden for her help with the data collection for Studies 9 & 10. While I designed the economic games used in Studies 7 & 8, I would like to thank Glyn Prichard for writing the computer code for these experiments.

David S. Gordon

Exeter, July 2014

1 Chapter 1: literature review

It is a (very) common sight on the motorways of the UK to see signs warning that the lane ahead is closed for maintenance and that drivers should change into an inner one. Most comply after the first few signs, but there are always some individuals who drive right to up the actual 'closed' barriers, passing the slower queuing traffic as they go, then expect to be allowed to change lanes at this last moment. However, this wish is rarely granted. Other drivers will go to great lengths, often risking a collision, to ensure such defectors remain stuck while the traffic they so arrogantly tried to get ahead of steadily rolls past. As any motorist will attest, there is an undeniable sense of joy that comes from passing such a scene.

The above is just a small example of what is referred to as 'moralistic' or 'costly' punishment, where an individual or individuals punish a violation of an established social norm in defence of the public good (as above), or punish an act of unfairness committed against an unrelated other. Importantly, such punishment is greatly beneficial to a group and to society at large, as it deters harmful social defections and encourages pro-social behaviour such as cooperation. In fact, while there are a number of mechanisms that can enforce pro-social behaviour between a small number of individuals, for example kin selection Hamilton (1964), reciprocal altruism (Trivers, 1971), or reputation building (Bird & Smith, 2005a), no mechanism has been as effective at encouraging cooperation as punishment for non-cooperation (Balliet, Mulder, & Van Lange, 2011): if "*covenants, without the sword, are but words and of no strength to secure a man at all*" (Hobbes, 1651/1996, Chapter 17, para 2), then punishment can be seen as just such a sword.

Furthermore, an environment of *"mutual coercion, mutually agreed upon"* (Hardin, 1968, p. 163) is also something that we actively want. A novel example appeared in 1993, when a rape occurred… in cyberspace [sic], in a text-based precursor of a modern online-game, when a player 'took control' of the avatars of several other players (Dibbel, 1999, Chapter 1). While

prior to this incident the community had been ardently against any form of control or codes of conduct, and although the act itself amounted in real terms to little more than a few lines of text, the outrage was palpable. Suddenly, when faced with a defector, there were calls for the ability to delete, exclude or otherwise remove players who violated hastily generated rules. More scientifically, it has been consistently experimentally demonstrated that participants become outraged at perceived unfair or uncooperative behaviour (Falk, Fehr, & Fischbacher, 2005; Trivers, 1971), and actively prefer environments where punishment is possible (migrating to them, Gürerk, Irlenbusch, & Rockenbach, 2006; or voting for them, Noussair & Tan, 2011). Outrage at acts of unfairness and anti-social behaviour, and the desire to punish those who behave in such a manner, seems to be a uniquely human trait (not shared by are nearest extant relatives, Chimpanzees, Jensen, Call, & Tomasello, 2013; Riedl, Jensen, Call, & Tomasello, 2012), and one that is likely responsibility for our highly cooperative behaviour.

However, it remains controversial how a tendency to perform such punishment could have evolved as it is costly to the punisher but is beneficial to the group as a whole. The apparent group-beneficial but individual-deleterious nature of punishment seems to put the behaviour at odds with prevailing evolutionary theory (as presented, for example, by Dawkins, 1976). Because of this, there is great disagreement as to how costly punishment, and the associated moral outrage, could evolve.

This thesis aims to offer a possible solution to this puzzle.

First however it will be important to explain in more detail why the evolution of punishment is such a puzzle. This chapter will discuss what the proximate costs to punishment are, and how the current theories have attempted to explain the evolution of punishment in light of these costs. The following chapter (Chapter 2) will then set out an alternative explanation for

the evolution of punishment, with the rest of the thesis devoted to empirically testing this explanation.

The nomenclature of punishment has fluctuated over the years, between ‘altruistic’ (for example, Fehr & Gächter, 2002), ‘costly’ (for example, Rockenbach & Milinski, 2006) or simply ‘third party’ (for example, Fehr, 2004) punishment. While the latter two are nominally interchangeable, ‘altruistic’ has fallen out of favour because the behaviour is no longer regarded as being so in the Hamilton (1964b) sense (Barclay, personal communication). However, on a proximate and mechanistic level, a distinction can made between situations where a 'disinterested' third party intervenes in an unfair dyadic conflict (see, Fehr, 2004; Pedersen, Kurzban, & McCullough, 2013) and when punishment is directed against defector from the public good (Fehr & Gächter, 2000). Strictly speaking, only the former can be referred to as ‘third party punishment’ while the latter has been referred to as ‘altruistic’, ‘moralistic’ or ‘costly’, and more recently as ‘second-party functional punishment’ (Jensen, 2010, p. 2639). This is because while the punisher in the latter was not directly harmed by the defection, they suffered because a defection lowers the group product: punishers in such public good situations are therefore not disinterested. It should be noted though that in the anthropological (see Mathew & Boyd, 2011), and the non-human animal literature (for example, Raihani, Grutter, & Bshary, 2010), punishment of defections against the group has been referred to as ‘third party punishment’ despite the punishers not being disinterested. Thus even the above distinction is not absolute.

Nevertheless, research on both punishment by disinterested third parties and ‘altruistic’ group members is interpreted as suggesting that humans have a unique desire to punish those who behave ‘unfairly’ (see Fehr & Gächter, 2002; Marlowe et al., 2008). The interpretation of data from both situations is comparable, if not identical, in terms of what it says about human behaviour. Thus, one is tempted to argue that ‘third party punishment’ can adequately cover

any instance of punishment where a punishing individual chooses to intervene, be it as a disinterested bystander or someone willing to confront a defection 'for the good of the group'.

However, in a recent critique of the field, Guala (2012) referred to punishment in both disinterested and intra-group situations under the umbrella term 'costly punishment'. Therefore, in this thesis, costly punishment will be used throughout to describe the act of punishing an anti-social, unfair, or otherwise norm-violating behaviour. However 'altruistic' will occasionally be used to express the idea that punishment, in the specific context under review, is implied to be truly altruistic in nature. 'Third party' will also occasionally be used when referring specifically to a situation where the punisher is deemed to be disinterested. Finally, costly punishment will occasionally be shortened to 'punishment', and this will always mean costly punishment; other forms of punishment (for example second party punishment) will always be labelled with their full title.

1.1 Cost of punishment

Individuals react very negatively to free-riding and unfair behaviour, and this reaction is a strong predictor of punishment (Falk et al., 2005; Fehr & Gächter, 2002; O'Gorman, Wilson, & Miller, 2005). Crockett, Clark, Lieberman, Tabibnia, and Robbins (2010) showed that the punishment of defectors is an immediate and impulsive act, and is both pleasurable to witness (Singer et al., 2006) and to do oneself (de Quervain et al., 2004). There have been a number of suggestions as to why such costly punishment occurs, and these have centred around the idea that humans are averse to inequality (Fehr & Schmidt, 1999; Leibbrandt & López-Pérez, 2011), or show an other-regarding preference (Camerer & Fehr, 2006) that compels us to punish those who behave in an anti-social manner towards others, or in a way that negatively affects the group as a whole.

It should be noted here that 'unfair', 'defector', 'free-rider' or 'norm violator' all refer to an individual who has behaved in an anti-social manner, and will be used interchangeably as appropriate. For example someone who does not contribute to the public good is punished for free-riding in a public goods game (for a review of public goods games, see, Casari, 2005), whereas in a third party punishment game (see, Fehr, 2004), punishment is directed at those who have made an unfair/unequal resource division.

Nevertheless, despite the 'moral outrage' (Trivers, 1971) felt at unfair behaviour, costly punishment, as one may expect from the name, is strongly affected by the proximate cost to the punisher, i.e. the amount of resources they must spend to punish (McCullough, Kurzban, & Tabak, 2013), with the majority of individuals punishing only when the cost is low. As shown by Dreber, Rand, Fudenberg, and Nowak (2008), the costs of punishment presents a problem for any theory attempting to explain the evolution of punishment, as the costs provide a strong selection pressure against the behaviour. The sections below represent a description of the proximate costs of punishment, and focuses on the costs only; how the evolution of punishment can be explained in the light of these costs will be discussed in Section 1.2.

1.1.1 Effectiveness of punishment

In 2000, an Israeli day-care centre, tired of parents being consistently late in picking up their children, decided to impose a small fine on parents who were late by more than 10 minutes. However rather than reducing lateness, the fine dramatically increased it. Whereas lateness had previously been seen as violating a social norm as it inconvenienced the staff, now that inconvenience was seen as a service, and one that busy parents were happy to pay for (Gneezy & Rustichini, 2000a). This nicely demonstrates that in order for punishment to successfully deter free-riding and defection, it must inflict sufficient costs upon the target as to make these behaviours untenable (for example, Andreoni, 1988; Gächter, Herrmann, &

Thoni, 2005; Gardner & West, 2004a; Price, Cosmides, & Tooby, 2002). Changing the behaviour of free-riders is certainly a motive behind punishment (Masclet, 2003) and explains why there is more punishment when groups are fixed and when there are additional resources at stake (Abbink, Brandts, Herrmann, & Orzen, 2010; Fehr & Gächter, 2000); as shown by Shinada, Yamagishi, and Ohmura (2004) if one must continually interact with a defector, it pays to alter their behaviour.

Ever since the earliest work in the area of punishment (see Ostrom, Walker, & Gardner, 1992; Yamagishi, 1988), experimenters have ensured that punishment inflicts sufficient costs on the target by making it 'effective'; that is, making the ratio between the resources spent on punishment and the damage inflicted on the target sufficiently large. This has the added effect of making it very cheap for individuals to punish. In their seminal work, Fehr and Gächter (2000) used a punishment system that slowly escalated in effectiveness (from 1 point spent by the punisher removing 1 of the targets points, to 10 points spent by the punisher removing 30 from the target), but most subsequent studies have used a fixed ratio of 1:3 (for a review see, Balliet et al., 2011; see also, Dawes, Fowler, Johnson, McElreath, & Smirnov, 2007; Masclet & Villeval, 2008). In their comparative analysis of different punishment ratios, Nikiforakis and Normann (2008) demonstrated a 1:3 ratio to be the optimum for encouraging punishment (and therefore cooperation) insomuch as above a 1:3 ratio (e.g. 1:4) cooperation is not significantly greater and the cost to the social product, i.e. the amount of resources destroyed by punishment, is excessive. This ratio has become standard for both public goods games and other punishment-based study designs (for example, Fehr, 2004; Pedersen, Kurzban, & McCullough, 2013).

In fact, the cooperation enhancing effect of effective punishment is maintained even when there is heterogeneity in the ability to punish (de Weerd & Verbrugge, 2011); for example if only a single group-member can punish effectively (Nikiforakis, Normann, & Wallace, 2009)

or at all (O'Gorman, Henrich, & Van Vugt, 2009). As long as there is some possibility of being punished cheaply and effectively, individuals will be more cooperative. This is either because free-riders fear being punished (Andreoni, 1988) or, as suggested by Fischbacher, Gächter, and Fehr (2001), because the knowledge that potential defectors will be deterred encourages others to cooperate as it lessens the fear of being taken advantage of. Equally, it is important to note that most, if not all, models of third party or costly punishment depend on effective punishment for the for behaviour to be evolutionarily stable (Boyd & Richerson, 1992; de Weerd & Verbrugge, 2011; Gardner & West, 2004a; Roberts, 2013; and many others). This is the case even when other possible mechanisms for the evolution of punishment are being investigated, for example reputation (Santos, Rankin, & Wedekind, 2011) or group-level selection (Gintis, 2000).

A question that arises therefore is how and why individuals could punish effectively outside the laboratory. One solution is ostracism. For any group-living animal, ostracism from a group tends to result in death (Cant, Hodge, Bell, Gilchrist, & Nichols, 2010; Wilson, 1980, p. 142) and Bowles and Gintis (2004) demonstrated that ostracism, while very costly to the target, can be considered cost-free to the punisher. As shown by Masclet (2003), even *if* ostracism is costly it can still enforce cooperation in an experiment, although the effects are not as strong as those seen when an actual monetary cost is inflicted on the target. Still, perhaps because of the threat from ostracism, Ostrom et al. (1992) and Masclet, Noussair, Tucker, and Villeval (2003) found that individuals do respond to verbal admonishments. Such an effect also occurs for other non-monetary sanctions; Barr (2001) found that shame, and the potential for shaming, act as punishment (see also, Jaffe, 2008), and others have found a similar effect for the threat of being gossiped about (Bazzan & Dahmen, 2010; Sommerfeld, Krambeck, Semmann, & Milinski, 2007).

However, there is a darker side to effective punishment; it encourages a great deal of anti-social and counter-punishment/retaliation. While the latter will be discussed in more detail below (1.1.3), anti-social punishment, where individuals punish high co-operators, seems especially to go against the common suggestion that costly punishment is an act of public good (Fehr & Fischbacher, 2003; Fehr & Gächter, 2002). Anti-social punishment regularly occurs in economic games (Barclay, 2006; Ostrom et al., 1992; Ottone, 2008), has been observed cross-culturally (Herrmann, Thoni, & Gächter, 2008), and can curtail the evolution of punishment in theoretical models (Dreber & Rand, 2012; Rand, Armao, Nakamaru, & Ohtsuki, 2010). Because of the frequency of anti-social punishment it has been suggested that a great deal of punishment might have a spiteful motive behind it (see 1.2.3). Interestingly however, Falk et al. (2005) have found that when the cost of punishment to the punisher rises, anti-social punishment all but disappears; only cooperative individuals are willing to punish at great cost to themselves (see also, Dawes et al., 2007; Egas & Riedl, 2008; Nikiforakis & Normann, 2008; Sigmund, 2007).

1.1.2 Resources and the net-cost of punishment

Effective punishment is not the only way in which the cost of punishment can be reduced. An alternative mechanism is for individual punishers to have higher overall resources available to them, so that while the absolute cost of punishment remains the same, the net cost would be lower for these individuals (de Weerd & Verbrugge, 2011; Frank, 1996). In fact, de Weerd and Verbrugge (2011) suggested that the cost to the punisher might be a more important factor in explaining the evolution of costly punishment than the effect on the target.

Although the effect of heterogeneity in resources has been little studied in relation to punishment, it has been investigated in relation to contributions to the public good. Results are mixed. Chan, Mestelman, Moir, and Muller (1999) found that participants with greater resources contribute more to the public good, but Buckley and Croson (2006) found the

opposite (for a review, see Ostrom, 2006). These studies did not include a punishment mechanism, but when punishment is possible the results are equally mixed. Generating resource heterogeneity by providing certain participants with higher initial endowments in a public good game, Burns & Visser (2006) found that participants who received lower endowments contributed more to the public good, whereas Reuben & Riedl (2013) found that those who received higher endowments contributed more. The same is also true for studies that provided different marginal returns from group cooperation; Reuben & Riedl (2013) and Tan (2008) found no difference in contributions between high and low earners, while Nikos Nikiforakis, Noussair, & Wilkening, (2012) and Reuben & Riedl (2009) found that high earners contribute more, but only when punishment was not possible.

Of the studies above, only Tan (2008) found that participants with more resources punished to a greater degree than those with fewer. The other studies did not report any differences in punishment behaviour between the different levels of resources held by participants. Resource level did not affect punishment most likely because punishment was effective, at least a 1:3 ratio (Reuben & Riedl, 2009) and up to a 1:5 ratio (Burns & Visser, 2006) being used; in the case of Nikiforakis et al. (2012) a small initial price allowed an unrestricted amount of punishment. In the Tan (2008) study, punishment was relatively ineffective (1:2). Therefore, in the majority of previous studies, punishment was still cheap, even for participants who received fewer resources; indeed, in the case of Burns and Visser (2006), low earning participants proved very willing to spend their resources to reduce the income of higher earners regardless of the latter's contributions to the public good.

Finally, an additional mechanism for lowering the net cost of punishment would be to distribute the cost amongst many individuals. Ostrom et al. (1992) found that when coordination was possible prior to beginning the experiment, there was no actual need for punishment as no participants defected. Equally, while all the research previously discussed

employed a peer-sanctioning mechanism whereby all members of a group can punish (for a review see, Casari, 2005), Traulsen, Röhl, and Milinski (2012) employed a pool-punishment system, whereby all participants could contribute to an additional punishment pool. When a threshold of resources was reached the pool would punish low contributors to the public good. Despite being more costly (if no-one free-rode, the contributed resources were lost) participants actually preferred the pool-punishment system, essentially a system where punishment was coordinated by a central authority. Finally, a model by Boyd, Gintis, and Bowles (2010) found that as long as a) potential punishers could signal to one another and b) punishment only occurred when a certain threshold of signallers was reach, a small initial number of punishers could lead to punishment being evolutionary stable. However, an issue raised by Boyd et al. (2010) is that such a mechanism may allow an individual to trigger punishment but not take part, and Peterson (2011) found that moral outrage can indeed be used in such a way. The emergence of coordination signals might therefore be problematic, as when there is a possibility for deception, 'cheap talk' is best ignored (Duffy & Feltovich, 2002).

1.1.3 Retaliation

Despite the relative paucity of work conducted on it, perhaps the greatest cost to punishment is from retaliation (Dreber & Rand, 2012). Also known as counter-punishment, retaliation occurs when an individual who has received punishment responds in kind. Simply put, individuals do not happily accept being punished, regardless of whether they deserved it due to their actions. To use an example from Janssen and Bushman (2008), there is a reason the police wear body armour when arresting criminals. Individuals do show a great desire to retaliate when punished (Falk et al., 2005) and even early studies into costly punishment reported that individuals seems to punish in response to having been punished themselves previously (Fehr & Gächter, 2000; Ostrom et al., 1992). Indeed Gächter et al. (2005)

suggested that much 'anti-social' punishment, where high contributors are punished, can be seen as defectors pre-empting the punishment for their own behaviour, essentially retaliating first. When allowed to actually engage in retaliation, i.e. given the ability to selectively target 'their' punishers, Nikiforakis (2008) found that individuals retaliate to such a degree that punishment no longer occurs. Cinyabuguma, Page, and Putterman (2006) suggested that retaliation could be beneficial, as it would allow *altruistic* individuals to punish anti-social punishers, but in practice it was primarily used to exact revenge for being punished in a previous round. Furthermore, when retaliation has been introduced into evolutionary models, it prevents the evolution of cooperation and costly punishment, as the latter is now *too* costly (Janssen & Bushman, 2008; Sigmund, 2007; Wolff, 2012).

Individuals seem to take the effect of retaliation into account when investing in punishment. In the laboratory, when retaliation is possible, participants either will not punish at all (Nikiforakis, 2008), or, as shown by Rockenbach and Milinski (2011), will punish defectors and then actively conceal their punishment. The desire to punish unfairness still exists, but retaliation poses too much of a cost for it to be translated into action. Furthermore, in non-economic studies, the effects of retaliation are also apparent. Jenson and Peterson (2011) found participants reported being less willing to confront a formidable defector, and Kawakami, Dunn, Karmali, and Dovidio (2009) found that in their confederate-based study, participants were unwilling to confront anyone for violating social norms. In fact, the lack of retaliation might be why laboratory experiments have been accused of hugely overestimating the willingness of participants to punish defectors (Guala, 2012), which draws into questions some of the theories derived from such methods.

Outside the laboratory, individuals will also actively hide punishment behaviour; Acheson (1988) described how fisherman would secretly cut the lines of others they believed to be violating fish quotas (see also, Ostrom, 1990), and the main reason crimes are not reported to

the police, an otherwise 'cost free' action, is the threat of retaliation (Miller, 2010; Tarling & Morris, 2010). Even when faced with a costly defector, individuals in pre-state societies are still very unwilling to engage in punishment; Mathew and Boyd (2011) for example describe the long process by which any decision to punish in one such society is reached, and this is not surprising given how ubiquitous retaliation and counter-retaliation is within such societies (Chagnon, 1988; Diamond, 2012; Hill, Barton, & Hurtado, 2009). According to Hill et al. (2009) the prevalence of retaliation, coupled with the lack of any formal sanctioning institutions, might be why there is very little evidence of costly punishment in non-state societies (see also, Marlowe et al., 2008).

Retaliation is a ubiquitous part of human behaviour, and as demonstrated by feuds (Nikiforakis et al., 2012; Zizzo & Oswald, 2001), one act of confrontation is always seen as deserving another. Given the human, and indeed non-human primate (Kazem & Aureli, 2005), propensity to take revenge for any perceived slight (Felson, 1982; Marlowe et al., 2010; and for a review, see McCullough et al., 2013), even the 'cost free' social punishments mentioned in 1.1.1 are potentially much more costly when retaliation is considered. As shown by Levine, Taylor, and Best (2011), even third parties attempting to reconcile belligerents rather than punish them are not immune from attack, and it is a sad fact that have-a-go-hero stories in the press end with 'fatally wounded' as often as they do 'chased away the perpetrator' (for example, intervention by armed bystanders, Branas, Richmond, Culhane, Ten Have, & Wiebe, 2009; Goodman, 2014). Fundamentally, it appears that humans do not like being punished regardless of whether we 'deserved it', so any attempt to understand the evolution of costly punishment behaviour must take into account the cost of retaliation.

1.1.4 Second-order free-riding

The final cost of punishment is in essence a result of all of the previously discussed costs. While an individual might be able to punish cheaply because their punishment is effective,

because they have relatively greater resource-levels than others, and have managed to avoid retaliation from the target of punishment, an individual who has spent any resources on punishment will lose out to second-order free-riders (Dreber et al., 2008; Kiyonari & Barclay, 2008). First identified as a potential problem by Yamagishi (1988), second-order free-riders are cooperative in that they contribute to the public good, but they do not invest in costly punishment. Therefore while all individuals benefit from cooperation, second-order free-riders do so more than the punishers because they have not paid the cost of punishment. As a result, over evolutionary time punishers are outcompeted and defectors can re-emerge (Helbing, Szolnoki, Perc, & Szabó, 2010) and while punishment of second-order free-riders has been suggested, this would lead to an infinite-order free-rider problem and yet more costs for the punisher (Sigmund, 2007).

Thus, any theory attempting to explain the evolution of costly punishment should take into account the costs above; how individuals as a whole or an individual singular can punish effectively, whether they have a lower net-cost of punishment, how punishers can avoid retaliation, and how any cost of punishment can be recovered in the face of second-order free-riding. As acknowledged in the opening section, the above represent the proximate costs of punishment only. The question for any theory, however, is how, in the face of such costs, the willingness to punish unfairness has evolved in humans.

1.2 Current theoretical explanations

1.2.1 Indirect Reciprocity

One solution to the costs of punishment is to assume that they cannot be directly overcome, and instead that any cost will be overcome through indirect reciprocity, whereby an actor's behaviour leads to a change in the future behaviour of conspecifics towards the actor, such as greater inclusion in cooperative activities or an increase in altruism directed towards the actor

(Nowak & Sigmund, 2005). While, as described by Johnstone and Bshary (2004), there is evidence that many animals take note of the behaviours of conspecifics, due to our greater social cognition humans are especially adept at image-scoring, i.e. tracking the behaviour of others we observe (Melis & Semmann, 2010; Nowak & Sigmund, 1998, 2005). From these 'scores', a reputation emerges. Indirect Reciprocity theory therefore proposes that those with a reputation as a punisher receive special benefits from others in the group.

The effects of reputation, and the human interest in maintaining our reputations, are well documented. Bateson, Nettle, and Roberts (2006) demonstrated that we are so sensitive to being observed that even subtle cues it is occurring, such as exposed to eye-like images, is enough to trigger pro-social behaviours (see also, Ernest-Jones, Nettle, & Bateson, 2010; Haley & Fessler, 2005). Our sensitivity to being watched is great enough that Levitt and List (2007) and Hilbe and Sigmund (2010) have questioned whether any decision taken in a lab experiment can be perceived as truly anonymous by participants. Furthermore, such is the effect of reputation that Kiyonari and Barclay (2008), and Rand, Ohtsuki, and Nowak (2009), have argued that indirect reciprocity from cooperative or altruistic acts alone can potentially support cooperation without punishment. However many others disagree (Gürerk et al., 2006; Herrmann & Gächter, 2009; Rockenbach & Milinski, 2006; Sigmund, Hauert, & Nowak, 2001), arguing that rewarding individuals for cooperation must take place continuously, whereas little actual punishment can be needed to ensure cooperation. For example, Ostrom et al. (1992) and others have found that after the first few rounds of a public goods game, no punishment is actually needed as no one defects.

A number of models have suggested that if punishers gain indirectly from their actions, specifically if they receive greater levels of altruism or cooperative offers from observers, then punishment can be evolutionarily stable (Frank, 2003; Gardner & West, 2004a; Panchanathan & Boyd, 2004; Santos et al., 2011), and Tennie (2012) suggested that we do

image-score for actual punishment behaviour. Indeed, Barclay (2006) and Fessler and Haley (2003) found that punishers are treated more altruistically than non-punishers and a number of studies have found that punishers themselves are sensitive to the presence of an audience (Bering, 2008; Kurzban, Descioli, & O'brien, 2007; but see Rockenbach & Milinski, 2011), which suggests that some sort of reputational gain is expected.

However, Indirect Reciprocity as a general theory does not make any assumptions about what sort of reputation is being gained per se, only that certain behaviours by an actor could lead to subsequent changes in the behaviour of eavesdropping conspecifics. Yet reputation is only useful so long as it allows individuals to make future predictions about the behaviour of conspecifics. So one question is, what does engaging in costly punishment predict about future behaviour of a punisher that warrants a positive change in behaviour towards them? The next two sections will discuss two theories that might provide an answer.

1.2.2 Costly Signalling

Costly Signalling Theory posits that individuals engage in immediately costly behaviours to honestly signal visually hidden qualities, such as genetic fitness. Perhaps the best known example is the handicap principle (Zahavi & Zahavi, 1997), according to which organisms produce ornaments that negatively impact their survival, but Costly Signalling can also be seen in 'show off' behaviour (Hawkes, 1991), i.e. purposely energetically and/or materially wasteful behaviour (Bird & Smith, 2005b; Bird, Smith, & Bird, 2001; Iredale, Van Vugt, & Dunbar, 2008) or behaviour that is especially risky (Farthing, 2005). Finally, costly signalling also includes conventional signalling (Maynard-Smith, Harper, & Brookfield, 1988), where the behaviour or ornament is itself cost-free, but the honesty of the signal is ensured by the cost-imposing behaviour it provokes in conspecifics.

Barclay (2006) and Nelissen (2008) have suggested that costly punishment might be a signal of one's pro-sociality. Generally speaking, individuals who punish are more cooperative than non-punishers (Barclay, 2006; Falk et al., 2005) and Egas and Riedl (2008) found this association to be especially apparent when the cost of punishment was high. Punishers are trusted more than non-punishers (Fessler & Haley, 2003), are seen as more group focused and 'nice' (Barclay, 2006), and are preferred choices for social partners (Farthing, 2005). Importantly, Nelissen (2008) demonstrated that the indirect rewards of punishment do seem to correlate positively with the actual cost of the behaviour, as one would expect from a costly signal. Thus, by signalling their trustworthiness by punishing, punishers can reap the indirect benefits of such sentiment in future dyadic interactions (Albert, Guth, Kirchler, & Maciejovsky, 2007; Berg, Dickhaut, & McCabe, 1995; Gächter, Herrmann, & Thoni, 2004; Rotter, 1980).

Furthermore by punishing an act of unfairness, punishers are (potentially) sacrificing the option to engage in selfish behaviour themselves, and this can be seen as further cost to punishment. While individuals react negatively to defections in dyadic interactions (for example, Fehr, 2004), there is seems to be a special disdain reserved for hypocrisy (Kurzban, 2012; Ohtsubo, Masuda, Watanabe, & Masuchi, 2010), as can be seen by the outrage that greets any authority figure caught violating even a relatively minor norm (for example, avoiding driving points, Laville, 2012). It appears that by 'moralistically' punishing certain behaviours, an individual is signalling they will not undertake such behaviours themselves (Peterson, 2011). Punishers might therefore be preferred social partners not because they themselves are trustworthy per se (indeed cooperative non-punishers are more well liked than punishers: Barclay, 2006), but because they have demonstrated a commitment to *public* fairness.

As shown by Fischbacher et al. (2001), cooperation is conditional: we want to cooperate but fear others will defect and, by punishing, punishers provide an environment whereby cooperation can take place because the threat of defection is lower. Despite the threat of anti-social punishment, Rockenbach and Milinski (2006) found that participants are very willing to migrate to environments where punishment is possible. Furthermore, individuals are more willing to condemn social norm violations when punishment is possible (Mulder, Verboon, & De Cremer, 2009) and, interestingly, Kim, Smith, and Brigham (1998) found the mere presence of a concerned third party makes individuals more willing to challenge unfairness themselves. As a result, punishers might find it easier than others to recruit social allies, something that would be vital given the role coalitions have likely played in human evolution (Gavrilets, Duenez-Guzman, & Vose, 2008; Pietraszewski, Cosmides, & Tooby, 2014), and suggests that punishment might be related to wider group dynamics rather than just signalling an individual punisher's 'niceness'.

A model by Gintis, Smith, and Bowles (2001) suggested that while punishment could be evolutionarily stable as a costly signal, what it signals need not be pro-social tendencies or result in group-beneficial behaviours (see, Boyd & Richerson, 1992). Engaging in costly punishment might send a very different signal, that the punisher themselves is formidable and not to be treated unfairly in future interactions. For example, Brandt, Hauert, and Sigmund (2003) found that participants were less likely to cheat a punisher out of fear of retribution, and Barclay (2006) also suggested that the increase in pro-social behaviour directed to punishers might be a due to this fear rather than positive regard. In fact, an alternative explanation for the paucity of costly punishment in non-state societies offered by Marlowe et al. (2008) is that, in such small communities, a reputation for formidability and for not allowing oneself to be treated unfairly can be established easily through eavesdropping on dyadic encounters. Costly punishment can instead therefore be seen as another form of

aggressive behaviour designed to signal the ferocity of the actor (Griskevicius et al., 2009), with any benefits being the result of that fear.

These two variants of a Costly Signalling theory approach to punishment are not mutually exclusive. It is perfectly possible that costly punishment could signal both formidability and that an individual is trustworthy and 'nice'. Sell, Tooby, and Cosmides (2009) suggested that when making welfare decisions about individuals, both their ability to inflict cost (formidability) and their potential usefulness is taken into account, and Petersen, Sell, Tooby, and Cosmides (2012) found this to be the case when making criminal-justice punishment decisions. Indeed, given the prevalence of inter-group conflicts in human societies (Chagnon, 1988; Keeley, 1996; Stanish & Levine, 2011) and in our evolutionary history (Choi & Bowles, 2007; Gavrilets et al., 2008; Mitani, Watts, & Amsler, 2010; Wilson & Wrangham, 2003), a formidable individual would be a very useful ally to have. However, Benard (2013) and Melis and Semmann (2010) found that we dislike aggression and aggressive individuals in general, and Hawley, Little, and Card (2008) suggested that we will not associate with such an individual unless they are also skilful and likable. While formidable males are preferred as sexual partners (Farthing, 2007; Penton-Voak & Perrett, 2000), Graziano, Jensen-Campbell, Todd, and Finch (1997) found this only to be the case if these individuals are also seen as likable. As proposed by Silk (2003), aggression has its uses but makes later pro-social associations difficult. Therefore costly punishment might be an effective way to signal formidability without the negative consequences associated with aggressive behaviour, but this this has not been empirically tested (a test of this is described in Chapter 4).

Thus Costly Signalling theory suggests that engaging in punishment acts as an honest signal, either of an individual's trustworthiness and honesty or as someone who is formidable and not to be cheated in the future. In this account, the costs of punishment are an integral part of

the behaviour rather than something to be overcome. However, the benefits from costly signalling via punishment as described above do not have to be specific to costly punishment; as stated by Barclay (2006), as long as conspecifics in our evolutionary past *did* respond positively to those who punish unfairness, then there was selection pressure in favour of individuals who punished an act of defection or unfairness. However, Costly Signalling theory does not explain *why* we would be adverse to unfairness in the first place, i.e. *why* we implicitly believe that punishment of defectors is a good thing and *why* we are angered by defections (Falk et al., 2005) and enjoy seeing defectors punished (de Quervain et al., 2004; Singer et al., 2006), especially when our closest extant relatives have no such compulsion (Jensen et al., 2013; Riedl et al., 2012).

1.2.3 Spite

An alternative theory for the evolution of costly punishment that may answer these questions has been suggested in the form of spite. Spite can be defined simply as the willingness of an individuals to harm another at their own immediate expense (Gardner & West, 2004b; Jensen, 2010; West, Griffin, & Gardner, 2007), which fits nicely with the proximate realities of costly punishment. Importantly, a spiteful explanation for punishment suggests a potential evolutionary origin of our apparent aversion to inequality (Fehr & Schmidt, 1999; Levine, 1998), which Costly Signalling theory cannot explain.

1.2.3.1 Spite and inequality aversion

Much work on spite has been conducted in the light of the ultimatum game (Camerer, 2003). Here, one player, the proposer, can split a stake any way they choose and send a share to another participant, the responder. The responder can either refuse or accept the decision, with the refusal resulting in a 'zero' score for both participants. While studies differ in their specifics, for example comparing the sex of participants (Eckel & Grossman, 2001), their personality (Osumi & Ohira, 2010) or the heritability of behaviour (Wallace, Cesarini,

Lichtenstein, & Johannesson, 2007), proposers consistently send almost equal splits of any stake, and participants consistently reject low offers (Henrich et al., 2005; but see, Lamba & Mace, 2013). While the rejection of any offer above zero is economically irrational in the immediate term, as long as others are watching it is best to reject low offers, lest one receive more unequal offers in the future (Nowak, Page, & Sigmund, 2000; Skyrms, 1996). In fact, as shown by Rand, Tarnita, Ohtsuki, and Nowak (2013), from a proposer/dictator point of view, all offers should be equal unless an individual has a reputation for accepting unequal offers. Individuals do seem sensitive to such outcomes as, for example, participants showed a very negative reaction to unfair offers when they believed they were playing with a fellow human, but not when their co-player was believed to be a computer (Van't Wout, Kahn, Sanfey, & Aleman, 2006).

Therefore, it has been argued that the recognition of *fairness* is beneficial for group-living animals as it allows an individual to prevent future exploitation (Brosnan, 2011), and as a result punitive responses to inequality, regardless of the immediate costs, would be under positive selection pressure (Johnstone & Bshary, 2004). Still, it should be noted here that while such spiteful behaviour might get an individual a reputation for formidability, its primary aim is to ensure that the initial unfair actor does not benefit from their actions; the emotional response to such actions appears to be due to the pleasure of punishment, rather than injustice at the offer (Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003). Thus evolutionary pressure is acting as much on individuals to behave fairly lest they lose everything to a spiteful response, and evolution has acted to coordinate both 'proposer' and 'responder' behaviours (Rand et al., 2013; Skyrms, 1996).

To put it in another more general way, if the actor loses out to others by some means (for example an unfair distribution of resources) it is in their long-term interest to lower the fitness gradient of their population through an immediate spiteful response (Gardner & West,

2004b). Nevertheless it has been suggested that such spiteful rejection of unfairness can account for the evolution of cooperation and punishment only if there is sufficient reputation building/indirect reciprocity (Gardner & West, 2004b; Johnstone & Bshary, 2004; Levine, 1998; Marlowe et al., 2010). However, in comparison to Costly Signalling, the reputation formed from Spite can be considered inadvertent. The actor is punishing to level the fitness gradient, and conspecifics are image-scoring the actor through eavesdropping (see Johnstone, 2001), whereas in the case of Costly Signalling, punishment is *intended* for observing conspecifics; as modelled by Rand et al. (2013) Spiteful behaviour is concerned with harming the target, it is simply also advantageous to know who will behave spitefully.

1.2.3.2 Envy and costly punishment

Spiteful behaviour therefore may have led to inequality aversion (Levine, 1998), as it was beneficial for individuals to respond to unequal allocation of resources and to ensure they themselves avoided such spiteful responses. Punishment might therefore be the result of a general aversion to inequality being extend out to interactions between conspecifics (Brosnan, 2011), resulting in what appears to be *altruistic* punishment (Fehr & Gächter, 2000; Fehr & Schmidt, 1999).

However, the motivation to punish might not be an aversion to inequality per se, but to *disadvantageous* inequality, where we ourselves are worse off, in which case it is properly referred to as *envy*. While I am not suggesting envy to be the same as spite per se, it can potentially be a sub-category of it. Certainly the patterns of behaviour associated with envy suggest equivalent motivations, i.e. the desire to reduce the fitness of others. For example, in public goods games, high earners are the primary targets of punishment regardless of their behaviour (Burns & Visser, 2006; Zizzo & Oswald, 2001). Also, Leibbrandt & López-Pérez, (2008, 2011), using a third party punishment game, found that punishment occurs even if the target made the most egalitarian decision possible. That resource disparity, not the 'fairness'

of behaviour, is the motivation for punishment was further demonstrated by Pedersen et al. (2013), who found that punishment was almost entirely motivated by envy of resources rather than actual anger. In fact, even bystanders, individuals who had no say in resources distribution, are punished if they are 'better off' than the punisher (Leibbrandt & López-Pérez, 2008).

Interestingly, it has been suggested that disadvantageous inequality aversion might explain the flat dominance hierarchies of non-state societies (see, Boehm, 1997). As argued by Gavrilets (2012), dominant individuals can monopolise resources and behave in a coercive manner, and so it is within an individual's best interest to notice this and spitefully attack better off individuals even if the punisher is not affected by the target's behaviour. Spiteful motivation behind inequality aversion may explain why 'moral outrage' is greater when the defector is formidable (Jenson & Peterson, 2011) or when the outraged individual is weak (Peterson, 2012); those unable to behave in an unfair or unequal way want to prevent others from doing so.

Costly punishment might therefore be grounded in this envious/spiteful sentiment. Indeed, while other primates, specifically chimpanzees, do not demonstrate any concern for others (Riedl et al., 2012), they do show negative responses to being disadvantaged themselves (Brosnan, Talbot, Ahlgren, Lambeth, & Schapiro, 2010). Thus, according to a spiteful account of the evolution of punishment, the initial aversion to inequality evolved to recognise exploitation in dyadic interactions (Johnstone & Bshary, 2004; Nowak et al., 2000) and as human social and coalitional psychology became more complex (Gavrilets, 2012; Gavrilets et al., 2008), spiteful punishment evolved as a more complex social tool to ensure others could not become too powerful. Indeed, as shown by Van De Ven, Zeelenberg, and Pieters (2010), the fear of being envied alone can promote generous behaviour by better-off individuals. That such targeted envious punishment, or the threat thereof, results in a reduced disparity in

resources between individuals that benefits the group as a whole (see Frank, 2003) is entirely accidental.

It should be noted that a great deal of spiteful punishment occurs even when the punisher is in an advantageous position (Falk et al., 2005; Fehr, 2004), and this is contrary to the idea that punishment is motivated by envy. However, this may partly be due to methodological difference in the different studies described. The studies that specifically demonstrate that punishment is motivated by envy make use of third party punishment games (for example, Leibbrandt & López-Pérez, 2008, 2011; Pedersen et al., 2013), where the punisher is an disinterested observer to the allocation of resources, and the simple mechanism (observing a split of resources) makes calculating any disadvantage easy. Studies that show general anti-social/spiteful behaviour are public goods games (as mentioned in 1.1.1, the majority of public goods game show spiteful behaviour) where any calculation is more complex. This complexity (coupled with a time-limit in most experiments) may explain why in the latter games individuals often appear to altruistically target free-riders. Public goods games generate a cooperative dilemmas due to the benefit of free-riding, thus even if identifying individuals who *actually* are better off than us might not be possible because, for example, there is a lack of full information (for example, Kamei & Putterman, 2012), it is possible that free-riders *probably* are better off. Indeed, as shown by Masclet and Villeval (2008), participants punishment of free-riders increases positively with the deviation of the target's contribution from their own. Envy suggests we punish 'successful' individuals rather than 'unfair' individuals; it is simply that, in public goods dilemmas, the two positively correlate.

Ultimatum-like games might show an inequality aversion because a balance must be reached in terms of the best mutual outcome when the receiver will behave spitefully (Rand et al., 2013), while punishment provides an individual with the opportunity to harm another at little cost to themselves. Thus costly punishment may be a useful general tool to maximise the

difference between oneself and other group members; the effect is just greater if one starts with those who have behaved unfairly or free-rode and are therefore the greatest distance away from the actor. Accordingly, 'envious' costly punishment can be considered a strategic application of a spiteful motivation, or, as noted by Jensen, "*Causing harm for harm's sake is a spiteful motivation, and it can be underpinned by a comparison of oneself to others"* (2010, p. 2643). This explains why most individuals punish when it is effective, as reducing disadvantageous inequality is easier, but also why some also punish when redressing unequal resource distribution is impossible (Dawes et al., 2007; Egas & Riedl, 2008).

Therefore, for a spiteful explanation for costly punishment, it does not matter whether the patterns of punishment show a specific sensitivity to disadvantageous inequality or just the desire to burn the resources of other (Zizzo & Oswald, 2001). Either motivation would explain why punishment still takes place when there is no direct chance of a reputation being formed (Fehr & Gächter, 2000) or where there is no possibility of altering the behaviour of social defectors (Fudenberg & Pathak, 2010). The primary aim of *spiteful* punishment is to lower the resources of the target rather than to compel them to cooperate. Spite does not require the target to do anything but suffer.

1.2.4 Strong Reciprocity

The theories presented above as to how costly punishment could be evolutionarily stable have relied upon one core assumption, that the punisher survives both the initial defection and the confrontation with the defector. This need not be the case. The risk from retaliation will be discussed later (2.1); here we turn to the possibility that the defection could result in the destruction of the group itself.

Intergroup conflict is as much as part of non-state human societies as it is of the modern world (Chagnon, 1988; Keeley, 1996; Stanish & Levine, 2011) and likely played a major role

in our evolution (Choi & Bowles, 2007; Gavrilets & Fortunato, 2014; Lehmann & Feldman, 2008). Indeed, Manson et al. (1991) have shown that Chimpanzees not only show territorial defence, but engage in behaviours such as raids into rival group's territory that have been likened to human warfare (see also, Mitani et al., 2010; Wilson & Wrangham, 2003). Such behaviour therefore clearly has a long evolutionary history. Importantly, human intergroup conflicts can escalate into genocidal conflicts (Chirot & McCauley, 2010; Keeley, 1996). Thus, defecting from the public good, and especially from tasks such as territorial defence, can have lethal consequences for the group as a whole. Furthermore, even without a direct threat from a neighbouring group, extinction could result from famine or environmental disasters caused by failure to contribute to the public good (Fehr, Fischbacher, & Gächter, 2002) or overexploitation of common resources (Hardin, 1968).

Therefore, instead of trying to explain the evolution of costly punishment from the point of view of the individual, we might concentrate on how punishment affects group-level survival. This group-selectionist explanation, Strong Reciprocity (Boyd, Gintis, Bowles, & Richerson, 2003; Fehr et al., 2002; Gintis, 2000), states that individuals are willing to sacrifice their own resources to reward co-operators and those who behave fairly, and to punish those who defect from the public good or behave unfairly, without any regard for the costs to themselves. Theoretical models have shown that groups with altruistic punishers, the term used correctly in the sense that no individual benefit or cost is considered, outcompete other groups either in direct conflict or when extinction rate is high due to natural disasters or famine (Boyd et al., 2003; Denga, Gintis, & Chua, 2011). While disadvantageous for the individual, punishment can drastically increase cooperation within groups, and the resulting increase in group-efficiency allows groups with punishers to succeed in conflicts against, or otherwise outcompete, groups without them (Abbink et al., 2010; Sääksvuori, Mappes, & Puurtinen, 2011).

Nevertheless, when punishment is possible a great deal of resources are destroyed by the actions of the punishers, and this has led some to conclude that punishment does not in fact help groups compete, or at least that the differences between punishing and non-punishing groups are negligible (Dreber et al., 2008; Ohtsuki, Iwasa, & Nowak, 2009; Sääksvuori et al., 2011). However, as shown by Fehr et al. (2002), when interactions occur repeatedly over time with the same individual, which likely represents the social environment of our evolutionary past, then groups with punishers become far more efficient. Masclet and Villeval (2008) and Gächter, Renner, and Sefton (2008) found this to be especially true when groups competed against one another, but it is also the case when the risks of inter-group conflict or natural disaster are low (Bowles & Gintis, 2004). Mappes and Puurtinen (2009) found that even without punishment being possible, groups are still more cooperative when faced with inter-group conflict. In fact, models with similar restrictions have suggested that such conflicts have resulted in 'parochial altruism', whereby individuals should behave altruistically towards in-group members and aggressively towards an out-group in order to with conflicts with the latter (Choi & Bowles, 2007; García & van den Bergh, 2010). This can be considered part of Strong Reciprocity insomuch as selection pressure favoured groups whose members where unconditionally altruistic towards one another, which, according to Gavrilets and Fortunato (2014), potentially includes behaviour such as punishing free-riders.

Essentially, Strong Reciprocity posits that random mutations gave rise to strongly reciprocating individuals. Different groups have a random number of strong reciprocators, and if there were a sufficient number of such individuals in a group (Carpenter, Bowles, Gintis, & Hwang, 2009), then cooperation can be enforced because strong reciprocators reward cooperative individuals (Bowles & Gintis, 2004) or punish defectors (Boyd et al., 2010) altruistically. Because such groups are successful at avoiding extinction, either due to natural events or through inter-group conflicts, the genes for strong reciprocity are

maintained in the population even though the behaviour they engender has a negative impact on the individual. Through this group-level selection, humans show a fundamental 'other regarding preference' (Camerer & Fehr, 2006) for the welfare of others and are willing to punish acts of social defection and unfairness while ignoring the proximate costs and benefits of costly punishment.

However, as supporters of the theory say, *"Strong Reciprocity cannot be rationalized as an adaptive trait by the leading evolutionary theories of human cooperation"* (Fehr et al., 2002, p. 1). This is a bold claim to make, and not surprisingly a number of researchers have questioned this Strong Reciprocity theory. One objection from West et al. (2007) is that the punisher does gain directly from punishment as, at a population level, even if a punisher is losing out relative to their local group, relative to a global population of non-punishers their group's success still gives punishers a greater fitness. Additionally, the costs of punishment could be recovered if it ensures the defector behaves cooperatively in a group containing a number of the punisher's relatives (Lehmann & Keller, 2006). Furthermore, an alternative model of intergroup conflict by Lehmann and Feldman (2008) suggested that group-enhancing behaviour can evolve individually if the actors gain something from the activity, for example by requisitioning additional mates through raiding. Finally, according to Baumard and Liénard (2011), the apparently altruistic punishment in non-state societies (see Mathew & Boyd, 2011) may in fact occur because free-riding in this instance hurts the punisher personally, so that punishment can be considered as revenge. These criticisms cannot be levelled at, for example, Spite, because there the theory provides its own mechanism for costs to be recovered, whereas the key theoretical point to Strong Reciprocity is that human cooperation and punishment evolved as group-level behaviour and individual actions should not be beneficial to their actor.

Nevertheless, given the prevalence of inter-group conflict in our evolutionary history (and present), it is plausible that behaviours and emotions that specifically enhance group-cohesion evolved. Shame, for example, is a very useful (Barr, 2001) and potentially cost-free (Bazzan & Dahmen, 2010) method of punishment. However, Jaffe (2008) suggested that shame could evolve at an individual-level if it initially evolved to avoid or mitigate punishment and antagonism, perhaps as an extension of primate submission and reconciliation gestures (see Cheney, 2011; Silk & House, 2011). More importantly, Gavrilets et al. (2008) demonstrated that competition between individuals in forming alliances could give rise to group-beneficial behaviour. Because larger coalitions win conflicts, it is in their members' best interest to behave fairly and police unfair behaviour, lest the coalition disband. A number of recent models have, for example, shown that the ability to withdraw from cooperation reduces unfairness (the 'loner' strategy, García & Traulsen, 2012; Sigmund, 2007) and Van Vugt, Jepson, Hart, and De Cremer (2004) experimentally found that participants would abandon groups containing anti-social individuals. The result of the process described by Gavrilets et al. (2008) is a coalition that encompasses a whole group. Thus, in principle, apparently group-efficiency enhancing behaviours are the result of individuals making selfish decisions, and there is no need for an idiosyncratic human-only selection mechanism.

Finally, Strong Reciprocity makes a very specific statement about the cost of punishment; *"[it] is a willingness to sacrifice resources for rewarding fair and punishing unfair behaviour even if this is costly and provides neither present nor future material rewards for the reciprocator"* (Fehr et al., 2002, p. 3). It would be more convenient to say that the costs should not matter. However, as detailed in 1.1, individuals are very sensitive to both the cost of punishment and its potential benefits, either due to the potential for indirect reciprocity (Kurzban et al., 2007) or extra resources from group success (Abbink et al., 2010; Sääksvuori

et al., 2011). On a proximate behavioural level, the incredible flexibility humans show in punishment and cooperative behaviour as single individuals (Kurzban & Houser, 2005; Sigmund, 2007), in response to specific local factors (Lamba & Mace, 2011, 2013) and cross-culturally in general (Gächter et al., 2005; Herrmann & Gächter, 2009; Herrmann et al., 2008; Iris, Herrmann, & Zeckhauser, 2009), is problematic for Strong Reciprocity. This is also problematic for Spite as theory as, as will be discussed in 2.1, there are many situations whereby individuals a) do not obey fairness norms, but b) this unfairness is accepted.

However, Strong Reciprocity does consider an important point about human behaviour ignored by other theories: that, like our primate forbears, our survival is tied up in the survival of our fellow group members. The web of associations and alliances an individual forms with others affects our lives and behaviour (for an introduction, see Christakis & Fowler, 2010), whether that individual is a Western primary school child (Gest, Graham-Bermann, & Hartup, 2002; Sapouna et al., 2011) or an indigenous herder in the Andean mountains (Lyle & Smith, 2014). Indeed, one of the most widely accepted theories as to the origins of our species' exceptional cognitive abilities is the need to navigate and manipulate the social environment (Byrne & Whiten, 1997; Dunbar, 1998). Given the importance of intra-group coalitions, as well as inter-group conflict, in our evolution (Cummins, 2005; Guala, 2012; Pietraszewski et al., 2014), it would not be surprising if a number of behaviours and emotions had co-evolved to benefit coalition members, including shame (Jaffe, 2008), generosity (Jones & Rachlin, 2006; Silk & House, 2011) and morality (Brosnan, 2011; Flack & De Waal, 2000; Silk & House, 2011; Trivers, 1971), as well as the use of costly punishment driven, perhaps, by a concern for others being treated unfairly. In his appraisal of the economic approach to behaviour, Guala (2012) acknowledged that such preferences can be demonstrated experimentally, and experiments are most effective when teasing out such preferences and biases. Certainly costly punishment, with its associated costs, does often

seem like a group-beneficial behaviour, even if the mechanism that ultimately produced it was not group-level selection (West et al., 2007).

1.3 Summary of current theories

At the end of 1.1, it was stated that any theory claiming to explain the evolution of costly punishment must account for the proximate costs of punishment behaviour itself, how there may be heterogeneity in these costs, and provide an ultimate explanation as to why humans would have evolved the dislike of anti-social behaviour from which punishment results. The three core theories discussed above, Costly Signalling, Spite, and Strong Reciprocity all cover different aspects of punishment, but fail, in my opinion, to sufficiently explain the behaviour.

Costly Signalling provides a good account of the proximate costs behind costly punishment, as the costs are needed in order to make punishment an honest signal. As a result, it can explain some of the variation seen in punishment behaviour, for example why (in lab experiments) many will punish if it is cheap and especially why punishment is sensitive to an audience and future association. Also Costly Signalling as a general theory can account for heterogeneity in punishment behaviour when/if it occurs outside of the laboratory: those who cannot afford the signal do not engage in the behaviour. However, current punishment research based on Costly Signalling does not explain what *actual* proximate biological or social state could allow an individual to signal effectively: only a physically fit individual could free-climb a sky-scraper, only a wealthy individual could own a yacht, but what allows an individual to costly punish, or rather, what prevents everyone from doing so? Finally, while the indirect benefits from signalling might allow punishment to be evolutionarily stable (Gintis et al., 2001), as stated by Barclay (2006) signalling theory does not offer an ultimate explanation as to why observers should find the punishers of anti-social conspecifics so appealing.

Spite provides an interesting, if bleak, theoretical account for the evolution of inequality aversion based in the long-term selfish benefits of inflicting immediately severe punishment and thereby lowering the fitness of others in response to being treated unfairly. On this account, costly punishment is a method to reduce the difference in resources between the actor and the target, it just so happens that 'unfair' individuals often have more than others. With such a motive, spiteful impulses could certainly explain, for example, the willingness to punish in anonymous and unstable conditions (see, Fehr & Gächter, 2000). The proximate costs of punishment are the price to pay for 'levelling the playing field' and are recovered in the long-term through this levelling. However, while Spite may be a good explanation for why we have the *impulse* to punish, it cannot necessarily account for when punishment does *not* occur. For example, unlike Costly Signalling, Spite does not provide a theoretical explanation as to why we might find variation in punishment behaviour between individuals, i.e. heterogeneity in cost (de Weerd & Verbrugge, 2011). For a specific example, if we all punished spitefully, there would be no second-order free-rider problem.

Finally, Strong Reciprocity provides an alternative theoretical account of the evolution of punishment and its associated emotions, in the form of group-level selection pressure acting to encourage pro-social behaviours. While Strong Reciprocity has been the dominant theory since the early 2000s (Fehr et al., 2002; Fehr & Gächter, 2000), as mentioned in 1.2.4 it has recently faced some serious challenges to its validity. Moreover, the theory is based in the fundamental principle that individuals should be insensitive to the immediate costs and benefits of punishment, but a great deal of experimental evidence suggests punishment is very sensitive to these factors. Nevertheless, as discussed in 1.2.4, Strong Reciprocity is the only theory that recognises the effect punishment has on group-level behaviour, and that group-dynamics and competition might play a role in more direct benefits from engaging in punishment behaviour.

All the theories above offer possible explanations for the evolution of costly punishment. However, in my opinion, all fail to sufficiently address the three main areas that need to be covered in order for an explanation to be valid; how the costs of punishment can be recovered, how there can be heterogeneity in the cost of punishment, and why humans show a unique sensitivity to unfair or anti-social behaviour in our social environment. This thesis aims to put forward an alternative theory that could potentially explain the evolution of costly punishment. As will be described in the next chapter, and tested in the empirical chapters, I propose that dominance is the key factor in the proximate cost of punishment, and that the behaviour itself has an evolutionary origin in dominance and status contests and social reasoning

2 Chapter 2: an alternative theory – Dominance

In Chapter 1, I summarised the major current theories for the evolution of costly punishment. This thesis offers an alternative explanation. I suggest that dominance, or rather the fact that humans live in societies in which dominance relationships play a significant role, is a key and thus far overlooked factor that can explain the proximate occurrence and the ultimate evolution of costly punishment. A number of theories (Machiavellian Intelligence, Byrne & Whiten, 1997; Dominance Theory, Cummins, 1996a; Cummins, 2005; Social Brain Hypothesis, Dunbar, 1998) suggest that the need to outwit conspecifics or otherwise navigate the social environment was one of the key driving forces behind the evolution of the human mind. This section will detail how dominance, or more specifically the social cognition required to navigate human social dominance hierarchies, provides a compelling explanation as to how our concepts of fairness and inequality evolved, and why they produce the emotions that lead to punishment. Furthermore, on a more proximate level, the characteristics of a dominant social position encompass many, if not all, of the factors that have been experimentally shown to encourage punishment. As a result the characteristics of a dominant position might allow individuals to mitigate the costs that prevent the evolution of punishment behaviour. Finally, by grounding costly punishment in dominance we can potentially establish a phylogenetic link to the behaviour of non-human animals, as without such a link it is not possible to produce a valid evolutionary explanation for the presence of costly punishment in humans (Pedersen et al., 2013).

The suggestion that costly punishment might be an overtly coercive act related to dominance and to dominance hierarchies is not wholly original; it is offered as a possibility by, for example Dreber et al. (2008, p. 350) and Rand et al. (2010, p. 630). However, as the page

numbers imply, such comments have been asides at the end of articles and have not, in my knowledge, been followed up or extrapolated upon. To the best of my knowledge, the discussion below is the first attempt to formally explain the evolution of costly punishment through dominance and status contests.

2.1 Defining dominance

First, it is important to define what is exactly meant by dominance and a dominant individual, as the concept has proven difficult to define. Dominance can be used as a relative measure to express the consistent outcome of antagonistic encounters between two individuals (Hand, 1986) or as a more general description of a individuals who is at the top of a dominance hierarchy ("dominance is a trait that conveys rank", Drews, 1993, p. 297). Still, in an attempt to produce a definition of dominance, Drews (1993) maintained the focus on relative interactions *"Dominance is an attribute of the pattern of repeated, agonistic interactions between two individuals, characterized by a consistent outcome in favour of the same dyad member and a default yielding response of its opponent rather than escalation"* (p. 308).

However, it has been argued recently by Hawley (2014) that such a definition is too concerned with the form of behaviour that characterise a dominant individual, specifically in terms of aggressive interactions, as opposed to the purpose of such behaviour, to exert control over resources (Hawley, 1999; Maynard-Smith & Parker, 1976) in order to increase reproductive fitness (Ellis, 1995). Therefore a much more functional description has also been proposed, with a dominant instead defined as an individual who has *"priority of access to resources, especially reproductive resources"* (Cummins, 1996a, p. 467), or an individual who has *"preferential access to any requisite that adds to the genetic fitness of the dominant individual*" (Wilson, 1980, p. 129). Importantly, this definition not only implies that the rank of an individual is evident by their access to resources, but also what the motivation for this high rank, for being dominant, is.

The advantage of this functional definition of dominance is that it allows a greater variety of ways that dominance can be achieved, which is important when considering humans as *"the greater the size of the brain, and the more flexible the behaviour, the more numerous are the determinants of rank* [dominance] *and the more nearly equal they are in influence"* (Wilson, 1980, p. 143). Nevertheless, in human and non-human animals dominance has been seen in the light of overt acts of, or threat of, aggression (Cummins, 2005), both to gain position (Griskevicius et al., 2009) and to maintain it (Clutton-Brock & Parker, 1995; Sell, Lovaglia, Mannix, Samuelson, & Wilson, 2004; Silk, 2003). Group living non-human primates do recognise non-physical aspects of a individuals social position, for example family relations (for example, in Baboons, Bergman, Beehner, Cheney, & Seyfarth, 2003; and vervets, Cheney, 2011) and known allies (Chimpanzees, De Waal, 1982/2007), but such examples can be seen as *derived dominance* (Chapais, 1992), whereby it is the threat of aggression from an individual's associates that allows an individual to access resources. The same is also true in humans. Hobbes (1651/1996) posited that human societies and social interactions are governed by hidden threats of force, be it from the local police or a vengeful Deity (McKay, Efferson, Whitehouse, & Fehr, 2010). A good specific example was given by Dunbar, Clark, and Hurst (1995) and their study of Viking Berserkers, warriors considered terrifyingly fierce even by Viking standards: should a family that contained a berserker aggrieve another family, the latter were far less likely to retaliate for fear of the berserker.

Recently however, it has been suggested that humans have unique way to gain social status, namely prestige (Cheng, Tracy, Foulsham, Kingstone, & Henrich, 2013; Henrich & Gil-White, 2001). These papers refer to prestige and *dominance*, but what they call dominance is referred to here as 'formidability', as they use dominance as a term for purely physical domination and coercion, whereas in this thesis dominance is used to represent sustained preferential access to resources, however this is achieved. While formidable individuals rely

on fear, intimidation and coercion to achieve their differential access to resources, prestigious individuals are seen as useful, with conspecifics offering or yielding resources to them in recognition of their knowledge and expertise. Individuals give prestigious conspecifics resources because they actively *want* to, as opposed to *having* to. As argued by Henrich and Gil-White (2001), the relationship between the prestigious individual and the subordinate is a much more reflexive, with the prestigious individuals having to offer something in return, such as tutelage.

This does raise some interesting questions for the definition of dominance, as a dominant position reached via prestige is achieved by a very different mechanism than formidability. Equally, prestige might be very domain specific, whereas a dominate-formidable individual may be recognised a wide variety of environments. For example, Sell, Cosmides, et al. (2009) found the recognition of physical strength to be fast, accurate and cost-cultural, whereas recognition of an individual's prestige would depend on a certain level of cultural knowledge; although while one might not recognise *why* an individual should be differed to, it is likely even a naive visitor would recognise that an individual *is* being deferred to (Cummins, 1996a).

However, as stated by Henrich & Gil-White themselves, their analysis relates to the process by which the recognition of expertise may have evolved as a unique trait in humans, rather than its application to specific individuals as such. It is likely that for the majority of human history formidability and expertise were strongly related, for example a skilled hunter is also a formidable opponent. Direct comparisons of prestigious and formidable individuals by Cheng et al. (2013) showed that both systems are part of human dominance and rank recognition and seem to work in parallel to one another (see also, Hawley, 2014). Furthermore, the same 'dominant and submissive' gestures and behaviours occur when

interacting with someone who is formidable or of higher status (Cheng, Tracy, & Henrich, 2010; Gambacorta & Ketelaar, 2013; Gregory & Webster, 1996).

In fact, keeping dominance as an umbrella term for prestige and formidability makes it very similar to more social psychological terms such as *leverage* and *power* (Lewis, 2002) as these concepts includes formidability, expertise, and current market value and bargaining power. Leverage, for example, recognises that at any given moment an individual might have a higher market value than others; a formidable individual might seem very worthy of 'freely offered deference' when an enemy is spotted on the horizon and a prestigious individual can use their position to make direct threats against those who are less prestigious, for example a threat of preventing career advancement. Furthermore, leverage results in power (Lewis, 2002), whereby powerful individuals (and groups, Sell et al., 2004) have the freedom of action to act in their own interests (Galinsky, Gruenfeld, & Magee, 2003). Specifically, power gives an individual the "*capacity to modify others' states by providing or withholding resources or administering punishments. This capacity is the product of the actual resources and punishments the individual can deliver to others"* (Keltner, Gruenfeld, & Anderson, 2003, p. 265). This definition acknowledges both the usefulness and threat that an individual might pose, and more so than *leverage*, recombines the distinctions made by Henrich and Gil-White (2001) and others between the 'ways to the top' of the human social system.

However, like other terms such as 'social dominance' (Sidanius & Pratto, 2004), the term *power* has certain sociological connotations (Lewis, 2002; Sell et al., 2004), the discussion of which is not relevant to the topic in hand. Equally, in terms of the nomenclature, while power can be used interchangeably with dominance in the social psychological literature, it is not commonly used this way in the biological or evolutionary literature (Cheng et al., 2013, Table 1). Also dominance is the term most often applied to non-human animal social behaviour, and as such its use stresses the core hypothesis of this thesis that costly

punishment is an evolved behaviour with an origin in dominance-based conflicts and social reasoning.

While one should be hesitant to offer a new definition of a much-used concept, combining the two definitions above, so to cover the range of ways in which the concept is currently used, seems apt. The concept of 'Dominance' in this thesis is therefore defined as the recognition that *"an individual has sustained priority of access to resources, especially reproductive resources, due to their capacity to modify others' states by providing or withholding resources or administering punishments"*. This recognises that the ultimate reason organisms strive for dominance is to acquire fitness enhancing resources (Cummins, 1996a, 2005), that dominant individuals have the capacity to influence the behaviour of others (Keltner et al., 2003), and that they can do so by through both coercive and cooperative means (Hawley, 1999; Henrich & Gil-White, 2001; Von Rueden, Gurven, & Kaplan, 2008). By accounting for these attributes, the definition also acts a description of position (rank). Perhaps most importantly for a thesis on costly punishment, this definition captures the idea that dominant individuals have many mechanisms at their disposal with which to inflict costs on others.

2.2 Why dominance?

In any social hierarchy, dominant individuals have a priority of access to resources, and it has been argued that dominance hierarchies therefore represent a set of basic implicit social norms surrounding this access (Cummins, 1996a, 2005). Dominance theory suggests that the need to maintain these norms placed strong section pressure on social cognition to recognise ones place in a social hierarchy, to recognise when these rules are violated and, potentially, to punish others when violations occur. Indeed, one of the most accepted theories as to our species' cognitive abilities attributes is the need to navigate and manipulate the social environment (Byrne & Whiten, 1997; Dunbar, 1998), and primates devote a great deal of time and energy to deceiving dominant individuals in regards to activities such as feeding and

mating (Le Roux, Snyder-Mackler, Roberts, Beehner, & Bergman, 2013; Whiten & Byrne, 1988). Specifically, this view posits that when perceiving a social situation, decisions should be made taking into account social ranks of those involved and the potential ramifications of making a grab for resources and/or social position.

At this point, the difference between dominance and dominance hierarchies should be noted, as dominance relationships between individuals, the ability of one to access resources or drive away an intruder without escalation, can occur without there being a strict dominance hierarchy, i.e. consistent and sustained, and possibly linear, rank differences between members within a group (Wilson, 1980, Chapter 13). Indeed primates are perhaps unique in the complexity of the rank-based interactions and affiliations the display (for a review, see Schino, 2001). However, the idea that it is in an individual's best interest to recognise the dominance relationships between them and a competitor (Cummins, 1996a) is valid regardless of whether these relationships can be integrated into a hierarchy encompassing the entire social group. The same is true of the advantages of being dominant, i.e resource access. Thus, although dominance is often used as a shorthand for dominance ranks within a hierarchy (for example, Sapolsky, 2005), a dominance analysis does not depend on the existence of such a hierarchy. In what follows however, I assume that dominance has a least some group-wide meaning, i.e. that all members of a social group would recognise certain individuals within it as dominant.

There is a great deal of evidence that reasoning about dominance and status affects our daily lives, or to state it another way *"dominance is so intrinsic to human social relationships that we don't even notice it"* (Maestripieri, 2012, p. 8). In everyday life we adjust our behaviour in response to dominance and status, for example, by adopting the mannerisms and speech patterns of higher status individuals (Gregory & Webster, 1996), and submissive gestures are made when facing a conspecific who is either more formidable or more prestigious (Ketelaar

et al., 2012). In fact, Gambacorta and Ketelaar (2013) found that when faced with (a picture of) a formidable male competitor, participants would engage in far less creative display. The opinions of dominant individuals hold more weight (Henrich & Gil-White, 2001), and those wearing badges of status (e.g. expensive clothes) have their requests complied with more readily than subordinate individuals and are also treated more generously (Nelissen & Meijers, 2011). Such accommodating behaviour towards dominant individuals includes treating them far more leniently when they violate cooperation norms or outright commit crimes (Eckel, Fatas, & Wilson, 2010; Petersen et al., 2012; Walker, 2013).

Conversely, a dominant position also alters how individuals behave. The sense of holding a dominant position, due to the presence of allies (Fessler & Holbrook, 2013) or priming (Watkins & Jones, 2012), lowers the perception of threat in the local environment, which in itself may explain why dominance is associated with norm violations (Piff, Stancato, Côté, Mendoza-Denton, & Keltner, 2012). Perhaps more importantly, as would be expected if our social reasoning is concerned with dominance and status, a dominant position alters the perception of fairness. Dominant individuals feel entitled to more resources (Sell, Tooby, et al., 2009) and, as shown by Pratto, Tatar, and Conway-Lanz (1999), are likely to favour resource distribution based on 'merit' as opposed to equality. Practically, this results in, for example, greater rejection of lower offers in the ultimatum game by dominant individuals (Burnham, 2007) and more self-serving behaviour by such individuals (Maner & Mead, 2010; Piff et al., 2012). More generally, Nikiforakis et al. (2012) suggested that individuals select norms of fairness and cooperation in a self-serving manner, i.e. we can switch our behaviour depending on the local situation, including when we rise and fall in dominance. While not a definitive list, the above does demonstrate how even subtle cues of dominance and status affect our everyday behaviour and, more importantly, perceptions of how we

should behave towards others. This obviously includes when and why we should engage in costly punishment of anti-social individuals.

2.2.1 Dominance and punishment in non-humans

Punishment, if not always an aggressive act, is certainly a confrontational one. While the majority of experimental work on punishment has used economic methods where punishment is somewhat abstract, outside the laboratory at some point punishment must involve one individual imposing, or threatening to impose, material or social costs on a defector or social-norm violator (see, for example, Levine et al., 2011). As stated by Pedersen et al. (2013) "*A link must be established between human [punishment] and non-human animal behaviour"* (p. 7), and across taxa confrontational and/or antagonistic behaviour is strongly associated with dominance. This in itself strongly suggested that the impulse to punish originated in dominance and status contests. Thus the following section will detail how dominance is associated with aggressive behaviour and how the rule that govern the relationships between dominants and subordinates might have resulted in an aversion to inequality that has led to the evolution of costly punishment in humans.

Firstly, aggression is used to maintain a dominant position and to ensure that the dominant individual has priority of access to resources. In an influential paper Silk (2003) argued that dominant individuals should *"Practice random acts of aggression and senseless acts of intimidation"*, as doing so continually reinforces their position; by engaging in low level aggression, they reinforce a credible threat of more extreme punishment (Cant & Johnstone, 2009). Indeed, while such behaviour could be seen as spiteful (West et al., 2007), it is advantageous for dominant individuals to remind subordinates of their relative social ranks to ensure later conflicts do not escalate.

However, this does not mean that a subordinate individual should simply accept their fate, as there are of course negative outcomes for being the first to back down in a conflict; one has to sacrifice resources to a more dominant individual or rank to a successful challenger. Thus dominance in humans and non-human animals is also strongly associated with functional punishment (also referred to as second-party punishment), the response to antagonism from others, as failing to respond to aggression results in the loss of status and an increased likelihood of future acts of antagonism. As models of animal contests by Maynard-Smith and Price (1973) demonstrated, one should always retaliate if possible. For example, after losing a conflict, many primate species have been shown to engage in redirected aggression (for example vervet monkeys, Cheney & Seyfarth, 1989; and for a general overview, see Kazem & Aureli, 2005), whereby the losing individual will attack either an individual subordinate to them, or relatives/allies of the victorious conspecific, with the aim of demonstrating they are, as it were, 'down but not out'. There are clear parallels with human society; in hunter-gather societies, raids are often launched on neighbours with the expressed purposes of revenge and deterrence (Chagnon, 1988; Keeley, 1996; Mathew & Boyd, 2011), and the same motivation lies behind many instances of inner-city gang warfare (Topalli, Wright, & Fornango, 2002). More individually, Kim et al. (1998) found that, when faced with an antagonistic conspecific, participants were more likely to take revenge when the former was subordinate, but less likely when the conspecific was dominant, and Felson (1982) found that insults and challenges, in both males and females, were responded to more aggressively when there was an audience.

So far we have discussed dominance in terms of direct aggressive encounters. However, such encounters only occur to ensure one has preferential access to resources, so similar results should apply to the direct and 'fair' distribution of resources. This does seems to be the case: human participants react very negatively when receiving 'unfair' offers from humans, but not

from a computer (Van't Wout et al., 2006) and low pay can be interpreted as 'insulting' (Gneezy & Rustichini, 2000b). Interestingly, both Brosnan et al. (2010) and Riedl et al. (2012) found that while chimpanzees generally do not reject unequal offers as long as the receiving individuals get something, dominant individuals, particularly males, would accept equal offers but reject low ones, as do dominant humans in similar experiments (Burnham, 2007). Dominance at its core concerns access to resources and therefore any unequal split can be considered a challenge to dominance. Thus, it is in a group-living individual's interest to respond to both direct aggression and other acts of *unfairness* (because random attacks in the form suggested by Silk, 2003, are pretty 'unfair'), lest they appear subordinate. This includes accepting unfair offers, as accepting a 'subordinate' share will lead to future unfairness from others (Nowak et al., 2000; Rand et al., 2013).

A dominance approach therefore produces a similar explanation for the evolution of an inequality aversion as spite (1.2.1.2), in that, all things being equal, one should always stand one's ground; certainly in principle the evolutionary pressure to develop a comprehension of 'fairness' is the same. However, things are not always equal; if they were, then dominance and status would not result in such consistent differences in resources and reproduction (Ellis, 1994, 1995). Attempting to assert oneself against a stronger opponent or coalition would simply result in injury or death (Maynard-Smith & Price, 1973). As is the case with many animals, challenging a stronger opponent is unlikely to occur when there have been previous encounters or when the dominance hierarchy of a group is well understood (Johnstone & Bshary, 2004; Maynard-Smith & Parker, 1976), which is especially true for human and non-human primates (Cheney, 2011; Cummins, 1996a). As mentioned previously, human participants will back down when faced with a stronger challenger (Gambacorta & Ketelaar, 2013), acquiesce to their demands (Nelissen & Meijers, 2011) or otherwise not respond to unfair behaviour (Eckel et al., 2010; Kim et al., 1998). Such behaviour is consistent with the

reasoning of dominance; i.e. recognising what one can or cannot do and deciding whether it is worth the risk, or conversely, recognising what dominance 'entitles' one to (Cummins, 1996a).

Thus, in terms of dyadic interactions, a dominance perspective on disadvantageous inequality aversion suggests why ultimately it is important to recognise one is being treated unfairly or unequally, as such behaviours indicate rank (Brosnan, 2006). Dominance also suggests why, proximately, the circumstances under which 'unfairness' will be tolerated or responded to. Indeed, a dominant position is one in which others accept the dominant individual's priority resource access without question (Drews, 1993).

Because dominance relationships carry with them implicit norms about resource distribution and behaviours (Cummins, 1996a, 2005), it is in a dominant individual's best interest to intervene to ensure these 'norms' of resource assess are observed across the group as a whole. For example, in many social species only a single dominant pair breed, and this is enforced through harassment and eviction by dominant individuals (banded mongooses, Cant et al., 2010; and for a general overview, see Cant & Johnstone, 2009), and in cleaner fish, larger individuals will punish cheating conspecifics (Raihani, Grutter, & Bshary, 2010), lest continued cheating drive clients away (Bshary & Grutter, 2002). Indeed, in humans, dominant individuals are far more sensitive to cheating by subordinates that vice-versa (Cummins, 1999; Lammers, Stapel, & Galinsky, 2010).

As with functional punishment, dominant individuals should act to protect their position through the use of costly punishment. Wong, Buston, Munday, and Jones (2007), demonstrated that dominant members of coral-reef fish queues (where group members wait their turn to inherit a dominant position) punish group members who grow too large with eviction before they can become a threat. In a similar vein, among fallow deer Jennings,

Carlin, Hayden, and Gammell (2011) found that intervention by dominant individuals in fights between two other males effectively prevented either benefitting from a winner effect, and reinforced the dominance of the intervener. In their review of 38 primate species, Bissonnette, Franz, Schülke, and Ostner (2014) found that one of the main methods of achieving a dominant position was to form a revolutionary coalitions and overthrow the dominant group member, and there is some evidence that dominant individuals will attempt to disrupt affiliative behaviour between conspecifics to prevent the formation of such coalitions (Chimpanzees, De Waal, 1982/2007; Barbary macaques, Widdig, Streich, & Tembrock, 2000). Interestingly, similar behaviour has been observed in humans; Maner and Mead (2010) found that when faced with a talented subordinate in an experimental game, dominant individuals attempted to exclude these subordinates from group tasks by withholding key information and preventing communication. The need to watch for future social competitors may also explain why the formidability of a defector raises the anger at their actions (Jenson & Peterson, 2011), and why we are more sensitive to subordinate cheaters (Cummins, 1999; Lammers et al., 2010), as cheating may suggest subordinates are no longer accepting their position. The examples above show that dominants in both humans and non-humans will engage in forms of punishment entirely for self-motivated reasons linked to maintaining a dominant position.

2.2.2 Dominance, inequality aversion and costly punishment

The previous section argued that aggression and intervention by dominant individuals to protect their priority access to resources occurs across taxa, even in species that are not considered particularly cognitively sophisticated. Simply put, punishment is the prerogative of dominant individuals (Clutton-Brock & Parker, 1995). However, due to the social complexity of human (and some other primate) societies, the reasoning about dominance and behaviour take on a more nuanced role when multiple levels of interaction become important.

For example, A recognises that B's behaviour towards C is *unfair*, and A cares because it suggests that B has risen in rank and that D, A's ally, might face a similar challenge in the near future. This is especially important when the role of coalitions is taken into account, and non-human primates do show transitive reasoning in regards to kin and alliances (for a review, see Cummins, 1996a), and will respond when coalition members signal they are under attack (Cheney, 2011). Corvids also show transitive reasoning ability (Emery & Clayton, 2004) and have also been shown to aid 'friends' under attack (Fraser & Bugnyar, 2011), and the evolution of such behaviours in primates and corvids has been linked directly to social complexity (Fraser & Bugnyar, 2011). Nevertheless, the level of social reasoning suggested by the example above might be developed only in humans (Brosnan, 2011; see also, Melis & Semmann, 2010).

Given the role of coalitions in human evolution (Gavrilets et al., 2008; Pietraszewski et al., 2014), it would be in the interest of dominant individuals to recognise when their allies were being treated unfairly by others (for chimpanzee examples, see De Waal, 1982/2007). By punishing unfairness a dominant individual might firstly protect their social ally, or rather social investment, who may have been chosen in the first place for their rank (i.e. their market value, Barclay, 2013; Barrett & Henzi, 2006). The positive social regard towards punishers found by Barclay (2006), Farthing (2005) and Fessler and Haley (2003) may also make it easier to recruit social allies to their cause. Thus, an other-regarding preferences (Camerer & Fehr, 2006) might have evolved to selfishly maintain a coalition (Gavrilets et al., 2008).

Furthermore, the need to recruit and placate coalition members might result in pressure for dominants to behave in a positive and generous manner. For example, it should remembered that dominant individuals can behave altruistically and cooperatively (Barclay & Willer, 2007; Hawley et al., 2008; Massen, van den Berg, Spruijt, & Sterck, 2010; Roberts, 1998),

and Fiddick and Cummins (2007) found that dominant individuals are expected to let subordinates free-ride to a certain degree. This also has connections to models of reproductive skew (Powers & Lehmann, 2014; Vehrencamp, 1983) whereby dominants are expected to concede some reproduction to subordinates to keep them in the group. This could easily be extended to include extending access to non-reproductive resources and allowing greater autonomy in behaviour, i.e. "my friends are allowed to behave unfairly". Finally, this also touches on biological markets and social bargaining (Barclay, 2013; Sell, Tooby, et al., 2009), with dominants, perhaps, perceiving subordinates as a resource that needs to be attracted or otherwise attained, and competing for them: in their description of prestige for example, Henrich and Gil-White (2001) suggested dominant individuals must actively earn the deference of their subordinates.

Fundamentally therefore, the need to maintain allies raises the interesting possibility that, at some point in human evolution, dominant individuals had to behave in a pro-social manner. Such pro-social behaviour may include punishing free-riders for the 'good of the group'. The reasoning is grounded in dominance and the interests of the individual involved, but the pressure of coalitional violence caused implicit norms of *acceptable* behaviour to be extended to include non-kin associates, i.e. the desire to protect others from harm.

Interestingly, it has been suggested that once coalitions became important to human sociality it was easy to supress those who tried to become too dominant, either through direct coalition aggression (Gavrilets, 2012) or through exclusion from resources in an environment where long-term storage is impossible and foraging unpredictable (Charlton, 1997), resulting in the apparent 'egalitarian syndrome' observed in pre-state societies (Boehm, 1997). This last point is perhaps the greatest advantage of a dominance approach to punishment and its associated emotions, because by grounding punishment in an evolutionary history of dominance and status contests rather than, for example, group-selection for 'niceness', we can potentially

explain more adequately why modern human societies *aren't* egalitarian. Or rather why, during the Neolithic when the wide-spread adoption of agriculture broke the constraints of an immediate return economy that had made the suppression of dominant individuals 'easy' (Charlton, 1997), human social structure once again came to resemble those of some of our primate brethren (Betzig, 2014; Powers & Lehmann, 2014; and see Turchin, Currie, & Whiteshouse, 2013, Figure 1, for a graphical representation).

In sum, for any animal that lives in a social hierarchy, by definition dominant individuals have priority of access to resources. Theoretically, for this to be the case individuals must be able to recognise their position in relation to others in the group and what this *entitles* them to (Cummins, 2005). As argued by Silk and House (2011), a sense of inequality would evolve as such a sense would be required to recognise when one received 'less' than others, whether this was appropriate for one's rank, and how one could behave to covertly circumvent any restrictions (Byrne & Whiten, 1997). Importantly, while it might be beneficial to either take as much as possible or respond spitefully to unfairness, dominance relationships will dictate when this is possible (Clutton-Brock & Parker, 1995). It is in a dominant's best interest not only to recognise when subordinates are violating the 'rules' of resource access and behaviour (for example by establishing forbidden affiliations), but also to intervene to prevent them.

Since dominant individuals in species with relatively strict dominance hierarchies (Fallow deer, Jennings et al., 2011; Coral reef fish, Wong et al., 2007) demonstrate such 'policing', punishment to suppress subordinates might not require complex social reasoning beyond basic recognition of social position. However in humans and some other primates, coalitions are an important part of social manoeuvring (Harcourt & De Waal, 1992), and therefore the recognition of 'fair' treatment and behaviour should become extended to coalition members (Brosnan, 2011; Cummins, 1996a). Finally, and possibly uniquely to humans, costly

punishment of anti-social individuals might have evolved to have become a tool to actively recruit and maintain allies (Barclay, 2006) and suppress those who try to benefit at the expense of others (Fehr, 2004), i.e. others who try to assert dominance, resulting in an apparently *altruistic* regard for the welfare of others.

2.3 Dominance and the costs of punishment

The previous section provided a theoretical rationale as to why dominance might have given rise to the evolution of costly punishment. More importantly, if costly punishment does have an origin in dominance, the advantages and motivations associated with a dominant position should exert an effect on proximate behaviour. As may be expected, there is little actual research directly examining the role that dominance and status might play in human costly punishment but, as detailed in the following section, many of the core costs of punishment can potentially be reduced or avoided completely by a dominant position.

2.3.1 Effectiveness and retaliation

Experimental economic games consistently find that the effectiveness of punishment, that is the ratio of resources spent on punishment by the actor to the amount removed from the target, is a strong predictor of both the occurrence of punishment and of a subsequent increase in cooperative behaviour (Balliet et al., 2011). Ever since the earliest work in the area of punishment (see, Ostrom et al., 1992; Yamagishi, 1988), experiments have employed mechanisms that make sure punishment is effective, and the majority of theoretical models of punishment also rely on the cost inflicted on the target of punishment being sufficiently severe (for example, Boyd & Richerson, 1992; Roberts, 2013). As long as some individuals (Nikiforakis et al., 2009), or a single individual (Baldassarri & Grossman, 2011; O'Gorman et al., 2009) can punish effectively, then punishment both occurs and promotes cooperation. Importantly, models have shown that such heterogeneity in punishment effectiveness can allow punishment to be stable over evolutionary time (de Weerd & Verbrugge, 2011), i.e. not

everyone needs to be able to punish effectively for punishment to evolve. However, there have been limited attempts to explain either how such effectiveness could biologically manifest itself, or how and why would certain individuals be able to punish more effectively than others? Such variation in the effectiveness of punishment can be explained by an individual's dominance.

Dominants have uncontested access to resources because, as discussed previously (2.1, 2.2) dominant individuals have the capacity to alter the behaviour of others by inflicting punishment. In fact, the establishment of a dominance hierarchy results in less overall aggression because, after a few initial encounters, individuals accept that other group members can assert power over them (Wilson, 1980). Sell, Cosmides, et al. (2009) demonstrated that among humans physical formidability is easily recognised (in males), and that formidable individuals – as measured by physical strength – showed a greater history of fights, and of winning these fights (Sell, Tooby, et al., 2009). In women strength also predicted success in fights, but not a history of such conflicts. This demonstrates a perhaps obvious point that physically formidable individuals can inflict physical harm on others. Interestingly, men who are physically imposing (Watkins et al., 2010) or who have been primed to feel dominant (Watkins & Jones, 2012) are less sensitive to dominance cues in others, i.e. are less sensitive to the threat others might pose. Thus, not only are physically formidable individuals more able to inflict costs on another, they are also less sensitive to the risks of confrontations. This may explain why, when punishment by third parties occurs in everyday life, the punishers are generally formidable (Huston, Ruggiero, Conner, & Geis, 1981).

In addition, dominant individuals have more social allies to potentially provide aid in conflicts. As with group-living non-human primates (Cheney, 2011; Widdig et al., 2000), dominance in human is not simply the result of personal formidability, but of the ability to

navigate the social world (Hawley, 2014; Hawley et al., 2008). Dominant individuals are sought out as social partners (Barclay, 2013; Fessler, Tiokhin, Holbrook, Gervais, & Snyder, 2013; Hawley, 1999) and being at the centre of a group is a strong negative predictor of victimisation (de Bruyn, Cillessen, & Weisfeld, 2012; Gest et al., 2002; Smith, Talamelli, Cowie, Naylor, & Chauhan, 2004). Fessler and Holbrook (2013) demonstrated that being surrounded by friends reduces the fear of a potential adversary, and when punishment occurs in non-state societies (if it occurred at all, see Baumard & Liénard, 2011), it is only after a great number of allies have been recruited (Mathew & Boyd, 2011).

Importantly, while we may expect that social allies will assist a dominant individual in any conflict, as our analysis of dominance also suggests, dominants can also punish by denying access to resources. Being dominant, for example, means to be central in a social network (for theory and discussion, see Freeman, 1979; Scott, 2007) and as a result an individual can act as a bottleneck between group members, a position that, as shown by Maner and Mead (2010), can be used to deny subordinates access to information. Furthermore, especially if an individual is prestigious (Henrich & Gil-White, 2001), the refusal to provide assistance, training or knowledge could be a very effective punishment. In fact, this logic suggests that when dominant individuals withhold cooperation or severe dyadic social connections with another, these might be more effective punishments than when subordinates attempt them. Ostracism occurs less when it is expensive (Masclet, 2003), and a 'real-life' manifestation of this cost would be the loss of a reciprocal partner. However because dominants have a greater number of outside options for cooperation than other (Cant, 2011), they should potentially be more able and willing to withdraw support (for an example of this bargaining power, see Camerer & Fehr, 2006). Finally, costly punishment, when it occurs in non-state societies, is highly coordinated (Guala, 2012) and while in such societies no one is 'in charge', dominant

individuals have a louder voice in group decision making (Boehm, 2000; Henrich & Gil-White, 2001).

Little has been said so far about retaliation. However dominant individuals should face less risk of retaliation than others because the will of dominants is respected more than that of others. Dominants are deferred to in social encounters (Gambacorta & Ketelaar, 2013; Gregory & Webster, 1996; Nelissen & Meijers, 2011), those overtly labelled as high status are not punished for non-contributions to the public good (Eckel et al., 2010), and subordinates will simply acquiesce to being treated unfairly by them in dyadic interactions (Kim et al., 1998). Indeed, individuals who are both high status and useful are treated more leniently for any crimes (Petersen et al., 2012) and Henrich and Gil-White (2001) mention that among the Aché, men overlooked liaisons between skilled hunters and their wives. The role of dominance in punishment and retaliation can be nicely illustrated by the case of Simon Singh, who was sued by British Chiropractic Association for questioning the evidence behind their claims (Singh, 2008). While Dr Singh ultimately triumphed, without the financial and high-profile support from The Guardian newspaper (Boseley, 2009), i.e. a strong ally, an individual who 'punished' in the public good would have been crushed by the retaliation of a more powerful coalition.

To return to our analysis of dominance, dominant individuals have the "*capacity to modify others' states by providing or withholding resources or administering punishments",* and this not only allows them to punish more effectively, but also means they have a much greater freedom of action (Galinsky et al., 2003; Keltner et al., 2003), whether their actions are pro- or anti-social. Subordinates will simply acquiesce to their demands for their own safety. Such acquiescence likely also extends to retaliation, because if subordinates are unwilling to respond antagonistically to social norm violations or unfairness by dominants, they are unlikely to risk further social, monetary or physical injury by attempting to counter-punish

the same dominant individual. Indeed, when punished, the more common response of subordinates is to attack someone lower down in the social hierarchy (Barash & Lipton, 2011, Chapters 1 & 4).

I would argue therefore that when economic experiments employ an effective punishment mechanism, they are simulating a dominant position. Animal conflicts take place as a series of escalating steps (Maynard-Smith & Parker, 1976; Maynard-Smith & Price, 1973), and one of the key advantages of dominance relationships is that such escalation does not take place; subordinates simply retreat. An effective punishment mechanism allows such an asymmetrical conflict and outcome; one individual punishes the other who must accept and yield (because the game mechanism means they cannot respond in any other way). One of the methodological advantages of economic games is that innate preferences can be teased out by the way they interact with the game mechanism (Camerer, 2003; Guala, 2012), and placing participants in a position where they can punish effectively with impunity taps into evolved dominance-based instincts surrounding norm violations. As a result many individuals will punish when punishment is both effective and free from retaliation, but when the 'dominant' position is lessened, for example by the inclusion of unrestricted retaliation (Nikiforakis, 2008), individuals are less willing to punish.

Individuals may universally become angry at unfairness or acts of defection, or indeed may not and simply accept inequality, but as detailed throughout this chapter, the relative status between actors dictates the behaviour of both the 'unfair' individual and the response from the aggrieved party. In fact, one of the advantages of a dominance-based approached tocostly punishment is that it provides a theoretical account of why we are so sensitive to the costs of punishment, because these costs are indicative of dominance and status.

2.3.2 Resources and the net cost of punishment

A number of models have suggested that as long as some individuals experience a lower net cost of punishment (Frank, 1996), specifically in the production cost of the behaviour, i.e. the amount of resource that must be expended per unit of punishment (de Weerd & Verbrugge, 2011), then costly punishment can evolutionarily stable. A reduced net production cost of punishment certainly applies to dominant individuals. By definition, dominant individuals have priority access to resources: dominant individuals believe they are more deserving of resources (Sell, Tooby, et al., 2009) and can behave coercively in order to attain them (Clutton-Brock & Parker, 1995; Hawley, 1999). Indeed, even without any direct antagonistic behaviour, Nelissen and Meijers (2011) found participants were more likely to give resources to dominants, and generally subordinates are willing to tolerate asymmetries in reciprocity to maintain a close relationship with them (for humans, see Barclay, 2013; and for other primates, see Schino & Aureli, 2009). One reason why such asymmetries are tolerated is because dominant individuals also have a higher value in the biological marketplace as a potential romantic or social partner. For example males are more desired as romantic partners if they are of high status (Buss, 1989; Li, Valentine, & Patel, 2011), and both males and females who are of 'high market value' tend to make more demands of any potential partner (Pawłwski & Dunbar, 1999). Individuals who are seen as formidable are also valued as a social ally (Bassett & Moss, 2004; Farthing, 2005; Fessler et al., 2013).

Furthermore, being dominant in a group implies that one is at the centre of the social network (Freeman, 1979; Krause, Croft, & James, 2007; Scott, 2007) and individuals benefit from such a position in informal hierarchies. This central position is beneficial for several reasons. a) It means more group members are socially close and Brañas-Garza et al. (2010) found participants in a dictator game are more likely to send generous offers to individuals whom are close to them in a network, with the generosity of offers decreasing with distance (Jones

& Rachlin, 2006). b) Participants will endure more discomfort for close reciprocal friends than they will for family members (Harrison, Sciberras, & James, 2011). c) The position provides more opportunities for cooperation with others and, more negatively, allows an individual to manipulate group interactions to ensure a favourable outcome for themselves (Dasgupta, 2011; Maner & Mead, 2010).

Finally, a further way that dominance could lower the net cost of punishment is through coordination. Ostrom et al. (1992) demonstrated that allowing communication can greatly enhance cooperation (for a review of communication, see Ostrom, 2006), and group efficiency can be improved by having a single coordinator or leader (Gillet, Cartwright, & Vugt, 2010). Given that groups tend to coalesce around dominant individuals (Hawley, 1999) and that in non-state societies dominant individuals have a greater say in group decisions (Boehm, 2000; Henrich & Gil-White, 2001) it is possible that dominants could fulfil such a coordination role. One issue with the Boyd et al. (2010) model of coordinated punishment is that individuals still have to pay a cost to signal honestly; however as dominant individuals can potentially punish independently, their initial behaviour could result in coordination and therefore cheaper punishment, and Przepiorka and Diekmann (2013) demonstrated that initial differences in the cost of punishment can in fact result in such coordination.

2.3.3 Direct benefits and second-order free-riding

It could be said that the main puzzle behind the evolution of punishment is second-order free riding (Dreber et al., 2008), as no matter how cheaply an individual can punish, they will always be outcompeted by others who cooperate but don't punish. However, this conclusion is based on the premise that individuals are homogeneous in the benefits derived from punishment, and this need not be the case. Dominant individuals are in a position to benefit disproportionally from the benefits of punishment. For instance, in non-human animals, the

punishment by dominants of non-dominant breeders (Clutton-Brock & Parker, 1995), of 'cheating' conspecifics (Raihani et al., 2010), to prevent associations between conspecifics (Widdig et al., 2000), and to police growth (Wong et al., 2007) all benefit the dominant individual greatly. Even when there is no clear direct benefit, dominant individuals can still monopolise group resources through coercion and intimidation (Clutton-Brock & Parker, 1995; Hawley, 1999) or through their closer connections with others in the group.

Thus, dominant individuals can potentially benefit disproportionately from any group cooperation and the increase in the social product that it brings. While, for example, all individuals might benefit from a collective project such as building a dam or irrigation system, the increase in productivity and benefits such a project may cause will vary amongst the population (Reuben & Riedl, 2013). Because dominants benefit from group success, they have a direct strategic motivation to act to police free-riders. Additionally, dominant individuals enjoy higher reproductive success than others (Barthes, Godelle, & Raymond, 2013; Ellis, 1994; Pawłwski & Dunbar, 1999) and because of this any increase in group productivity will be multiplied via indirect fitness benefits. It also potentially gives more dominant individuals 'more to lose' when faced with, for example, inter-group conflict. Consistent with this, individuals will generally punish altruistically when faced with inter-group competition (Abbink et al., 2010) and dominant individuals will behave in a more pro-social fashion when faced with such a conflict (Maner & Mead, 2010).

A recent model by Gavrilets and Fortunato (2014) suggested that when individuals can sequester a disproportionate share of resources, they should actively contribute more to the public good. Interestingly, this model also suggested that one reason subordinates may remain in a group despite the skew in resource distribution is so they can free-ride on this public good activity (see also, Roberts, 2013). Furthermore, it seems that such behaviour is expected, with leaders being expected to punish free-riding on collective action (King,

Johnson, & Van Vugt, 2009) and it is the case that those in a better position to punish do so (Nikiforakis et al., 2009; Przepiorka & Diekmann, 2013). In fact, when only a single individuals in a group could punish effectively (Baldassarri & Grossman, 2011; Nikiforakis et al., 2009; O'Gorman et al., 2009), they did so almost exclusively altruistically, i.e. when in a tacit position of leadership, individuals behaved in a pro-social way. Potentially therefore, costly punishment might could be seen as a behaviour a dominant individual has to engage in to 'justify' their social position and continued priority of access to resources. This once again hints that costly punishment, superficially for the 'good of the group', is actually another mechanism by which to navigate complex social relationships and maintain one's dominant position.

However, few of the studies that have directly manipulated the benefit from group success have shown punishment to correlate with a higher level of benefit (Tan, 2008), partly because low earners are very willing to target higher earners (Burns & Visser, 2006; Zizzo & Oswald, 2001). However, when the punishment of high earners was forbidden, which simulates the privilege resulting from a dominant position, higher earners did punish free-riding more than low earners (Noussair & Tan, 2011). This suggests that dominants will punish for the good of the group, but perhaps only when in a secure position. For example instability in a dominance hierarchy produces more negative behaviour directed at subordinates (Fast & Chen, 2009; Georgesen & Harris, 2006), and more direct self-serving tendencies by dominants (Georgesen & Harris, 2006; Maner & Mead, 2010).

It needs to be stressed again that part of the rationale for economic experiments is that individuals bring their implicit biases and social norms into the laboratory with them and these biases are highlighted by participant responses to the game mechanisms (Guala, 2012; Levitt & List, 2007). Therefore, although it is hard to disentangle disproportional benefit from punishment effectiveness (see 1.1.2), there is some evidence that when individuals gain

a disproportionate benefit from any group activity they are more willing to punish free-riding than when this is not the case. Thus, when individuals are given the attributes of a 'dominant' position, they show the behaviour that would be expected if dominance exerts an influence on costly punishment behaviour.

2.3.4 Indirect Benefits

Dominant individuals, as shown above, may be motivated to punish because they derive a direct benefit from maintaining the public good. In addition, substantial indirect benefits are available to anyone who engages in costly punishment. As shown in 1.2.1 and 1.2.2, punishers are both well liked (Barclay, 2006; Nelissen, 2008) and seen as formidable (Brandt et al., 2003). Importantly, it has been have suggested these indirect benefits are generated because punishment can act as a costly signal of ones pro-sociality (Farthing, 2005; Fessler & Haley, 2003). Given these benefits, as raised in 1.3, the question remains as to why *everyone* would not punish if such positive reputations were available, i.e. what factor prevents someone from 'affording' the signal? This factor, I suggest, is dominance.

While there are potentially ways the production cost of punishment could be 'free', for example through condemnation (Masclet et al., 2003) or gossip (Bazzan & Dahmen, 2010), punishers will likely face retaliation for any such attempt. Yet one of the logical consequences of a dominant position is that dominant individuals do not suffer aggression for their actions, altruistic or otherwise. Thus punishment could be considered a conventional signal, i.e. a signal where the costs are generated from the reaction they provoke in conspecifics (Maynard-Smith et al., 1988). The fact that only dominant individuals can punish suggests that costly punishment could be a signal *of* dominance.

Punishment as a costly signal of dominance has important ramifications for any punishment behaviour. While any individual could attempt to punish and may even be successful in doing

so, unless they are actually in a dominant position the repercussions could be dire, especially if, as suggested by Tennie (2012), individuals are image-scored as 'punishers' in the way they are as co-operators or defectors. For example, establishing oneself as a 'protector' could make an individual the first target during any anti-social action. Furthermore, animal experiments (Molles & Vehrencamp, 2001; Tibbetts & Izzo, 2010) have shown that false signalling of dominance, as costly punishment by a subordinate would be, results in far more aggression directed at a subordinate individual that would otherwise be the case. Indeed, Anderson, Srivastava, Beer, Spataro, and Chatman (2006) demonstrated that humans are very aware of our own position in a social hierarchy and Anderson, Ames, and Gosling (2008) showed that individuals are socially punished for seeming to 'over step themselves' in social interactions. Furthermore, differences in physical strength and social power will likely mean a subordinate would *fail* to punish successfully and, beyond the direct retaliation a subordinate individual might receive, such failed attacks in primate societies often elicit further aggressive behaviour from a dominant individual for a period of time after (Clutton-Brock & Parker, 1995).

Costly punishment therefore may signal dominance in general but also, uniquely for an antagonistic behaviour, that one is dominant but (potentially) willing to forgo some of the more coercive behaviours associated with such a position, and as such is trustworthy (see, for example, Barclay, 2006). Furthermore, as discussed in 1.2.2 and 2.2.2, a reputation for costly 'moralistic' punishment might help recruit and maintain social alliances, as punishment shows that an individual is willing to police group behaviour and protect weaker members. Interestingly, a recent model by Schoenmakers, Hilbe, Blasius, and Traulsen (2014) demonstrated that if punishment is considered a costly signal of a willingness to police, punishment behaviour encourages others to commit to punishment (see also, Kim et al., 1998; Mulder et al., 2009). So by punishing, a dominant might encourage others to do so as well.

However, conventional signals are only honest when an individual cannot walk away from the environment (Számadó, 2011b). In the small scale societies of our evolutionary past, a punisher would be stuck with those they punished and were signalling to, and this is still likely the case today in terms of our workplace, friends and other formal and informal networks (Christakis & Fowler, 2010; Dunbar, 2010). Thus the signalling and subsequent reputational benefits of punishment will only be available if the individual in question can *consistently* absorb or mitigate the immediate costs of punishment. An individual should only attempt punishment if they think they can not only win, but win repeatedly.

Finally, taking a dominance perspective on costly punishment highlights an issue with previous experimental and theoretical research to punishment, that punishment is *always* successful, i.e. if a participant makes a decision to assign deduction points this will automatically punish the target. While, for example, Levati, Sutter, and Van Der Heijden (2007) and Kamei and Putterman (2012) tested the effects of imperfect *information* on punishment, to my knowledge no one has tested imperfect *success*. If we see punishment behaviour as fundamentally a dominant behaviour, behaviour that is confrontational with the intention to inflict a cost and/or enforce resource distribution norms on the target, then success is not guaranteed. To use spite as an example, while it may be ideal to respond aggressively to inequality (e.g. Nowak et al., 2000), dominance relationships may make this unwise. Indeed, while redirected aggression (for example, Aureli, Cozzolino, Cordischi, & Scucchi, 1992) could be seen as spite, it is spiteful behaviour grounded in dominance and dominance relationships.

2.4 A note on sex differences

The section that follows on from this will detail how the relationship between dominance and costly punishment was investigated. However, one factor that is not present in the

empirical chapters is an analysis of the role sex differences might play in costly punishment. This is due to both theoretical and practical reasons.

Firstly, in terms of dominance, while overtly dominant and status-seeking behaviours are stereotypically seen as male, a number of recent reviews (Cummins, 2005; Hawley, 2014; Hawley et al., 2008), have suggested that sex differences in status seeking and contests have been overestimated. As stated in 2.1, it has been suggested that definitions of dominance have been too concerned with aggressive interactions as opposed to the purpose of such behaviour, to exert control over resources (Hawley, 1999; Maynard-Smith & Parker, 1976), in order to increase reproductive fitness (Ellis, 1995). As access to resources, as opposed to competing for status as value in unto itself, is suggested to motivate females (Buss, 1989; Kwang, Crockett, Sanchez, & Swann, 2013), females should also wish to be dominant in a hierarchy. Indeed, female conflict over resources has been shown to mirror male conflict over reputation/loss of status (Griskevicius et al., 2009), females are just as willing to act antagonistically as males in same-sex confrontations (Felson, 1982) and dominant females show the same pattern of self-serving biases as dominant males (Sell, Tooby, et al., 2009). As argued by Cummins (2005), the main sex difference may be that, compared to males, females are more likely to act covertly to achieve dominance (see also, Griskevicius et al., 2009). Fundamentally, the differences lie in means by which males and females achieve dominance, not the desire for it.

Secondly, in terms of punishment, reviews of the literature (for example, Balliet et al., 2011; Guala, 2012), do not consider sex to be an important factor, and to the authors knowledge only one study that has investigated sex found an effect (Barclay, 2006); here males and females punished with equal frequency, but the former punished more severely. While physical difference between males and females, i.e. the ability to inflict physical costs and withstand physical retaliation, might in practice result in males being more likely to punish

outside of the laboratory (see, Huston et al., 1981; Levine et al., 2011), males and females are equally as likely to report anti-social behaviour (Borofsky, Stollak, & Messe, 1971): there is no sex difference in the 'moralistic' desire to punish. This may explain the lack of sex differences in economic games, as the game environment provides an implicit and unassailable dominant position from which to pour retribution on to those who violate fairness norms. Therefore, we would not expect sex differences in economic games (Chapters 6 & 7).

This is not to say sex would definitely not have an effect on costly punishment. Indeed, the role of culture, sex and expectations of behaviour regarding costly punishment could be a thesis in itself. As noted above, physical differences between males and females might bias the behaviour towards males in everyday life, even if this is entirely due to rational calculations of risk rather than differences in the desire to punish. Furthermore there are also potential cultural differences. Lowe, Levine, Best, and Heim (2012), for example, suggested that uncertainty about norms of inter-sex conflicts change the patterns of bystander intervention when two females are in conflict and a third party is male. Equally, Eisenegger, Naef, Snozzi, Heinrichs, and Fehr (2010), found that cultural beliefs about the role of testosterone in behaviour meant females who believed they received the hormone (it was a placebo) behaved more negatively than females who actually, but unknowingly, received testosterone.

Thus, while the vignette based studies were presented as gender neutral, it is likely that some cultural biases were present; specifically the characters may have been implicitly assumed to be male. Nevertheless, it is in the best interests of both males and females to acquire resources (Ellis, 1994), to recognise when they are being treated unfairly (Brosnan, 2011), to monitor the social environment and respond to changes in the social hierarchy (Cummins, 1996a, 1999), and to recognise the advantages to both associating with individuals who costly

punish and to signalling one's own pros-sociality through the behaviour (Barclay, 2006; Farthing, 2005). Therefore, while any possible cultural and biological effects of sex cannot be dismissed, the study of any effects is beyond the purview of this thesis.

Finally, there were also practical considerations. For one, to tease out the relationship between dominance and costly punishment, the studies below use multi-factorial designs to which sex would be an additional variable. Given that there is little evidence sex would affect punishment behaviour in experiments, it would not be a useful factor to include in these models. Perhaps more pressingly, as is the case in many UK universities, the psychology undergraduate cohort is predominantly female and therefore directly testing for sex effects, or performing post-hoc exploratory analyses based on sex, was not possible due to the skew in recruitment. Thus, while a sex difference cannot be ruled out, for the theoretical and practical reasons mentioned above, such sex differences were not addressed in this thesis.

2.5 Testing the relationship between dominance and costly punishment

Dominance potentially provides a theoretical explanation as to why humans have evolved an other-regarding preference (Camerer & Fehr, 2006), because of which we become angered by, and are willing to punish, those who behave unfairly, those who free-ride on group activities or otherwise act in an anti-social manner. In essence, reasoning about our place in a dominance hierarchy and how this affects entitlements and behavioural proscriptions (Cummins, 1996a) has resulted in an evolved sense of what is 'fair' (Brosnan, 2006, 2011) that goes beyond simple equality for ourselves. This is comparable with a spiteful account, which suggests that it is beneficial to avoid exploitation and harm those who attempt it (Gardner & West, 2004b; Nowak et al., 2000; Rand et al., 2013). Indeed, given that many animals that do not live in complex hierarchies show spiteful behaviour (Jensen, 2010; Johnstone & Bshary, 2004), spite could be a precursor to the formation of reasoning about 'norms' of behaviour (see, Cummins, 1996b). Because in more socially complex species a

dominant position depends on alliances and coalitions, more complex transitive reasoning is needed to recognise and punish those who 'cheat' both the dominant individual and their allies. Finally, in humans, this results in behaviour that suppresses more/too dominant individuals to *prevent* such individuals exploiting ourselves and our allies (Gavrilets, 2012).

More importantly, as argued in previous sections, a dominance-based approached costly punishment can potentially provide a theoretical account of why we are so sensitive to the proximate costs of punishment, something which is lacking in other theoretical explanations (see 1.3). Because the sense of entitlements and behavioural proscriptions are bound to dominance, so that a subordinate who felt they could openly monopolise mating would soon discover the error in their thinking, our concept of 'fairness' changes with our proximate social position, i.e. our ability to acquire, defend or hold resources (Fessler & Holbrook, 2013; Maynard-Smith & Parker, 1976; Peterson, 2012; Sell, Tooby, et al., 2009). Therefore, many of the manipulations of proximate costs, benefit and situations in experiments on costly/third party punishment can be seen as simulating the effects of dominance and status, for example the ability to punish with impunity, to establish a (dominant) reputation or ensure the success of one's own group in conflicts.

However, few studies have directly examined whether punishment is affected by dominance and status. The literature and theories discussed above suggest a number of potentially fruitful avenues for research into the role that dominance and dominance-reasoning might play in human cooperative and punishment behaviour. As a first step in such research, to investigate the basic premise and to see if dominance could be established as a credible theory, a series of 10 studies were devised to evaluate whether dominance has played a role in the evolution of costly punishment. The literature and theories presented in Chapter 1 and 2 provide the background for this research series as it was finally developed. The theory described in the present chapter was developed in reference to the results of earlier studies.

The later studies themselves were, in turn, developed in reference to earlier results and to the literature and theory that these earlier results pointed towards, i.e. there was a reflective relationship between the theoretical development and the experimental studies. Accordingly, the earliest of the studies (Chapters 3&4) were framed mainly in terms of the theories set out in Chapter 1; the later studies are more clearly tied to the ideas in the present chapter. All, however, contribute to the development of a dominance theory of costly punishment, and provide evidence in its support. The studies themselves are presented in broadly chronological order, and their progression represents the development that took place over the course of the research. A brief summary of the studies therefore follows.

2.5.1 Chapter 3: Measuring punishment behaviour

Chapter 3 investigated whether individuals took the presence of an audience into account when making punishment decisions, and whether punishers were well liked. It also investigated, specifically in response to the model by Boyd et al. (2010), whether a cheap signal of a willingness to punish would encourage participants to punish more readily. Furthermore, this chapter also discusses the usefulness of the questionnaire/survey method for investigating costly punishment.

2.5.2 Chapter 4: Perceptions of punishers

Chapter 4 investigated whether costly punishment can actually be considered a signal of dominance and pro-sociality, and whether the latter was unique to costly punishment, or caused by the 'warm glow' felt when anti-social individuals get their comeuppance. Chapter 4 also further investigated under what circumstances the signalling benefits of punishment are generated, specifically whether they are dependent on the success of any intervention.

2.5.3 Chapter 5: Dominance rank and observer perceptions of punishers

Chapter 5 investigated whether the reputational benefits investigated in Chapter 4 were dependent on the dominance of an intervening third party, i.e. whether only dominant individuals could actually access the indirect benefits of punishment. Chapter 5 also investigated participants expectations of dominant individuals, i.e. are they expected to punish, and how status affected the perceived risk of retaliation. Chapter 5 also investigated how the rank of both the defector and punisher affected observer perceptions of costly punishment.

2.5.4 Chapter 6: Dominance and the behaviour of punishers

Chapter 6 attempted to simulate one aspect of a dominant position by allowing certain individuals to benefit disproportionally from any group success. It invested whether individuals would be more willing to punish when the net cost to punishers was low. Chapter 6 also investigated whether the way in which these additional resources were generated affected behaviour; would better-resourced individuals be more willing to punish free-riding when they were explicitly benefitting at the expense of others? Finally, Chapter 6 also investigated whether punishment was used strategically to ensure continued benefit for the 'dominant' punishers.

2.5.5 Chapter 7: Dominance and behaviour - naturally occurring dominance

Chapter 7 tested whether being in a dominant position affected an individual's sensitivity to unfairness in the local environment. Specifically Chapter 7 investigated whether an individual's position in an informal social hierarchy affected their cooperative and punishment decision making. Chapter 7 also investigated whether, based on position in an informal social hierarchy, punishment behaviour would be affected by the possibility of gaining a reputation as a punisher.

2.5.6 Chapter 8: General discussion

Finally, Chapter 8 discusses whether the studies presented in this thesis provide evidence that dominance is associated with costly punishment and why this is the case. It will be shown on a proximate level that dominance is an intrinsic part of costly punishment, both in terms of those engaging in punishment itself and in our perceptions of punishers. As such, regardless of the ultimate cause of the human other-regarding preference, dominance is a factor that should be considered in any future model, experiment or theoretical explanation of costly punishment. However, the general discussion will also show that dominance can itself provide a coherent explanation for the ultimate cause behind the human desire to punish unfairness, and that dominance is a fundamental aspect of the socio-cognitive reasoning surrounding fairness. With dominance established as a viable hypothesis for the evolution of costly punishment, Chapter 8 will offer future research suggestions to further probe this relationship.

3 Chapter 3: measuring punishment behaviour

3.1 General introduction

Human cooperation may be unique in the natural world: while traditional theories of kin selection, direct reciprocity and signalling/reputation building have been used to explain cooperation in humans and, to an extent, in other animals (Bird & Smith, 2005b; Hamilton, 1964; Trivers, 1971), no other creature displays our willingness to cooperate with non-kin individuals (Melis & Semmann, 2010).

One, perhaps the main, reason for our consistently high level of cooperation appears to be punishment: specifically punishment to uphold norms of fairness and egalitarian sentiment at their own apparent expense (Gintis, 2000). This *altruistic* punishment not only drastically increases cooperation between both strangers and partners (Fehr & Gächter, 2000; Ostrom et al., 1992) but is something individuals both desire to have (Gürerk et al., 2006) and find rewarding when it occurs (de Quervain et al., 2004). Because of this, the costly punishment of anti-social behaviour is now seen by many as not only a human universal behaviour but one that is unique to the species (Fehr & Gächter, 2000; Marlowe et al., 2010).

However the evolution of this behaviour presents something of a puzzle. While punishment can be evolutionarily stable once established in a population as it becomes essentially cost-free (Boyd et al., 2010; see also Masclet et al., 2003 for this in practice), before this point is reached the individual cost of the behaviour means that punishers can be outcompeted by defectors and non-cooperators, and by second-order free riders; those who cooperate but don't punish (Dreber et al., 2008; Yamagishi, 1988). However, punishment could potentially be evolutionarily stable if the net cost of punishment could be reduced, either by punishers individually recuperating the cost of their behaviour (i.e. through means that would not

benefit the group as a whole, unlike improvements to overall group efficiency, which would also benefit free-riders) or if the cost of the behaviour itself could be reduced.

The current chapter investigates two such possible mechanisms that costly punishers could employ to alter the cost of punishment. Study 1 investigated whether punishers would vary their punishment behaviour depending on the potential for reputational gains, and addresses the questions concerning indirect reciprocity and signalling that were raised in 1.2.1, 1.2.1 and 2.3.4. Study 2 investigated whether punishers would alter their behaviour in response to a signal that others would also be willing to punish and thus share the cost. Study 2, as well as addressing indirect benefits, examines how the cost of punishment could be spread between multiple punishers (1.1.2/ 2.3.2).

This chapter also examined whether a questionnaire/vignette method can be effectively employed to study costly punishment. The majority of work in this area has been conducted using economic games (but see Mathew & Boyd, 2013; O'Gorman et al., 2005). However the vignette method, as well as being used consistently in social psychology, has also been used to study other phenomena related to human evolution (for example altruism, Barclay, 2010; formidability, Fessler et al., 2013; mate choice, Iredale et al., 2008; and for norm violation and a general discussion of the use of vignettes, see Wilson & O'Gorman, 2003) and thus can potentially be applied to the study of costly punishment. The advantage of the vignette method is that conditions can be systematically varied to provide clear data without violating the proscription against deception in the economic literature (Bonetti, 1998), and that these scenarios can be varied in ways that would be difficult to simulate within an economic game. Nevertheless an attempt was made to construct each scenario around the logic of a public goods game; that is to say involving a small group of people in which one member fails to contribute to a common good.

There are however some issues with this method. First, a number of reviews have suggested that a performance-related payment method (based on cooperation/punishment decisions) produces the most robust decisions from participants (Etzioni, 2010; Perc & Szolnoki, 2010; but see Guala, 2012; and, Levitt & List, 2007) and a vignette method would make this impossible. Equally, individuals will often indicate more willingness to punish unfairness when the situation is hypothetical rather than immediate; this is true both for experimental economics (Pedersen et al., 2013) and for wider research into social norm violations and moral dilemmas (for example Kawakami et al., 2009; Patil, Cogoni, Zangrando, Chittaro, & Silani, 2014). However it should be noted that the 'strategy method' (see Fischbacher & Gächter, 2010), despite being criticised for inflating punishment behaviour, is commonly used in economic games and is often required to make an experiment feasible. Furthermore the issue of inflated rates of punishment has been directed at economic experiments in general (Guala, 2012). Thus a vignette method should at least been seen as no worse than the alternatives.

3.2 Study 1: is punishment behaviour sensitive to reputational gains?

3.2.1 Reputation

Humans are very adept at image scoring, i.e. keeping track of an individual's reputation (Nowak & Sigmund, 1998); indeed, there are great advantages to associating with like-minded others (Santos, Pacheco, & Lenaerts, 2006) and the ability to predict (and manipulate) future behaviour is one of the strongest explanations for the evolution of human cognitive ability (Byrne & Whiten, 1997; Dunbar, 1998). In terms of cooperative and/or altruistic behaviour, individuals behave far more pro-socially when they believe the situation places their reputation under threat (Bateson et al., 2006) and conversely will reduce such behaviours when the likelihood of future interaction with a specific individual is low and when groups are unstable (O'Gorman et al., 2005). If the purpose of altruism is to secure a

positive reputation, there is little point in behaving altruistically when anyone who observed the behaviour is no longer around.

Therefore, one possible way in which individuals can make up the costs of punishment is through some sort of reputational gain. It has been suggested that we image-score punishment in a similar fashion to altruism (Tennie, 2012) and there is some evidence that individuals like those who punish, and that this positive affect translates into actual material rewards (Barclay, 2006; Nelissen, 2008). Indeed, a number of models have shown that such indirect benefits from an act of punishment can make the behaviour evolutionarily stable (Panchanathan & Boyd, 2004; Sigmund et al., 2001). If such reputational gains were important in enabling the evolution of costly punishment, we may expect punishment behaviour to be sensitive to conditions that would maximise the potential for such reputational gain, and this does indeed seem to be the case. When punishment decisions are non-anonymous, individuals are far more willing to engage in punishment (Bering, 2008; Kurzban et al., 2007) as more individuals are present to witness the behaviour. Also when groups are stable, that is to say the same individuals will interact repeatedly, individuals are far more willing to punish non-co-operators (Masclet & Villeval, 2008; Ostrom et al., 1992).

3.2.2 Type of reputation

One fundamental question it is important to address is what sort of reputation an individual would get from an act of punishment. It has been suggested that punishment might act as a costly signal of an individual's commitment to fairness norms; bearing the cost of costly punishment demonstrates, for example, you are an honest person (Nelissen, 2008), and indeed, generally those who punish are also highly cooperative (Falk et al., 2005). Such a reputation may help recruit cooperative partners or coalition members (something potentially vital in our evolutionary history; Gavrilets et al., 2008) as, for example, we prefer to be in an environment where *someone* will punish unfairness or defections (Gürerk et al., 2006;

Rockenbach & Milinski, 2006). Indeed, one of the more direct proximate motivations of punishment is to change the behaviour of the defecting individual (Fudenberg & Pathak, 2010) or to remove them from the group altogether (Bowles & Gintis, 2004; Masclet, 2003). The fear of defectors inhibits cooperation (Fischbacher et al., 2001), so by signalling a willingness to prevent this, punishers provide an environment safe from such threats and are thus seen as people worth associating with: it is unlikely to be coincidence that the traits attributed to punishers, such as trustworthiness and being group-focused, are those demanded of leaders (Hogg, van Knippenberg, & Rast, 2012).

Alternatively, engaging in punishment might be less about signalling hidden pro-social personal characteristics and more about signalling personal formidability. While non-human animals show only limited evidence of costly punishment, the few examples of a non-human animal appearing to punish norm violations occur only where there is a large asymmetry in dominance (Flack, de Waal, & Krakauer, 2005; Flack, Girvan, De Waal, & Krakauer, 2006; Wong et al., 2007). Thus punishment might enable an individual to signal that they are not to be treated unfairly in future dyadic interacting (Barclay, 2006). In fact, Marlowe et al (2008) suggested that one reason for the lack of third party punishment in small scale societies is that, due to eavesdropping on dyadic interactions, a "don't mess with me" reputation can be easily established without an individual involving themselves in the conflicts of others (See also, Rand, Ohtsuki, et al., 2009).

3.2.3 Personality and punishment

Individuals experience a great deal of 'moral outrage' (Trivers, 1971) when observing an unfair interaction and this emotional response is a strong predictor of punishment (de Quervain et al., 2004; Falk et al., 2005). Because of this we may expect personality traits that lower social affect to also lower punishment behaviour, traits such as those defined as the

'Dark Triad' of personality: Psychopathy, Narcissism and Machiavellianism (Paulhus & Williams, 2002).

These personality traits are especially characterised by low affect, low empathy for others and short-term strategizing (Paulhus & Williams, 2002), and when individuals high in trait-Psychopathy received low offers in an Ultimatum Game they accepted these offers more often than others (Osumi & Ohira, 2010). In fact high-trait individuals showed behaviour far closer to the Nash equilibrium than other participants by accepting far more of these lower offers, so we may expect that high Dark Triad individuals will also not punish in a third party situation. Nevertheless there is limited and inconsistent data as to how levels of Psychopathy and how other 'Dark Triad Traits' would affect behaviour in a punishment situation. For instance those high in the Dark Triad traits should forgo punishment as they are not emotionally aroused by unfairness, i.e. they don't care about other people. However both Gunnthorsdottir, McCabe & Smith (2002) and Fehr & Schneider (2009) demonstrated that, albeit in a different context, individuals with high Machiavellianism in particular were sensitive to potential future gain and the loss of reputation. Furthermore, two of the traits of Psychopathy are aggressiveness and impulsivity so we may expect high Dark Triad individuals to punish indiscriminately. The current study therefore included a measure of the Dark Triad to see how these personality traits affected punishment behaviour. However while the Dark Triad may affect punishment behaviour, it is not possible to make a definitive prediction about the direction of any effect.

3.2.4 The current study

Using fictional vignettes, the current study investigated whether punishment behaviour would be sensitive to the potential for reputational gain. The vignette scenarios were manipulated to vary the stability of the groups participants were described as being a member of, and the extent to which their actions would be observable by others. It was predicted that participants

would be most likely to engage in punishment when the groups were described as stable and when there would be an audience to that behaviour.

3.3 Method

3.3.1 Participants

Participants were recruited from the psychology department of the University of Exeter via the university's internal email system. A total of 86 participants, 29 Males (M age = 27) and 58 females (M age = 24) with an overall age range of 18 – 46 completed the questionnaire and answered the manipulation check questions successfully (3.3.5). A further 22 participants did not complete the questionnaire or failed to correctly answer the manipulation checks. Recruitment was conducted between December 2010 and January 2011.

3.3.2 Materials and procedure

The survey consisted of two sections. The first section presented participants with an experimental vignette and the second section collected personality and demographic information. The survey was conducted using the web-based application SurveyMonkey (www.surveymonkey.com) and was presented to participants in the order shown below.

3.3.3 Group stability and Audience to punishment

Participants were presented with a short vignette asking them to imagine that they were part of a 5-person university study group tasked with producing a presentation for a course module, a project worth 50% of the marks for that module. The vignette then described how one member of the group failed to complete their assigned task. It was stressed the lack of contribution from this group member was not due to any extrinsic factors such as ill health (for the full vignette, see Appendix A). Participants were then asked the following questions.

a) How angry are you with the group member who did not work on the presentation? On a scale of 1 – 7 (1 being not angry at all and 7 being extremely angry)
b) How much would you like to see this person sanctioned in some way for their actions? On a scale of 1 – 7 (1 being not at all and 7 at lot)

The second part of the vignette described the participants preparing to approach the course tutor in order to request that the mark received by the defecting group member be reduced. Participants were then asked the following questions.

a) How likely is it you would do the above? On a scale of 1 – 7 (1 = never, 7 = being definitely)
b) How likely do you feel it is that your actions would be supported by the group? On a scale of 1 – 7 (1 never and 7 being definitely)
c) How much would you wish the non-contributor's mark to be reduced by? On a scale of 0 (no change) to 100 (no marks at all)

The study used a 2x3 design with the stability of the group and audience for the punishment (who else knew the participant was about to approach the tutor) being manipulated. In the case of Group Stability, participants were either informed in part 1 of the vignette that the group would be permanent for that year or that it was a one-off group for that assignment only. For Audience, in part 2 participants were informed they had either approached the tutor privately, had told the other group members what they were planning, or had let it be known to the whole course that they were requesting a grade reduction for the defecting member.

The final part of the vignette described how another member of the group, Avery (chosen from an online list of the top 20 androgynous names; accessed November 2010) told the participant and other group members that they had informed the tutor already. Participants

were then asked to answer the following questions on a scale of 1-7 (1=strongly disagree, 7=strongly agree).

a) Avery is a trustworthy individual
b) Avery is group focused
c) Avery is a 'nice' person
d) I would happily work with Avery in the future
e) I would invite Avery to a social occasion

These 5 items were adapted from Barclay (2006) and had a high reliability index (α=0.88). They were therefore collapsed into the single variable "Attitude to Avery"

3.3.4 Personality Measure: The Dirty Dozen

The Dirty Dozen (Jonason & Webster, 2010) is a 12-point scale developed to measure the "Dark Triad" (Paulhus & Williams, 2002) of personality traits: Machiavellianism, Narcissism and Psychopathy. The scale provides a compressed alternative to individual measures of these traits as together amount to 120 items.

The measure consists of statements such as "I tend to lack remorse" to which participants indicate their level of agreement using a 7-point scale, (1=strongly disagree, 7=strongly agree). There is a minimum score of 12 and a maximum of 84. The alpha reliability for this measure was 0.86.

3.3.5 Manipulation check and demographic questions

Following completion of the personality measure, participants were asked the manipulation check question. This question asked them to identify, from a list of four possible choices, what occurred in the scenario. Data from any participant who did not answer this question correctly were excluded from the study. Finally, participants were asked to indicate their sex

and age. Following completion of this information, participants were shown a debriefing screen explaining the study and thanked for their participation.

3.4 Results 1: behaviour of participants

3.4.1 Punishment

There was a strong correlation between the amount individuals were willing to punish and the other variables (Anger at defection, desire to see that person punished, willingness to engage in punishment, and feeling of support; see Table 3.1). These variables correlated highly with one another and had a high reliability index (α=0.74) and therefore were collapsed into a single variable "Outrage". This new variable was also strongly correlated with the amount participants wished to punish the defector (r_s=0.53, N=87, p<0.001)

3.4.2 Group Stability

Group Stability did not affect Outrage at the defection (stable, M=4.3, SD=1.3, unstable, M=4.4, SD=1.1; $F_{1,85}$=0.38, p=0.54), however there was a trend towards Group Stability affecting the amount of punishment participants wished the defector to receive (stable, M=32.4, SD=29.3; unstable, M=44.0, SD=33.1; $F_{1,85}$=30.9, p=0.08) with participants wishing to reduce the defector's mark by a greater amount when the group was unstable As shown in Figure 3.1, there was an uneven distribution in punishment (although statistically the data are normally distributed; ks=1.320, n=90, p=0.061), with the majority of individual choosing to reduce the defector's mark by 0%, 50% or 100%. To analyse the data further therefore, a median split was carried out on the data and, as shown in Figure 3.2, participants in the stable group condition were more likely to punish below the median amount of punishment (X^2_1=3.98, p=0.046).

Table 3.1: the relationship between punishment and the emotional response to a defection

	Anger	Desire for punishment	Willingness to punish	Feeling of support
Amount punished	**0.321****	**0.47****	**0.40****	**0.46****
Anger		0.68**	0.42**	0.31**
Desire for punishment			0.51**	0.32**
Willingness to punish				0.28**

(N=87, **<0.001, two tailed)

3.4.3 Audience

As the Audience manipulation was presented after participants had responded to 'anger' and 'desire to see defector punished' items, it was not appropriate to use the 'Outrage' variable. Therefore, the audience effect on the amount of punishment, willingness to punish, and feeling of support were analysed separately. Audience did not have any effect on participant responses (MANOVA, $F_{2,84}$=0.62, p=0.94) and nor were there effects on any individual items. Participants were also equally likely to punish above or below the median amount

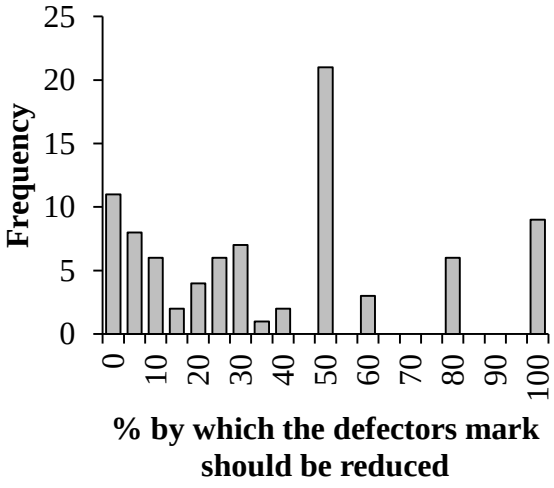


Figure 3.1: distribution of the percentage participants wished the defector's marks to be reduced.

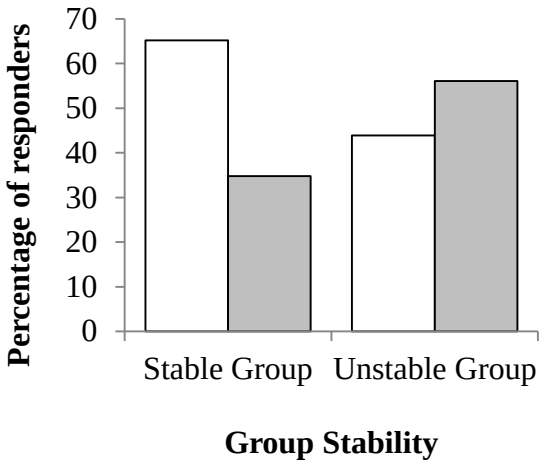


Figure 3.2: percentage of participants who punished above (grey) or below (blank) the median amount of punishment.

regardless of the potential Audience (X^2_2=3.81, p=0.144)

3.4.4 Group Stability and Audience

As above, the amount of punishment, willingness to punish, and feeling of support were analysed as individual variables. There was a significant overall interaction between Group Stability and Audience (M ANOVA, $F_{6,160}$=2.67, p=0.016), however there was no significant effect of this interaction on individual variables. An interaction between Group Stability and Audience also did not influence whether participants punished above or below the median amount (Wald X^2_5=8.91, p=0.11).

3.4.5 The Dark Triad

There was no correlation between Dark Triad scores and the amount participants were willing to punish the defector (r_s=-0.14, N=87, p=0.21) or the amount of 'Outrage' participants felt at that defection (r_s=-0.06, N=89, p=0.59). Individuals who punished above the median amount were no more or less likely to exhibit Dark Triad traits than those who punished below (Mann-Whitney, U=743.5, N=47/39, p=0.14).

3.5 Results 2: perception of a punishing other

3.5.1 Punishment and Outrage

As shown in Figures 3.3 and 3.4, participants liked Avery (the group member who punished the defector) more when they themselves were outraged at the defection (r_s=0.33, N=87, p=0.002) and when they wanted the defector to be punished by a large amount (r_s=0.27, N=87, p=0.013). When entered into a Stepwise regression model only Outrage predicted attitude to Avery (Adjusted R^2= 0.16, $F_{1,87}$=17.88, p<0.001).

A mediation analysis was conducted with Outrage as the predictor of participant's attitude to Avery, with the amount of punishment participants felt the defector should receive as the

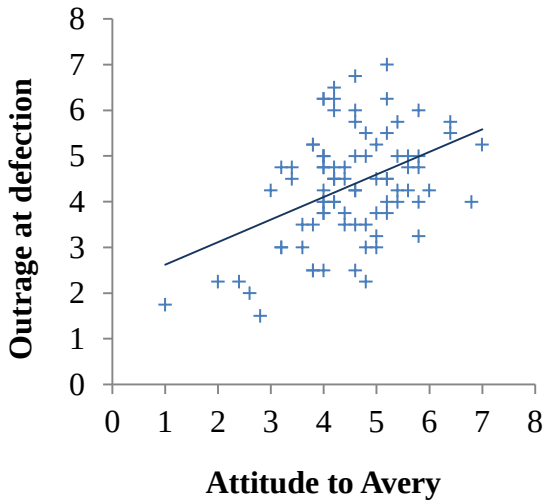


Figure 3.4: relationship between participant's outrage at a defection and their attitude to a punisher.

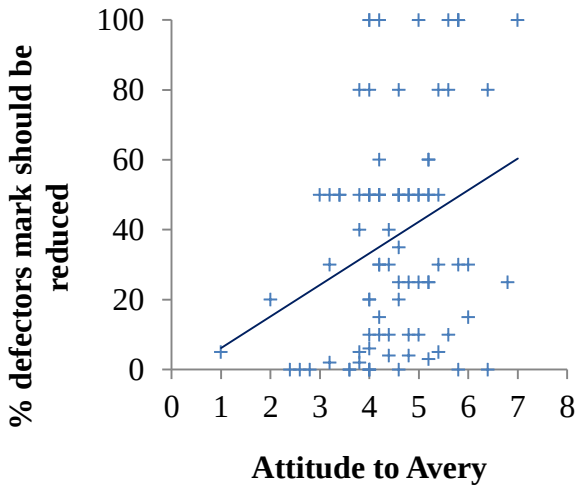


Figure 3.3: relationship between the severity of the punishment demanded by participants and their attitude to a punisher.

mediator. There was no mediation by Punishment on the relationship between Outrage and Attitude to Avery ($CI_{95\%}$= -0.05, 0.16).

3.5.2 Audience, Group Stability and the Dark Triad

Participants' attitude to Avery was not affected by how stable the group was ($F_{1,81}$=0.06, p=0.80) or the audience the participants themselves would have faced had they punished ($F_{1,81}$=0.52, p=0.60). There was also no interaction effect of Stability or Audience on participant's Attitude to Avery ($F_{2,81}$=0.82, p=0.45). There was also no relationship between Dark Triad personality traits and attitude to Avery (r_s=-0.003, N=87, p=0.98).

3.6 Discussion

3.6.1 Group Stability

One of the more direct proximate motivations of punishment is to change the behaviour of the defecting individual (Fudenberg & Pathak, 2010; Masclet, 2003). Thus by initially investing in punishment, punishers ensure cooperation in the future. It was predicted therefore that in a situation where the participants expected to work with the same individuals again, that they would be more willing to punish someone who defected from a group task.

However, the current study found the opposite; participants were more willing to punish when groups were unstable. While this affect was marginal, it does contradict previous findings. One explanation may be the fear of retaliation from the defector. Both theoretical and experimental studies that have demonstrated that the prospect of retaliation reduces costly punishment (Nikiforakis, 2008; Rand et al., 2010) and even without the explicit ability to counter-punish, defectors who have been punished will use subsequent punishment rounds to retaliate if the person who punished them is identifiable (Cinyabuguma et al., 2006). Indeed outside the laboratory, one of the primary reasons members of the public do not inform the police of criminal behaviour is the fear of being identified by the criminal fraternity (Tarling & Morris, 2010). This may be why other studies have found that individuals are just as willing to punish when the behaviour is anonymous (Fudenberg & Pathak, 2010) and will actually pay to hide their punishment (Rockenbach & Milinski, 2011). Thus, while participants in both conditions felt equally outraged at the act of defection, and outrage is strongly connected to punishment (Marlowe et al., 2010; Reuben & Van Winden, 2008; Sigmund, 2007; Van't Wout et al., 2006), only in a situation where it was unlikely that participants would have to face the target of their punishment were they willing to punish in line with their emotional reaction. When retaliation was possible, i.e. they would meet the person again, participants had to reconsider whether to punish.

Another explanation may be that participants did *not* want a reputation as a punisher. It has been suggested that we reputation-score punishers in a similar fashion to co-operators (Tennie, 2012) and there may be negative, as well as positive, repercussions to being seen as a 'punisher'. In the most minimal sense, costly punishment is a dyadic interaction between a defector and a third-party who has voluntarily entered into a conflict with the defector (although in the case of the current scenarios, via the course tutor) The outcomes of such interactions can imply dominance and status (Jones, DeBruine, Little, Watkins, & Feinberg,

2011; Tiedens & Fragale, 2003). Being perceived as dominant can have negative consequences on the opinion others form (Stirrat & Perrett, 2010; Vaillancourt & Hymel, 2006) and there are mechanisms that ensure individuals cannot 'false-signal' dominance or status (Anderson et al., 2008; Számadó, 2011a; Tibbetts & Izzo, 2010). Indeed Számadó (2011b) demonstrated that signalling displays are kept honest by long term commitments, thus when groups were stable participants would have to bear the social 'badge' of any actions they.

Thus, while individuals may want to be well-liked by others they may not wish to be seen as formidable or to assume the role of 'policeman', and with these labels the retaliation and conflict they could invite. This is not an aspect of punishment that has received much attention, but examining the social consequences of punishment (e.g. on the perceptions of observers) may help explain why people do or do not punish, especially in everyday life.

The results of the current study did not contradict all previous research as, employing a similar methodology O'Gorman et al. (2005) failed to find any effect of group stability on punishment behaviour. While O'Gorman et al (2005) attribute this to a group selection explanation for human sociality, the way in which both that result and the current study differ from the literature could be due to the vignette method. On the one hand, punishment in both studies was 'cost free' in the sense that the hypothetical scenario did not describe an explicit imagined cost on the participant for their choices, and the cost of punishment is another strong predictor of its occurrence (Falk et al., 2005; Nikiforakis & Normann, 2008). Indeed many models suggest the cost to the punisher, as opposed to the effect punishment has on the target, may be the most important factor in the evolution of costly punishment (de Weerd & Verbrugge, 2011; Frank, 1996). Thus, the results of the study could be attributed to the fact participants did not face any actual costs to their behaviour and may suggest, that in terms of investigating punishment *behaviour* a survey/questionnaire approach may not be appropriate.

However, a caveat to this is that both theoretical models and experiments have demonstrated that punishment does not have to be costly to be effective at deterring free-riding and promoting cooperation (Bowles & Gintis, 2004; Masclet et al., 2003)

One reason for the above criticism is that individuals *over* estimate their willingness to punish (Pedersen et al., 2013), so the question remains why, rather than group stability having no effect, individuals punished more when groups were unstable; one explanation may also be retaliation. One limitation of economic experiments is they artificially limit the information and behavioural options open to participants. A case in point is that, despite being a consistent motivation of human individual and group behaviour (Barash & Lipton, 2011; Mathew & Boyd, 2011), retaliation was not considered as a part of cooperative/punishment behaviour until recently and when it was included the results were dramatic for punishment (Nikiforakis, 2008). By asking participants to take the perceptive of the vignette character (see Alexander & Becker, 1978), the current study asked them to invoke their individual experiences and perceptions of this social behaviour, informal peer-sanctioning, which would include any physical or social repercussions for 'moralistic' intervention. In fact, this is a principle that often underpins studies that employ economic games as a method (Levitt & List, 2007). While this is not to say the criticisms of the questionnaire approach are not valid, or that in the future other methods would be more appropriate to measure actual punishment behaviour, it does suggest that the use of vignettes can be considered a legitimate way to study evolved human behaviours.

3.6.2 Audience

One cannot establish a reputation without there being an audience to one's behaviour and a number of studies have suggested that when an audience is present individuals are more willing to engage in punishment (Bering, 2008; Kurzban et al., 2007). Such an effect makes sense if costly punishment can be used to signal hidden qualities about the individual such as

their commitment to fairness (Nelissen, 2008) or unwillingness to tolerate any unfairness directed against themselves in future social interactions (Barclay, 2006; Baumard & Liénard, 2011). However the opposite has also been found. Individuals are sometimes more willing to punish when the target of punishment would only be informed at the end of the experiment (Fudenberg & Pathak, 2010) and, interestingly, individuals will pay additional sums to have their punishment decisions kept hidden from the target and the group as a whole (Rockenbach & Milinski, 2011). With this in mind it was predicted that the audience for any punishment would have an effect on punishment behaviour, but without a predicted direction. However the study found not such effect in either direction.

While this finding contradicts the suggestion that individuals will aim to maximise (or minimise) any reputation gained from an act of punishment it does conform to an alternative theory for the evolution of costly punishment, Strong Reciprocity. This theory suggests that group-level processes selected for individuals who would act in the best interest of the group without concern for their individual well-being or in-line with their individual fitness (Gintis, 2000). The results therefore may indicate participants were demonstrating a general concern for the welfare of the group and egalitarian 'fairness' rather than acting to 'show off' their individual prowess. Thus it was more important that a social defector was punished for their actions than that the punisher received attention and acclaim for doing so.

A more mundane and methodological explanation may be that the manipulation was unsuccessful in making the participants feel suitably 'anonymous'. Certainly, in the economic literature at least, there is debate as to how truly anonymous conditions are (Levitt & List, 2007), with some researchers making a distinction between anonymous behaviour (where no one knows who carried out an action) and secret behaviour (where no one knows an action took place). The two can produce very different results (Winking & Mizer, 2013). However, there were significant effects of group stability on punishment behaviour and this may be

because "you will always work in the same group" is fairly unambiguous and the consequences of any action are easier to imagine – i.e. "I will have to continue to work with the person I just punished". However, the idea that once it was apparent someone had told the course tutor that this behaviour would remain secret may have been harder to imagine: this especially applies to whether participants could really imagine a difference in anonymity between the "working group only" and "full course" conditions.

3.6.3 Attitude to punishers

Perhaps the most interesting, and certainly the clearest, results of the study are those regarding the participants' attitude to another punisher, "Avery". There was a strong relationship between how outraged participants were, their willingness to punish and their attitude to Avery. These results agree strongly with the findings from the economic literature, firstly in that emotional response to defection influences punishment decision-making (Dawes et al., 2007; Falk et al., 2005), and more importantly here, that individuals who punish also like those who punish (Barclay, 2006). This does suggest that the results gathered by the vignette method can be seen as comparable to the economic literature in regards to how participants perceive other punishers.

The results of the study regarding the relationship between Outrage/Punishment and attitude to Avery suggest a number of mechanisms by which a punisher could gain from an act of punishment. Firstly, they could suggest that costly punishment operates in a like-attracts-like fashion that has been observed in cooperation (Albert et al., 2007; Fehl, van der Post, & Semmann, 2011), with individuals preferring to be in groups where everyone else will also punish. As one of the main theoretical problems with the evolution of punishment behaviour is the cost to the punisher, both in terms of the production of the behaviour (de Weerd & Verbrugge, 2011) and the potential for retaliation (Dreber & Rand, 2012), pooling punishment effort with other like-minded individuals would be advantageous. Boyd et al

(2010) suggested that punishment can evolve if there is a way to signal a willingness to punish and, Mathew & Boyd (2011), in their study of non-state tribes, found that before punishment occurs, a great deal of discussion takes place to ensure a large number of people agree with this decision. Experiments have also shown that, despite the costs of such mechanisms, participants prefer environments where the costs of punishment are pooled (Traulsen et al., 2012). This suggests that individuals may in fact be conditional punishers as well as conditional co-operators (Fischbacher & Gächter, 2005; Peter, Ottone, & Ponzano, 2010) and that engaging in punishment is as much a coordination signal as it is a costly signal of one's character.

However, while there was a relationship between participants' punishment behaviour and their attitude to Avery, this relationship disappeared when participants' outrage was considered as a variable: even individuals who didn't punish the defector, despite being outraged, still liked Avery. Thus a second explanation is that the reputational benefits of punishment are not so much "like-attracts-like" as "the enemy of my enemy is my friend". In a very Hobbesian fashion, despite the threat of perverse or anti-social punishment (Cinyabuguma et al., 2006; Herrmann et al., 2008), the reduction in group efficiency punishment causes (Traulsen et al., 2012), or simply the lack of freedom to free ride, individuals prefer environments were punishment is possible (Gürerk et al., 2006; Rockenbach & Milinski, 2006). Engaging in punishment therefore may not be a signal to attract other punishers *per se*, but a signal to others of your willingness to uphold fairness norms yourself, which will help attract cooperative partners. As individuals tend to fear defection, and reduce their cooperation as a result (Fischbacher & Gächter, 2010), joining the group of someone with a reputation for enforcing cooperation and fairness will likely be advantageous for all. Studies (for example, O'Gorman et al., 2009) have shown that a single punisher can be as effective as many. Why outraged participants liked Avery is the reflection

of why they didn't punish as expected; participants were unwilling to don the 'punisher' badge themselves, but want to be around an individual who was willing to.

The advantage of the latter explanation is that, while in both explanations punishment acts as a costly signal, in the case of the latter explanation the returns would not be diminished by second-order free-riding. In the former case however the cost of punishment cannot be recovered if others do not join in with the behaviour. Indeed, Boyd et al (2010) comment that one issue of punishment as a coordination signal is that someone could incite punishment but not actual take part, i.e. use 'moralisation' strategically (Peterson, 2012). However, if punishment is a costly signal of other qualities, then any cost is recuperated through gains in reputation, be it egalitarian intent or formidability. This indicates that while potentially costs could be reduced if other individuals also punish (and this may certainly happen), that is not the primary mechanism by which the costs of punishment are reduced or offset.

3.6.4 Dark Triad

The study did not find any evidence that Dark Triad personality traits had any effect on punishment behaviour, nor were they associated with the level of anger participants felt at an act of defection. As stated in 3.2.3, on the one hand Dark Triad traits are associated with low social affect (Paulhus & Williams, 2002) and more *rational* economic behaviour (Osumi & Ohira, 2010), but on the other they, and especially Machiavellianism, are associated with social manipulation and reputation management (Gunnthorsdottir et al., 2002; Paulhus & Williams, 2002). Therefore no predictions were made about the direction the effect of Dark Triad traits on punishment might take. Nevertheless it was predicted that they would have an effect on punishment, and this was not the case. While emotion might play a role in punishment behaviour (Falk et al., 2005), as discussed in the sections above (2.6.2, 2.6.3) and as will be elaborated upon in Study 2, individuals might demonstrate more strategic concerns when choosing to engage in punishment.

3.6.5 Limitations

One direct limitation of the study, which may be responsible for the contradiction in results between group stability and audience, is the ordering of the questions. The group stability manipulation appeared first in the text and after this participants answered a series of questions. It was only following this that participants were informed about the audience for punishment, and they may not have attended to this new information. While a general manipulation check was carried out ("*What happened in the scenario"*) the questionnaire did not include a specific manipulation check to ensure participants noticed the manipulation. Additionally, the study did not include the option to give a qualitative explanation for their answer, which might have provided further evidence of success/failure of the experimental manipulations and whether factors such as group opinion or the potential for retaliation were part of participants' decision making.

Another limitation, and one inherent to the survey method, is the lack of *actual* cost to the punishment decision making. Fundamentally participants are being asked to imagine what they would do '*if*' something occurred, and they may not be correct or honest in predicting their own behaviour. This is one reason why in behavioural economics there is a consensus that any pay off must be dependent on a participant's actual behaviour (Hertwig & Ortmann, 2001). This is not a problem that is easily solved, but others (e.g. O'Gorman et al., 2005) have argued convincingly that the fictional vignette approach does have merit. Indeed, the most concrete results from the current study were generated from the *attitude to a punisher* part of the questionnaire and these results match those from economic experiments, experiments that asked for responses to economic behaviour in terms of social opinions (Barclay, 2006) and actual monetary reward (Nelissen, 2008). This suggests that survey method would be more

suitable to investigate participant opinions about *others* who engage in costly punishment, rather than predictions about their *own* punishment behaviour.

Finally, the study did not measure whether the actions of Avery also led to him/her being seen as more formidable. It is important to test whether this occurs as one of the reasons we have suggested for the unstable group results of punishment behaviour is that participants did not want to be seen as formidable, even if they would also be treated well by others (Nelissen, 2008). This is also important as such a result may help explain the evolutionary origins of the behaviour as, for example Pederson et al (2013) suggested that no theory about the evolution of costly punishment could be considered accurate unless there was a clear connection to the behaviour of non-human animals. Given much aggressive behaviour in non-human animals, and certainly primates, is linked to dominance and status contests (Clutton-Brock & Parker, 1995; Silk, 2003), that punishers are perceived to be more formidable may suggest punishment has its origins here rather than in the maintenance of cooperation.

3.6.6 Conclusion

The study investigated whether participants would alter how much they believed they would punish a social defector depending on the stability of the group and the audience to the punishment. The audience for punishment had no effect on predicted punishment behaviour, and while group stability did affect this, the direction of the effect was contrary to predictions. Participants were more willing to punish in the unstable condition, and we suggest this was because participants were unwilling to endure the possible negative responses that punishment would bring in a stable group. Nevertheless, the latter result was marginal and we suggest that an alternative explanation for both results might be that (with numerous caveats) the use of vignettes is not appropriate to investigate punishment *behaviour.*

The study also found that participant anger and punishment behaviour strongly affected their opinion of a punishing other, Avery. We suggest that punishment can act as a costly signal, either to fellow punishers as a way to coordinate and reduce the cost of subsequent punishment, or to attract cooperative partners by signalling that the punisher will police fairness in their vicinity. These results also conform to the findings of experiments employing different methodologies and this suggests that vignettes could be a useful and practical tool in investigating the reputational rewards of punishment, as they are in other areas of social and evolutionary psychology.

3.7 Study 2: actions speak louder than words: the response to deceptive and non-deceptive signalling of punishment behaviour

Punishment, or the threat thereof, can be a powerful motivator for encouraging cooperation; groups not only cooperate more when punishment is possible (Fehr & Gächter, 2000), but such groups (eventually) also out-compete those where punishment is not possible (Gächter et al., 2008). However, while advantageous at a group level, the cost to the individual who engages in punishment is such that explaining how the behaviour could evolve has been problematic (Dreber & Rand, 2012; Dreber et al., 2008).

One potential solution to this problem is to distribute punishment between many individuals. Early research into the effect of punishment on cooperation found that when individuals were given a brief period prior to testing to discuss their responses to free-riding, more individuals were willing to punish free riding in the experiment itself (Ostrom et al., 1992). Indeed, when the punishment of social defectors occurs in non-state societies, it does so after a long consultation process amongst many individuals (Mathew & Boyd, 2011). More recently, Traulsen, et al (2012) showed that participants preferred a pool-punishment mechanism, where participants paid a small amount to a ‘punishment pool’ that was automatically used to punish low contributors if there were sufficient contributions to the pool, over a peer-

punishment system that is the standard in public goods games. This finding corresponds closely to a theoretical model by Boyd et al (2010) which demonstrated that costly punishment could be evolutionarily stable if punishers could somehow signal their willingness to punish a social transgression, but only actually (collectively) punished when a certain threshold of willing punishers was reached.

3.7.1 Honest signalling

One issue, however, is what form this signal could take for it to be an honest signal of intent. One potential solution is that engaging in punishment may itself act as an honest signal, due to the production costs (Frank, 1996) and potential repercussions (Rand et al., 2010; Tibbetts & Izzo, 2010). These costs would exclude anyone unwilling, or unable, to actually take part in punishment. For instance, while anti-social punishment does occur, the majority of individuals who punish are also highly cooperative (Barclay, 2006; Lehmann, Rousset, Roze, & Keller, 2007), with anti-social punishment disappearing entirely when punishment is no longer cheap (Falk et al., 2005). Study 1 in this chapter found a strong positive association between punishment behaviour and the attitude to punishing individuals, suggesting punishers may be seeking one another out based on their actual behaviour. This curtails the need for an independent signal of intent, as a reputation for punishment may itself act as the signal (Barclay, 2006), and mirrors findings from cooperation research suggesting that like-minded individuals cluster together (Albert et al., 2007; Skyrms & Pemantle, 2000). However, this creates a paradox whereby in order for punishers to signal and coordinate their punishment, and thus enable punishment to be evolutionarily stable, they have to punish independently first.

Therefore, as noted by Boyd et al (2010), if punishers are to signal their desire to punish, and if the signal is to be effective at both promoting mass action and not harming the producer, it must be low cost or cost free. A candidate for this type of signal may be language. Recent

work has demonstrated that the ability to publicly identify free-riders enhances cooperation (Bazzan & Dahmen, 2010) and it has been suggested that language evolved to aid the transfer of socially salient information (Dunbar, 2004). Indeed, while those who engage in social gossip are generally disliked, the gossiper can gain a positive reputation if the content relates to social defections (Peters & Kashima, In Press). So, vocalising anger or outrage over another's behaviour could act as a signal to others. There is, however, the issue of deceptive signalling. Language itself is very cheap to produce; an individual could pay a small cost to trigger others into punishment without paying the full cost of being involved. This is not an unlikely scenario as it has been suggested that moral outage at social defections may be an act of last resort by individuals who cannot defend themselves directly from exploitative or unfair behaviour (Peterson, 2012).

3.7.2 Conventional signalling and retaliation

A solution to this problem is that, while being cheap to produce, a vocal signal of outrage may be honest due to the response it provokes in conspecifics. For example, studies on territorial calls in birds have demonstrated that while vocalisations cost little enough energetically as to be considered cost free, they act as honest signals due to the antagonistic response they elicit from neighbours (Molles & Vehrencamp, 2001). In terms of punishment specifically, retaliation to an *act* of punishment has been shown to severely curtail punishment behaviour (Janssen & Bushman, 2008; Nikiforakis, 2008; Rand et al., 2010), and this effect may also apply to any sort of precursory behaviour. Firstly, verbal challenges are seen as effective punishment (Masclet et al., 2003; Ostrom et al., 1992), clearly suggesting that verbal challenges are perceived as punishment, and it has been found that verbal insults or challenges are taken very seriously, and are as likely to lead to physical fights as physical challenges (Felson, 1982). Thus, denouncing someone for their behaviour will likely result in a similar reaction to that triggered by actual punishment. Vocalising a willingness to punish

might, therefore, be a conventional signal, cheap to produce but with the potential reaction of the target ensuring its honesty.

Even if the retaliation costs are not as certain for signals of intent to punish as they are for punishment behaviour, there are other costs to signalling a willingness to punish. One such cost is that is forces the signaller to inhibit their own behaviour. We dislike hypocrites far more than plain defectors (Kurzban, 2012), so the act of signalling outrage at another's behaviour is costly in the sense that it prevents the signaller from taking part in the punished action themselves (Peterson, 2011). Equally costly might be the response of the other individuals who engaged in punishment after any signal was sent. In this regard, the situation can be seen as a prisoner's dilemma interaction where, following a mutual signal to punish, individuals can either cooperate (follow through with their intention) or defect (stand back and let others punish); and individuals react very negatively and seek retribution when they are defected against (Fehr, 2004; Nowak et al., 2000). In fact, such breaking of social contracts is viewed far more negatively by observers than, for example, the inequitable division of resources (Fehr, 2004). The advantage of this explanation is that it allows for the case where punishers can coordinate punishment without the target knowing: while someone willing to give a dishonest signal may never face retaliation from the target, they would still face a response from their duped comrades.

3.7.3 Personality and punishment

Study 1 did not support the idea that punishment behaviour would be related to Dark Triad personality traits. Measures of these variables were included again in the present study in case the results of Study 1 were idiosyncratic. However, their failure prompted the consideration of other individual difference variables that might be relevant, and one that commanded attention was dominance. Firstly, dominant individuals are in a position that inherently reduces the cost of punishment. Dominant individuals have access to greater

resources (for example, because others wish to associate with them, Barclay, 2013; and because they can monopolise group resources Cheney, 2011) and therefore the net cost of punishment is lower for dominants. Secondly, dominant individuals are less likely to face the threat of retaliation from the target of punishment; as in other forms of social interaction, subordinate individuals might simply submit to the demands of a more powerful individual (Clutton-Brock & Parker, 1995; Eckel et al., 2010). Finally, physical or social dominance is a strong predictor of anger (Sell, Tooby, et al., 2009) which in turn can lead to punishment behaviour. Indeed, research into welfare trade-offs has suggested that being in a dominant position calibrates an individual's expectations of how they should be treated and what they deserve, which manifests as anger when these expectations are not met (Sell, Tooby, et al., 2009).

Therefore, when faced with a social defection or free-riding, we may expect a more dominant individual to be more willing to punish because the production and/or retaliation costs are less for them, and because they will react more negatively to any behaviour that affects them personally (Brosnan, 2011; Cummins, 1999). Equally, and for the same reasons, we might expect more dominant individuals to react negatively to being 'tricked' by someone false-signalling a willingness to engage in punishment.

3.7.4 The Current Study

The current study therefore tested how punishers respond to signals from another individual. Firstly, it examined whether individuals were more willing to punish when others have signalled that they too wished to punish a defector. Secondly, the study investigated how participants would respond when this other individual acted contrary to or in accordance with their previous signal. Informed by the results of Study 1, as well as asking participants to indicate their desire to punish and their opinion of the other individual on 7-point scales, the current study also included a short qualitative section where they were asked to explain their

actions. This was to ensure that, compared to Study 1, there would be less conjecture as to the motivations of the participants.

3.8 Method

3.8.1 Participants

A total of 76 participants, recruited from the psychology departments of three UK universities via each of their internal email systems, the University of Exeter (n=33), Manchester Metropolitan University (n=34) and the University of York (n=9), successfully completed the survey. Nine of the participants were Males (M age = 26) and 67 were females (M age = 23); the overall age range was 18 – 51. 42 additional participants were excluded as they failed the manipulation check (see 3.8.6). All data were collected between April and May 2011.

No significant differences were found between the three departments on any of the measured variables, so the university participants attended was not included in further analyses.

3.8.2 Materials and procedure

The survey consisted of three sections: the main experimental treatment containing the signalling manipulation, a section that collected personality information, and a section collecting demographic information that also contained the manipulation checks. The survey was conducted using the web-based application SurveyMonkey (www.surveymonkey.com). The survey was presented to participants in the order shown below.

3.8.3 Signalling and punishment scenario

Participants were presented with a short vignette asking them to imagine themselves travelling to a ski resort with other students of unspecified gender after winning a prize draw. It was stressed that while everyone was from their university, no one on the trip knew one

another. Participants were informed that a member of their 5-person chalet ("Charlie"[1]) was consistently refusing to clean up after making a mess in the kitchen (for the full vignette, see Appendix A).

Participants then encountered the 'signalling' manipulation; they were informed of a conversation they had with another housemate ("Alex"[1]) who either said he found Charlie's behaviour be to unacceptable and said they should confront Charlie together; said he was annoyed but didn't want to get involved; or conversed about an unrelated matter (neutral control). Following this, participants were shown the following statements and asked to indicate their agreement, on a scale of 1 (not at all) to 7 (very):

a) How angry are you at the behaviour of Charlie?
b) How likely is it you would confront Charlie about his behaviour?
c) If you confronted Charlie, how likely is it that other people in the chalet would support you?

Following this, participants encountered the honest/dishonest manipulation. Participants were informed they had indeed confronted Charlie about his behaviour and had looked for support to Alex, who either did support them or did not, giving the study a 2x3 design. Participants were asked to indicate their opinion of Alex using the same 'likability' items as detailed in Study 1 (*2.3.3*). As in Study 1, these items had a high reliability index in this study (α=0.91) and were collapsed into the single "attitude to Alex" variable.

Participant were also asked to rate their feelings towards Alex using the Ekman emotions. They were asked for their agreement with the following statements, on a scale of 1 (1=strongly disagree, 7=strongly agree):

[1] A pilot study was conducted which determined these to be the most androgynous names. For the sake of clarity, male personal pronouns will be used when referring to these characters

a) I am angry at Alex
b) I am disgusted by Alex
c) I am afraid of Alex
d) I am happy with Alex
e) I am saddened by Alex
f) I am surprised by Alex

Participants were then presented with the final part of the vignette containing the critical question. The passage informed participants that it was the last day of the trip and they were the last to leave the chalet. Having locked the door, and carrying their luggage, participants were told they noticed a souvenir Alex had bought but had clearly left behind. They were asked to rate, on a scale of 1-7 (1=not likely at all, 7=very likely) how likely it would be that they would go back for his property. Participants were also asked to give a qualitative explanation for their decision.

3.8.4 The Trait Dominance-Submissiveness Scale (TDS)

The Trait Dominance-Submissiveness Scale (Mehrabian, 1994) is a 26-item scale designed to measure trait dominance independent of arousal or extraversion and contains questions such as "When I am with someone else, I usually make the decisions". The version used here asked participants to indicate their agreement with such statements on a 9-point scale (1= Very strong disagreement, 9= Very strong agreement) with a '1' response subsequently being scored as -4 and a '9' response scored +4. The alpha reliability for this measure was 0.91.

3.8.5 The Dirty Dozen

As with Study 1, the current study also examined any potential effects of anti-social personality traits on punishment behaviour. The Dirty Dozen (Jonason & Webster, 2010) is a 12-point scale developed to measure the "Dark Triad" (Paulhus & Williams, 2002) of

personality traits: Machiavellianism, Narcissism and Psychopathy. The scale provides a compressed alternative to individual measures of these traits which together amount to 120-items.

There were no significant results associated with the Dirty Dozen personality measure. Because the non-significant results replicated the findings of Study 1, the non-significant results of the current study will not be discussed further.

3.8.6 Comprehension and manipulation check questions

Following completion of the personality measure, participants were asked two questions about the scenario. One was a comprehension check which asked them to identify what Charlie was doing, and the second was a manipulation check to ensure they had noticed Alex's honest or deceptive signal to support them.

3.9 Results

3.9.1 Reaction to Charlie

As shown in figure 3.5, participants felt more supported when Alex also indicated his willingness to punish ($F_{2,73}$=4.15, p=0.02). Bonferroni-corrected pair comparisons found a significant difference in feeling of support between the help and refuse conditions only (p=0.016). Anger was strongly correlated with a willingness to confront Charlie (r_s=0.34, N=76, p=0.002), but was did not correlate with Support (r_s=0.02, N=76, p=0.84).

A step-wise linear regression carried out with Alex's signal, Support and Anger included as the predictor variables found that Anger accounted for 11% of variation in confronting/punishment behaviour (Unadjusted R^2=0.112, Adjusted R^2= 0.10, $F_{1,74}$=9.337, p<0.001). Alex's signal and Support were excluded from the model. This suggests that while

participants did respond to the manipulation, as suggested by the variation of support, this was not factored into the decision to confront Charlie about his behaviour.

3.9.2 Attitude to Alex

Whether Alex signalled that he would help (M=4, SD=1.6), refuse to help (M=4.1, SD=1.2) punish Charlie or did not mention the subject (M=3.9, SD=1.2) had no effect on how positively participants viewed him ($F_{2,70}$=0.37, p=0.69). As shown in Figure 3.6 participants did view Alex positively if he assisted them in the punishment ($F_{1,70}$=21.34, p<0.001). As also shown in figure 3.6, participant attitude to Alex was affected by an interaction between his signal and his behaviour ($F_{2,70}$=3.70, p=0.03), with participants viewing Alex more positively when he signalled his refusal or signalled nothing and then also refused to engage in the actual punishment of Charlie. That is to say, participants liked Alex more when his signal was honest, even if this meant he did not assist in their punishment of Charlie.

A step-wise linear regression was carried out with Alex's signal, Alex's behaviour, Support, Anger and willingness to punish entered into the model. Alex's behaviour accounted for 20% of the variance in positive attitude to Alex (Adjusted R^2= 0.20, $F_{1,74}$=19.71, p<0.001; β=0.46,

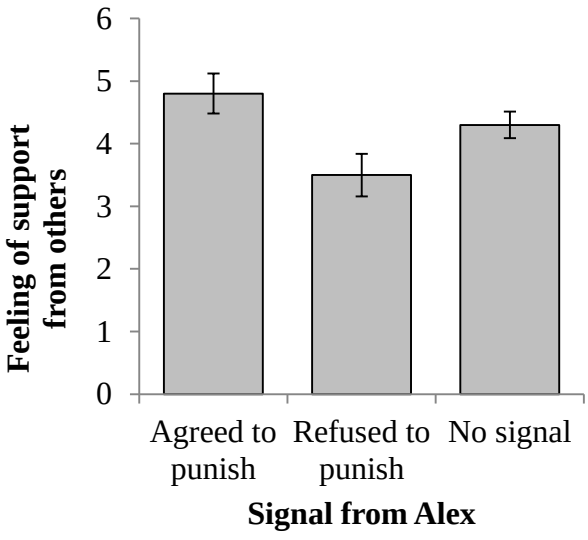


Figure 3.5: how supported participants felt in response to the signal of a conspecific.

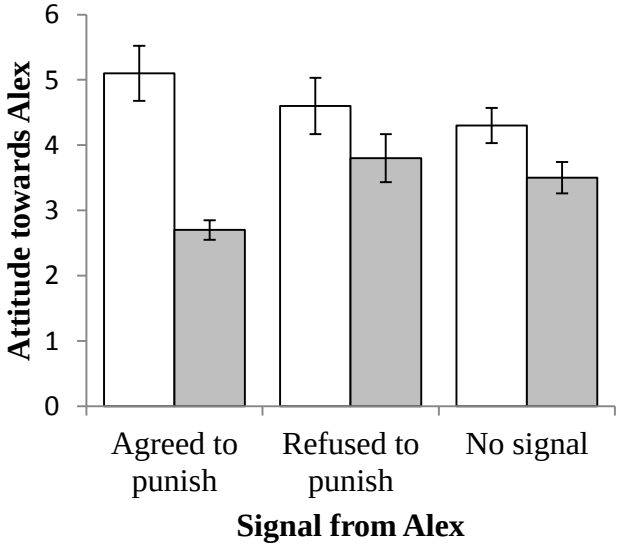


Figure 3.6: attitude to Alex when he did (blank) or did not (grey) assist in punishing Charlie in relation to the former's initial signal.

p<0.001), with Anger accounting for a further 4% (Adjusted R^2= 0.24, $F_{1,74}$=12.89, p<0.001; β=0.23, p=0.028).

3.9.3 Emotional response to Alex (The Ekman emotions)

There was no overall effect of Alex's initial signal on the emotional response to him (MANOVA, $F_{12,132}$=0.637, p=0.808), nor did the initial signal affect any emotional individually. However, as shown in Figure 3.7 Alex's actions did affect the overall emotional response to him (MANOVA, $F_{6,65}$=0.637, p<0.001), with participants being less happy ($F_{1,70}$=15.21, p<0.001), more disgusted ($F_{1,70}$=17.37, p<0.001) more angry ($F_{1,70}$=41.42, p<0.001) and more saddened ($F_{1,70}$=57.66, p<0.001) by Alex when he failed to aid in the punishment of Charlie. Participants' emotional response to Alex was also affected by an interaction between his signal and his action (MANOVA, $F_{12,132}$=2.59, p=0.004), however this effect was driven by surprise at Alex's behaviour ($F_{2,70}$=13.58, p<0.001, see Figure 3.8). With "surprise" removed from the analysis, there was no longer a significant overall interaction between signal and action on emotional response (MANOVA, $F_{10,134}$=1.004, p=0.443). This suggests that while participants clearly noticed any conflicts between Alex's

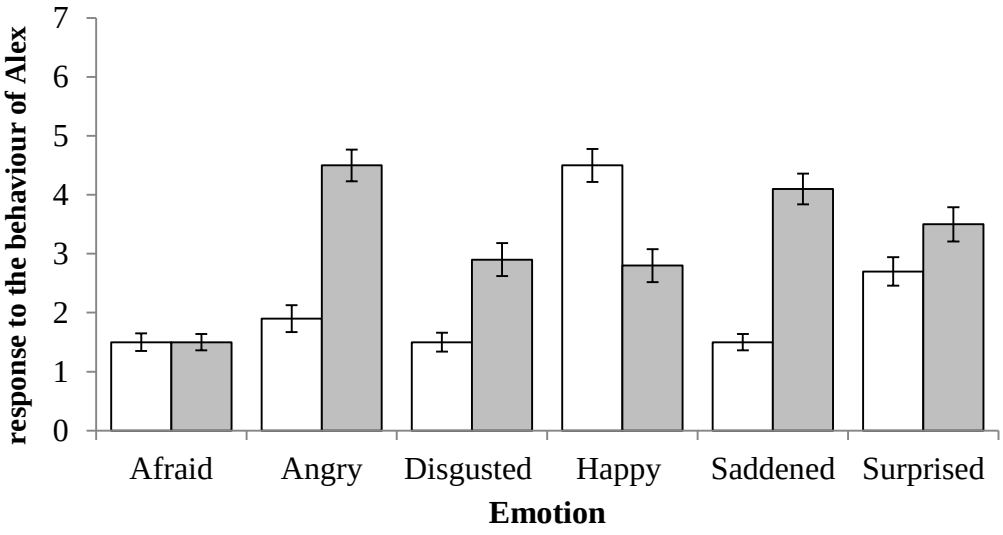


Figure 3.7: emotional reaction to Alex whether he punished (blank) or did not punish (grey). Bars = 1 standard error.

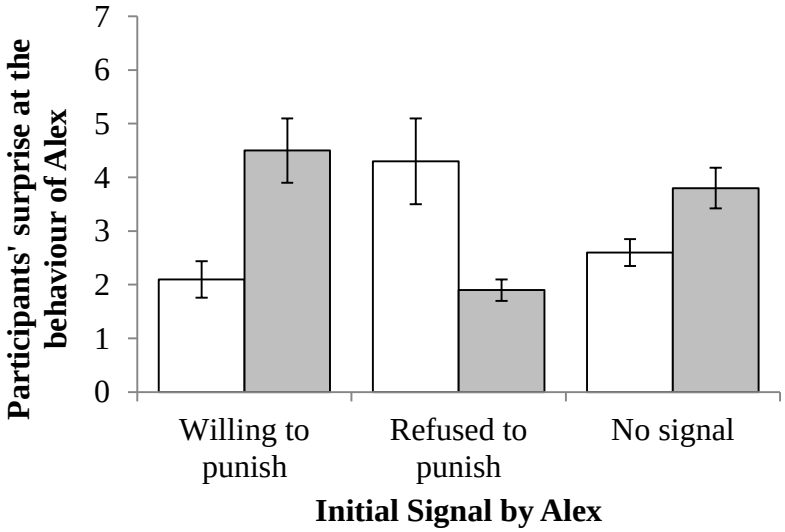


Figure 3.8: participant's surprise that Alex did (blank) or did not (grey) punish in relation to his initial signal.

signals and behaviour, their emotional response was driven by his actions alone.

3.9.4 Punishment of Alex

Participants were willing to help Alex by retrieving his lost property regardless of his initial signal (help, M=5.9, SD=1.5; refuse, M=6.1, SD=1.2, no signal, M=6.3, SD=1.2; $F_{2,70}$=0.54, p=0.59), his actual behaviour (help, M=6.2, SD=1.2; refuse, M=6.0, SD=1.2; $F_{1,70}$=0.66, p=0.42), or how these interacted ($F_{2,70}$=0.47, p=0.76). In addition, there was no relationship between the anger participants may have felt at Alex and their willingness to help him (r_s=-0.11, N=76, p=0.36) or any other variables. It should be noted that the mean response to this question was high, 6.1 of a maximum of 7; participants were simply overwhelmingly helpful.

Participants were also asked to give a qualitative explanation for their behaviour towards Alex. There were six common themes identified in participant responses (see Table 3.2) with "Moral" and "Quid Pro Quo" being the most common overall and occurring across both of Alex's potential behaviours. Many participants suggested they would want someone to do the same for them in the future; i.e. expected some form of direct or indirect reciprocity (Quid Pro Quo). Indeed while the prevailing response to Alex's refusal to help was "Two wrongs do

not make a right" some participants did indicate that a desire to be seen as the "better person" was behind their altruism.

Interestingly participants did take context into account when Alex did *not* signal any behaviour (the neutral condition). Participants made reference to Alex potentially being nervous or shy so that it would be wrong to punish him for his behaviour. Overall participants did favour helping Alex, but did so for very different reasons between the conditions. They were basing their behaviour more on Alex's "actions" when he had supported them, but when he did not, on general social norms surrounding appropriate behaviour.

Table 3.2: summary of most common participant responses to Alex's behaviour

Signal	**Help**		**Refuse**		**No Signal**	
Action	**Help**	**Refuse**	**Help**	**Refuse**	**Help**	**Refuse**
"Quid pro quo"	3	2	3	4	6	1
"Moral"	4	5	3	4	6	7
"Action"	2	2	1		3	
"Better Person"						3
"Effort"					4	1
"Understanding"						5

"Quid pro quo"	Participants indicated helped Alex as "they would like someone to do that for them"
"Moral"	Participants indicated it was the "right thing to do" or something they should do
"Action"	Participants indicated their behaviour was in direct response to Alex's actions
"Better Person"	Participants indicated that helping Alex would show what a better person they were
"Effort"	Participant indicated helping Alex would be little effort
"Understanding"	Participants specifically indicated they understood why Alex may have acted as he did

3.9.5 Dominance

3.9.5.1 Attitude to the defector (Charlie)

There was a significant positive correlation between dominance and a willingness to confront Charlie (r_s=0.47, N=76, p<0.001), however there was no positive correlation between participant anger at Charlie's defection and dominance (r_s=0.015, N=76, p=0.90)

Dominance was added to a step-wise regression model with support, anger and Alex's signal and this revealed dominance to be individually to be responsible for 20% of variance in willingness to confront (Adjusted R^2= 0.209, $F_{1,74}$=20.821, p<0.001), with anger and dominance predicting 31% of variance (Adjusted R^2= 0.299, $F_{2,73}$=16.983, p<0.001). This suggests that trait-dominance had a more powerful influence on a willingness to confront than emotion or the potential for support from others.

3.9.5.2 Attitude to Alex

Trait-dominance only correlated significantly with how likable participants believed Alex to be (r_s=-0.227, N=76, p=0.015), with more dominant individuals liking Alex less. It did not correlate with any of the emotional responses to him. A step-wise regression model was created with Anger at Charlie, Dominance and Alex's actions and, as shown in Table 3.3, while dominance was added into the model, it only explained 4% of the variance in attitude to Alex. While dominance could affect how participants respond to defection (see Sell, Tooby, et al., 2009) trait-dominance did not mediate the relationship between Alex's behaviour and their opinion of him ($CI_{95\%}$ =-0.012/0.03).

Table 3.3: model summary

Model	variables	*Adjusted r2*	*F*	*df*	*P*	*β*
1	Behaviour of Alex	0.2	19.71	1,74	<0.001	0.46
2	Behaviour of Alex	0.26	12.9	2,73	<0.001	0.46
	Anger at Charlie					0.23
3	Behaviour of Alex					0.44
	Anger at Charlie	0.31	11.02	3,72	<0.001	0.24
	Dominance					-0.23

3.9.5.3 Punishment of Alex

Trait-dominance did not correlate with the decision to retrieve Alex's property (r_s=0.126, N=76, p=0.27) and nor did dominance moderate any relationship with this decision and Alex's behaviour ($CI_{95\%}$=-0.02/0.02).

3.10 Discussion

Boyd et al (2010) suggested that if individuals could cheaply coordinate punishment prior to the action itself, then punishment behaviour could evolve even when initially rare. The primary aims of the study were to investigate the behavioural realities of this model, i.e. whether individuals would alter their punishment behaviour in response to a signal from another, and whether there would be any negative consequences to signalling a willingness to engage in punishment and then withdrawing from any subsequent action. The current study failed to find any evidence that a cheap verbal signal from one individual altered the punishment behaviour of participants, and while there was some evidence suggesting participants disliked false-signallers, their response to Alex was primarily driven by his behaviour alone. Furthermore, the study did not find any evidence that a deceptive signaller would actually be punished for their actions.

3.10.1 Signal of punishment and participant behaviour

In response to some of the limitation of Study 1, the first question to address is whether participants responded to the experimental manipulation. All participants included in the analysis of Study 2 successfully passed the manipulation check, and importantly, seemed to respond to the initial set of questions as one would predict: as shown in Figure 3.5 they felt far more supported by others when Alex signalled he would assist in the punishment of Charlie. However this did not affect their punishment behaviour, for which the only significant predictor was the anger felt at the defection. This support the finding that emotional response is the strongest predictor of punishment behaviour (Falk et al., 2005), which itself may help explain why there tends to be 'over punishment' when multiple individuals can punish (Peter et al., 2010); individuals are more concerned with ensuring a defector is punished than with effective coordination. And while coordination can help lower the costs of punishment, others have shown that costly punishment can be both effective and evolutionarily stable even if there is no coordination between punishers (Bowles & Gintis, 2004). After all punishment is seen as a 'human universal' (Fehr & Gächter, 2002) driven by the 'moral outrage' that acts of defection induces; perhaps participants 'would have confronted Charlie anyway' and thus paid little attention to this signal alone.

For the results of the Boyd (2010) model to be fulfilled, individuals would have to respond to relatively-cost free signal by another. However, participants in this study simply did not respond to the verbal (potentially cost-free) signal from Alex when making their punishment decisions.

3.10.2 Response to honest or deceptive signalling

The study found some evidence that participants responded negatively to false signalling: when Alex signalled a willingness to punish and then refused, he was disliked more than when he honestly signalled his intention not to help. Indeed, participants were surprised when

Alex went against his word which, if nothing else, demonstrate that participants had attended to the second piece of information about Alex's behaviour in the vignette.

However, despite false signalling having a negative effect on social attitude to the signaller, participants were primarily concerned with the signaller's actual behaviour: whether Alex engaged in punishment with them or not. Participants demonstrated significant differences in attitude and emotional response to Alex depending on whether or not he joined them in the actual confrontation. This potentially indicates a like-attracts-like property amongst punishing individuals (Albert et al., 2007; Skyrms & Pemantle, 2000) with those who punish being liked far more by individuals who also punish. While potentially in a future situation the presence of a known punisher would encourage others to also join in, this can be considered a secondary effect of a separate motivation to punish.

This may be why participants did not respond to the verbal signal from Alex: given the costs involved in actual punishment behaviour, from either the production of that behaviour (Frank, 1996) or the risk of retaliation (Dreber et al., 2008; Nikiforakis, 2008), only actually engaging in punishment may be (potentially) costly enough to signal something about the individual. Individuals do, for example, pay close attention to the actual cost of any punishment (Nelissen, 2008) and there is a is a strong connection between costly punishment and cooperative behaviour (Barclay, 2006; Lehmann et al., 2007). Furthermore, when punishment is very costly, only moralistic punishment of free-riders occurs (as opposed to a mix of anti-social, spiteful and random punishment, Falk et al., 2005). Therefore, only by actually punishing can an individual send an honest signal of a commitment to fairness and pro-social social norms. In fact, Duffy and Feltovich (2002) demonstrated that when there is any ambiguity in the motives of a signaller, for example here perhaps an attempt to deceive the receiver (the participant) into punishing, receivers will rely on past actions alone. Thus, only

a reputation for actually engaging in punishment acts as a predictor of future social behaviour: actions speak louder than words.

3.10.3 Punishment of a deceptive signaller

While participants responded far more to the actions of Alex, they did demonstrate some reactions to the interaction of signal and behaviour, with Alex being significantly more disliked when he signalled a willingness to assist in punishment, but subsequently withdrew that support. This negative response in social attitude demonstrates a potential cost of the false signal as, for example, there is a strong association between (a lack of) positive attitude to someone, trust in them and cooperation (Albert et al., 2007; Gächter et al., 2004; Rotter, 1980). While not tested explicitly in this study, the negative attitude may have led to future social ostracism of Alex by participants. Indeed, individuals tend to self-assort in a 'like-attracts-like' fashion when given the opportunity to adjust their social ties (Fehl, van der Post, & Semmann; Rand, Arbesman, & Christakis, 2011; Santos et al., 2006), and so the cost to Alex's deceptive signalling may be in the form of the loss of long term social opportunities rather than in a direct response to his action.

Regardless of any long-term social costs, what the study did not find was any direct retaliation against Alex for any combination of signal and behaviour: participants overwhelmingly did not take advantage of the opportunity to punish Alex by withdrawing cooperation. An explanation for this is offered by the qualitative information collected at the end of the study, as participants did not want to violate wider norms of cooperative behaviour (Moshagen, Hilbig, & Musch, 2011). While no specific method of analysis was applied to this data, in the majority of cases no deeper investigation was needed as participants made quite definite statements, such as wishing to be seen "*as the better person"* and the desire not to be caught "*stooping to their level*". This suggests that participants were more concerned with their reputation as a 'nice', rather than antagonistic, person.

An alternative explanation for the lack of a response to Alex's false signalling may be the threat of retaliation from Alex. Opportunity for retaliation significantly curtails punishment behaviour (Janssen & Bushman, 2008; Rand et al., 2010), and long-term vendettas are a feature of human societies regardless of how such vendettas started (Topalli et al., 2002; Zizzo & Oswald, 2001). Thus participants may have either feared a response from Alex or at least wished to avoid a string of tit for tat punitive actions. Conversely, the lack of overt punishment from participants may be due to a *lack* of retaliation in the vignette: the scenario did not describe what became of the initial confrontation, and participants who received the deceptive signal might have responded far more negatively to being 'tricked' had the punishment resulted some form of retaliation from Charlie, i.e. they had they paid an explicit cost for being abandoned by Alex.

3.10.4 Trait-Dominance

When the response to the initial (i.e. Charlie's) defection was considered, dominance appeared to have an important impact on social decision-making. This is interesting in itself as it suggests that while anger at a defection may be important in driving punishment behaviour (Falk et al., 2005), there are other factors that curtail actual punishment behaviour. We suggested in Study 1 that one reason participants' preferred to punish anonymously, despite being angry, was an unwillingness to bear the reputational marker that engaging in punishment might provide. Here, dominant individuals were more willing to confront Charlie about his behaviour, with dominance being a greater predictor of this confrontation even than anger.

This result is interesting as it might suggest that costly punishment behaviour may be primarily carried out by dominant individuals. This is important as the main results of the current study suggest that the costs of punishment cannot be reduced by simply signalling a desire to punish, and a dominant position provides an individual with numerous ways to

punish at a lower cost without the need for the involvement of others. For example, dominant individuals possess greater resources, are surrounded by individuals willing to provide assistance in exchange for contact (Schino & Aureli, 2009) and, as we are unwilling to challenge more dominant individual regardless of their behaviour (Egas & Riedl, 2008; Henrich & Gil-White, 2001; Kim et al., 1998), being dominant might lower the cost of retaliation. If nothing else, dominant individuals enjoy a freedom of action not open to others and consequently the defection of others from an agreed action is not so great a risk. While this is, for the moment, purely speculative, it does hint at an alternative explanation for the emergence of costly punishment in humans, one decoupled from enforcing cooperation (Rand et al., 2010).

3.10.5 Limitations

One limitation of this study is that participants were only asked whether they would confront a social defector (Charlie), and not how much punishment they would inflict on him. This procedure was chosen as, as also seen in the results of Study 1, asking for actual punishment 'amounts' might not be appropriate for a questionnaire/vignette method. In an attempt to overcome this limitation, the decision was taken to remove the idea of quantifying punishment by making the situation more 'social', i.e. describing the sort of low-level norm violation likely to occur in everyday life with the potential for the informal peer-sanctioning that economic experiments, for example public goods games, try to simulate. The vignette in the current study was developed to represent a common action problem (Hardin, 1968) familiar to many students and non-students alike: the use and maintenance of communal space.

Indeed, a related limitation is that there was no actual 'punishment' inflicted on Charlie and as such any findings cannot be compared to the results of studies using explicit costs and effects of punishment. However, again, we would argue that many of the costs to such

behaviour, either to the punisher or the target, are likely to be expressed socially (Van Vugt, 2006), for example through humiliation (Barr, 2001), rather than in the form of physical violence. In fact, 'non-monetary' or 'verbal' punishment is seen as punishment even in the economic literature (for example, Masclet et al., 2003; Ostrom et al., 1992).

Nevertheless, while we believe this means the results of the study can be considered alongside other methods, as opposed to being artefacts of the experimental design, we accept that when participants are not forced to pay a physical cost for their intervention, they believe they would be more willing to engage in punishment and pay the costs of such punishment then they actually would (Pedersen et al., 2013).

3.10.6 Conclusion

The study attempted to investigate whether a signal of a willingness to punish from another individual would affect the punishment behaviour of participants, and whether this signal could be considered honest due to the penalties inflicted upon deceptive signallers. The study found little or no effect of the signal or its honesty on either participant behaviour or their opinion of the signaller: how this signaller acted was seen as far more important than the promises they made. The finding suggests that participants make judgements of others based only on their willingness to engage in actual costly behaviour, and suggests that only a history and reputation for actual punishment behaviour might affect any future decision making. This suggests that any theory trying to explain the evolution of costly punishment will have to explain how any one individual can bear, or otherwise recover, the costs of the behaviour. The results for trait-dominance may suggest that dominance and status could be one such factor that may explain this.

3.11 General discussion

The initial aim of the chapter was to assess two possible mechanisms by which the net cost of costly punishment could be reduced; by adjusting punishment behaviour in response to the potential for reputational gain, or by coordinating punishment with another individual. No evidence for either mechanism was found. In Study 1 participants did not adjust their punishment behaviour in response to an audience, and participants actually preferred to punish when groups were unstable. In Study 2 participants did not adjust their punishment behaviour in response to the punishment-signal of another individual, nor did they punish an individual for deceptively signalling their intent to engage in punishment.

The results of Study 1 do contradict some research in this area (Bering, 2008; Kurzban et al., 2007) but can be seen as partially supporting other research that suggests individuals are actually happy to punish anonymously (Fudenberg & Pathak, 2010) and will pay additional costs in order to hide punishment (Rockenbach & Milinski, 2011): individuals, it seems, want to punish the social defection, but only when the chances of meeting the target were low. These results suggest that participants were not taking into account the possibility of gaining a positive reputation for the act of punishment and we theorised that instead they may have been reacting instead to the possibility of gaining a *negative* reputation which, specifically, might invite future antagonism or retaliation from the target (See Dreber & Rand, 2012; Nikiforakis, 2008).

The results of Study 2 did not support the model of Boyd et al (2010) who suggested that punishment could evolve if punishers could coordinate. If this had been the case we would have expected participants to be sensitive to signals from others, or indeed to their own feelings of being supported by others, when making their punishment decisions. This mistrust of any signal by another might be explained by participant's unwillingness to punish deceptive signalling. While there may be downstream costs (for example ostracism, Masclet,

2003) for this deception, a signal of a willingness to punishment alone was not seen as honest.

3.11.1 People like punishers

While the studies did not demonstrate that participants varied their punishment behaviour in order to most effectively generate a positive reputation, both did support the idea that punishers do gain indirectly from an act of punishment (Nelissen, 2008; Sigmund et al., 2001); participants really liked individuals who punished. In Study 1, there was a strong relationship between how angry participants were at a social defection and how much they liked a punishing other, and in Study 2 participants based their opinion of another punisher solely on this action and not how they had previously signalled their behaviour.

Importantly, the results from Study 1 suggest this is not just a result of 'like attracting like', as the positive opinion was related to anger rather than actual punishment behaviour. Instead this result supports the theory that engaging in punishment acts as a costly signal of other pro-social or other regarding tendencies (Nelissen, 2008); indeed participants in Study 2 only responded to Alex based on his behaviour, i.e. when he had displayed the costly punishment signal. There are good reasons why individuals would wish to associate with a punisher: environments where punishment occurs possible are more cooperative and efficient (Fehr & Gächter, 2000; Gächter et al., 2008; Rockenbach & Milinski, 2006) and so associating with someone who is willing to uphold cooperation norms would be beneficial. Equally, signalling a concern for the welfare of others provides any target of aggression with a tacit ally; the presence of social allies reduces the perceived threat of an opponent (Fessler & Holbrook, 2013) and Kim (1998) demonstrated that in the presence of (in their terminology) a 'justice-minded' third party, low status victims were more willing to resist antagonism from higher status individuals. Given that anger at social norm violations might be a result of fear (Jenson & Peterson, 2011; Peterson, 2012), and that unfairness or antagonism is likely to be instigated

by high status or formidable individuals (Clutton-Brock & Parker, 1995; Griskevicius et al., 2009; Piff et al., 2012; Sell, Tooby, et al., 2009; Silk, 2003), for many a punisher would be especially welcome as an associate.

3.11.2 Dominance equals more likely to punish

This does however raise a certain paradox in the data. As shown by previous research, the actions of a punisher generated a positive reputation, but in Study 1 participants responded in a manner that would indicate they wished to avoid any such reputation. The results of Study 2 may suggest a possible reason for this, as here punishment was associated with dominance.

This is important for two reasons. Firstly, engaging in any sort of antagonism, even 'moralistic' aggression, can be seen as a dominant act (Clutton-Brock & Parker, 1995; Jones et al., 2011); and one possible indirect benefit of engaging in punishment might be a reputation for formidability (Barclay, 2006). Participants may not have wished to gain this sort of reputation due to the potential for future antagonism (Tibbetts & Izzo, 2010). A more important reason however is threat of retaliation. Retaliation might be the primary cost to costly punishment (Dreber & Rand, 2012) and experimentally it dramatically curtails punishment (Nikiforakis, 2008). While revenge may be a dish best served cold, in terms of ancestral human informal peer-sanctioning, the response to punishment is likely to be immediate (for a modern example, see Levine et al., 2011). Success in conflicts is often determined by dominance and formidability (Clutton-Brock & Parker, 1995; Maynard-Smith & Parker, 1976; Sell, Tooby, et al., 2009) and so, while reputation may offset the cost of punishment indirectly and in the long-term, this will only occur if the punisher survives the attempt at punishment itself.

In fact this may be why individuals might not wish to gain a reputation as a punisher, as such a reputation for formidability and dominance will simply invite aggression from others

wishing to challenge them. Equally, as mentioned in 3.11.1, there may be negative social repercussions from falsely *recruiting* allies by engaging in costly punishment when one has no intention and/or ability to actually punish in the future.

3.11.3 Abandoning the Dark Triad

As discussed in 3.2.3 there has been little research conducted into the effect that the 'anti-social' personality traits of the 'Dark Triad' have on punishment behaviour. Costly punishment seems to be drive by anger (Falk et al., 2005) and can be considered an impulsive act (Crockett et al., 2010). Therefore, due to deficits in empathy and impulse control that individuals with high Dark Triad trait exhibit (Paulhus & Williams, 2002), it was reasonable to investigate whether these individuals differences might have affected punishment behaviour. However the studies in the current chapter did not find any evidence that the Dark Triad personality traits affected punishment. While a lack of an effect of the Dark Triad on punishment could be due to the use of vignettes (see 3.6.5 and 3.10.5), the results of the other variables tested did conform to results gathered by economic games, for example the attitude to a punisher. Due to the lack of any effect therefore, the Dark Triad line of enquiry was abandoned.

3.11.4 General conclusion

The two studies in the current chapter did not find evidence that punishers respond to the opportunity to gain a positive reputation or responded to an attempt to coordinate punishment. However, the pattern of results seen in Study 1 & 2, has given rise to a potentially more interesting, and un-researched, question; the role that dominance and status might have played in the evolution of costly punishment.

4 Chapter 4: perceptions of costly punishers

Chapter 3 revealed that punishers are well liked, and that dominance might be an important factor in punishment behaviour. Building on these results, Chapter 4 investigates the potential indirect benefits of punishment through Costly Signalling (see 1.2.2 and 2.3.4), and specifically whether punishment can signal dominance. Study 3 investigates whether participants judge a punishers differently compared to those who engage in other forms of confrontational behaviour in terms of their likability and dominance. Study 4 investigates whether acts of punishment are used to make dominance-rank judgements about those involved, and whether the success of punishment and the risks posed to the punisher by an aggressor affect the social judgments of the punisher.

4.1 General Introduction

Punishment has been consistently shown to be one of the main factors that ensures cooperation between groups of individuals (Balliet et al., 2011). Costly punishment is effective at promoting cooperation even if it is delayed (Fudenberg & Pathak, 2010) or given in a verbal form only (Masclet et al., 2003), and the mere presence of a third party significantly increases both fair behaviour and, conversely, the unwillingness to accept unfair behaviour (Kim et al., 1998). The punishment of anti-social or 'unfair' others has also been claimed to be a universal human behaviour (Fehr & Gächter, 2002) and the desire to punish seems to be an automatic response (Crockett et al., 2010). Nevertheless, there is continuing debate as to how punishment and the associated moral sentiment could initially evolve because it imposes costs on the punishing individual while the benefits are shared amongst the group as a whole (Dreber et al., 2008).

4.1.1 Reputation and costly punishment

This puzzle could be solved if there was some way for punishers to recuperate the costs of punishment through indirect benefits from their actions (Panchanathan & Boyd, 2004; Santos, Rankin, & Wedekind, 2010). One such indirect benefit might be a reputation as an honest and trustworthy person. It has been suggested that punishment might act as a costly signal; where one engages in a risky or otherwise energetically or materially costly behaviour to demonstrate an otherwise unobservable trait (Bird et al., 2001). Here, bearing the cost of punishment demonstrates, for example, you are an honest person who values fairness (Nelissen, 2008). Punishment as a signal for this trait does seem to be accurate as generally those who punish are indeed also highly cooperative, especially when punishment is very costly (Falk et al., 2005). Costly punishment therefore signals an individual is trustworthy and making such a signal so could allow other such individuals to self-assort with one another and enjoy the cooperative benefits this allows (Santos et al., 2006; Wang, Suri, & Watts, 2012).

More speculatively, in that it has not been explicitly tested, a reputation as a punisher may also help recruit cooperative partners or coalition members (something potentially vital in our evolutionary history, Gavrilets et al., 2008) as, for example, individuals do prefer to be in an environment where punishment might occur (Gürerk et al., 2006; Rockenbach & Milinski, 2006). In fact, individuals are willing to pay far above what would be an efficient amount to maintain an environment where punishment occurs (Traulsen et al., 2012). This might be because while individuals wish to cooperate, they fear the cost of defections (Fischbacher et al., 2001) and one of the more direct proximate motivations of punishment is to change the behaviour of the defecting individual (Fudenberg & Pathak, 2010) or to remove them from the group altogether (Bowles & Gintis, 2004; Masclet, 2003). By taking action to prevent free-riding, punishers provide an environment safe from such threats and are therefore seen as

people worth associating with and/or following: it is unlikely to be coincidence that the traits attributed to punishers, such as trustworthiness and being group-focused, are also those demanded of leaders (Hogg et al., 2012). With this in mind the results of Rockenbach & Milinski (2006), that individuals prefer an environment where punishment is possible, could be reinterpreted to suggest individuals prefer to be in an environment where *someone* will punish social defection. Thus a reputation as a punisher allows the punisher to recruit social allies more effectively because, as well as signalling their own altruistic and cooperative tendencies, it may also suggest they will intervene to ensure any individual in their vicinity is treated fairly and that any defectors are removed.

Alternatively, engaging in punishment might be less about signalling pro-social personal characteristics and more about signalling personal formidability. Firstly, costly punishment can be considered a confrontational act that at some point must, by definition, involve an individual inflicting a cost upon a defector or aggressor, and most antagonistic actions are instigated by dominant individuals (Clutton-Brock & Parker, 1995; Silk, 2003). Indeed, those who feel high status or formidable are far more willing to both use and approve of the use of force (Sell, Tooby, et al., 2009). Equally, such aggressive actions are used by dominant individuals to maintain their position (Silk, 2003) and, while non-human animals show only limited evidence of 'altruistic' punishment, the few examples of punishment in the non-human literature are conducted by dominant individuals only (Flack et al., 2005; Flack et al., 2006; Wong et al., 2007), with the apparent purpose of maintaining their social rank. Thus punishment could be another form of aggression used as a signal of position and to demonstrate personal formidability (Barclay, 2006). In fact, Marlowe et al (2008) suggested that one reason for the lack of costly punishment in small scale societies is that, due to eavesdropping on dyadic interactions, a "don't mess with me" reputation can be easily established without an individual involving themselves in the conflicts of others.

It should be noted that the reputation gained from an act of costly punishment need not only be either as a fair and trustworthy person or as a formidable person; it could be both. For example, research on welfare trade-off ratios, the process by which we make resource allocation decisions are made (see Sell, Tooby, et al., 2009), splits the factors in this process into two broad categories: the potential benefit the recipient provides to us, and their ability to inflict costs upon us. Thus an act of punishment would provide social gains to a punisher because, on the one hand, they are seen as beneficial to be around as their actions indicate they are trustworthy and are willing to defend group norms and eliminate free-riders, and on the other hand they have signalled their individual formidability or willingness to use force and thus should be treated fairly or even with deference.

In summary, an analysis of the current literature suggests that engaging in costly punishment could act as a signal of both pro-social personal characteristics and personal formidability/dominance. These indirect benefits could provide a means by which punishers recuperate the cost of punishment. However, this will only be the case if a) punishers are indeed judged to be both formidable and likable, b) if such judgements in response to costly punishment specifically, as opposed to any individual who is victorious in a conflict generally, and c) if observers do make rank dominance judgments while observing punishment. To investigate these questions, the current studies used a vignette-based method to measure the social judgments made by uninvolved observers about individuals who engage in costly punishment.

4.2 Study 3: both loved and feared: costly punishment is perceived differently from other agonistic behaviour

This exploratory study investigated whether observers do in fact make judgements about the likeability and dominance of an individual after observing them engage in punishment. More specifically, the study investigated how judgements about punishers are different from

judgements about individuals who engage in other types of aggressive behaviour, i.e. whether any judgements of dominance or reputational benefits are related to the punishment itself or, more generally, to an effect of aggression and/or winning a physical contest.

4.3 Method

4.3.1 Participants

414 (132 male) undergraduate students from the University of Exeter, UK, successfully completed the survey. Participants were recruited via email using an existing 'paid participant' list. As an incentive to take part, any participant who completed the survey was entered into a prize draw for a number of store vouchers worth £10 (about US$13). The mean age of participants was 22 years. 25 participants failed manipulation check questions and their data was excluded from all analyses. Recruitment took place between October 2011 and February 2012.

4.3.2 Materials and procedure

The survey was administered online. Participants followed an email link and were presented with a survey consisting of two sections. The first section contained the experimental vignette, presented as a news website-style article. To keep with the 'news site' aesthetic and the wider aims of the study, the article included a picture of its subject, a male identified only as 'John Taylor'. The name was fake but the picture was chosen from a set of photos collected for a previous study (Gordon & Platek, 2009) as the face received neutral ratings in regards to attractiveness and trustworthiness. Once participants had finished reading the article they were presented the second section of the survey which contained a series of questions concerning John (for the full vignette, see Appendix B).

4.3.3 Experimental Scenario

Participants were presented with one of four possible articles concerning the actions of John. In the Third Party Punishment condition, John was described as having successfully intervened to stop the mugging of an old man late at night - 'third party' is used here as John can be considered to be 'disinterested; in the Second Party Punishment condition, John was described as having successfully fought off a mugger late at night; in the Bar Fight condition, John was described as having been involved in a bar fight of indeterminate cause, although it was made clear that alcohol was not involved and that John 'won' the fight; and in the Control condition John was described as having witnessed a flash-mob. In all three experimental conditions the assailant who fought John was described as "a 6ft muscular male".

The scenarios also manipulated the formidability of John. John was described as a "keen amateur boxer" (strong), as someone who "had never been in a fight in his life" (weak) or was given no additional description (neutral). Thus, study had a 4x3 between-subjects design. As the 'weak' description would have made no sense in the 'control/flash mob' condition, it was changed to "on his way back from a beauty salon". While the author acknowledges the stereotyping this description represents, the stereotype did produce a response consistent with the other 'weak' conditions.

4.3.4 Likability and dominance questions

Participants were asked the same set of questions as in Studies 1 and 2 regarding how likable John was perceived to be. They were asked to rate John on a scale of 1 (strongly disagree) to 7 (strongly agree) as to how trustworthy, group focused, 'nice' he was and whether they would work and socialise with him. In the current study the five items had a high reliability index (α=0.91). Therefore they were collapsed into a single 'likability' variable for all future analyses. While there was no reason to believe sex would affect likeability per se, a separate

analysis found that the sex of the participant did not affect likability or interact with the types of scenario presented to participants. Therefore sex was not included in the analyses below to conserve power.

Male participants then answered a further set of questions concerning how dominant they perceived John to be, on a scale of 1-7 (1=strongly disagree, 7=strongly agree), on how threatening, intimidating, dominant, antagonistic or aggressive he was. In the current study the five items had a high reliability index (α=0.86) and were therefore collapsed into a single 'perceived dominance' variable for all future analyses.

As part of the wider aims of the study, female participants (n=282) were asked questions concerning their willingness to be romantically involved with John (these data are not reported here). In order to keep the questionnaires to a similar length for both sexes, females were not asked to judge John for perceived dominance.

4.4 Results

4.4.1 Likeability

As shown in Figure 4.1, John was seen as more likable in the Third Party Punishment condition than in the other conditions ($F_{3,399}$=36.72, p<0.001). John in the Bar Fight condition was the least liked. Bonferroni-corrected pair comparisons found significant differences (all p<0.001) between all comparisons of Article-types except between the Control and Second Party conditions (p=1.0). As shown in Figure 4.2, John was seen as more likable when he was depicted as 'weak' ($F_{2,399}$=4.40, p=0.013). Bonferroni-corrected pair comparisons found that the weak John was seen as more likeable than neutral John (p=0.031), however there were no significant differences between weak and strong John (p=0.45) or strong and neutral John (p=0.99). Likeability was not affected by an interaction between the article type and John's formidability ($F_{6,399}$=0.65, p=0.65).

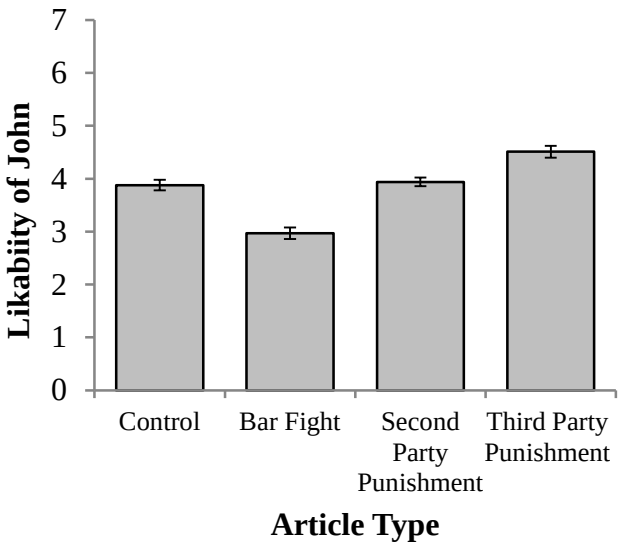


Figure 4.2: likability of John across different antagonistic encounters. Bars = 1 Standard Error.

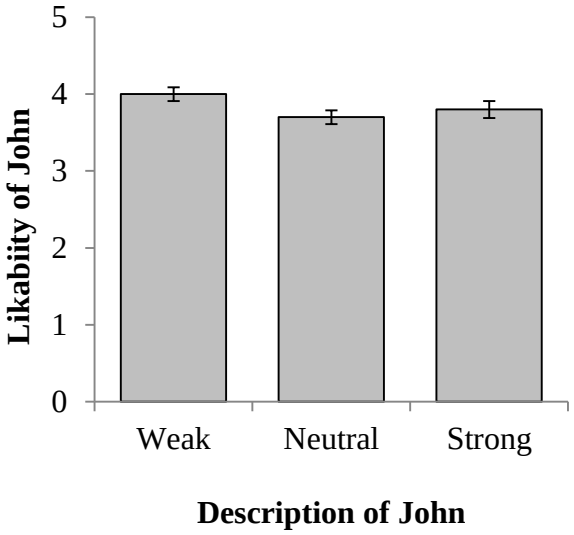


Figure 4.1: likeability of John depending on how formidable he was described as being. Bars = 1 Standard Error.

4.4.2 Perceived dominance (male participants only)

As shown in Figure 4.3, John was judged as more dominant in all the experimental articles compared to the Control condition ($F_{3,120}$=6.15, p=0.001). Bonferroni-corrected pair comparisons found significant differences between the Control Article and the experimental conditions (Bar Fight, p=0.003; Second Party Punishment, p=0.009; Third Party Punishment, p=0.011), but no differences in judgments of dominance between the three experimental conditions (all p=1.0). John's described formidability did not affect how dominant he was seen to be ($F_{2,120}$=0.45, p=0.64), nor was there an interaction effect of article type and formidability on perceived dominance ($F_{6,120}$=0.83, p=0.54).

4.5 Discussion

These results show that the increase in likability of punishers cannot be explained alone by them winning an altercation or by the 'warm glow' that may accompany seeing an offender receive retribution (de Quervain et al., 2004; Singer et al., 2006). This is because when John fought off his own attacker, he was seen as no more likable than in the control article where John did nothing. This is probably because second party punishment is driven more by a desire to protect oneself, or for personal retribution and to save face (for example, Topalli et

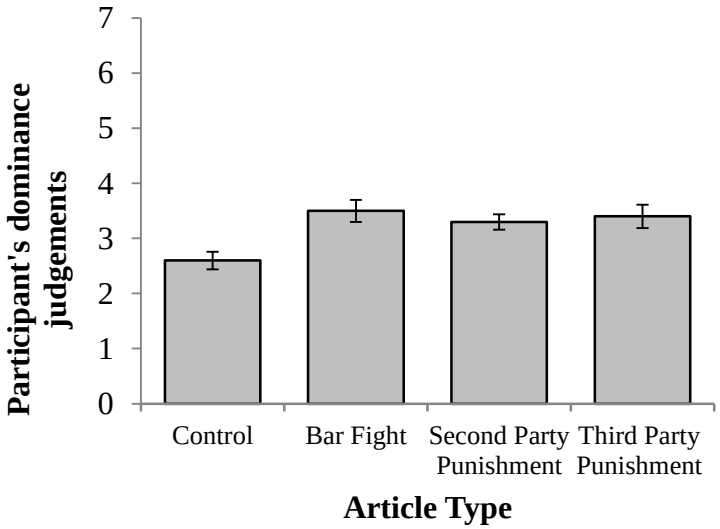


Figure 4.3: perceived dominance of John across different antagonistic encounters. Bars = 1 Standard Error.

al., 2002) and as a result is far more common than costly punishment (Fehr, 2004). Thus defending yourself says little about your qualities save your ability to fight back. This suggests that, in opinion of observers/receivers, engaging in costly punishment is seen as signalling additional information about the punisher, such as their commitment to fairness or as being someone with whom it may be worthwhile to associate. Indeed, while there can be sex differences in how violence is perceived (Griskevicius et al., 2009), in the present study both males and females made similar judgements about the likability of John.

Judgements about dominance were however dependent solely on the aggression in the encounter rather than on the context, i.e. John was seen as equally dominant whether he acted as a third-party or was involved in a fight with an indeterminate cause. This is unsurprising as engaging in aggressive behaviour can signal dominance (Silk, 2003) and perceiving dominance from an interaction can be seen as a reasonably objective process; it is in our interests to make accurate observations of the social hierarchy (Cummins, 1996a) and the outcome of a confrontation can be easily recognised (Jones et al., 2011). It is telling that when judging dominance, the participants did not take John's reported formidability into account; their judgement was based entirely on the outcome of the conflict. The dominance

data came from male participants only, however for the social-cognition reasons mentioned above, and because it has been shown that males and females agree on male formidability (Sell, Cosmides, et al., 2009), it is unlikely there would have been sex differences in dominance judgments in this study.

By comparing the judgements of a punisher to other aggressive acts, this study demonstrated that engaging in costly punishment specifically provided the punisher with positive reputational benefits. This study also demonstrated, in males at least, that engaging in punishment can make one seem more dominant without the negative social consequences associated with other forms of aggressive behaviour, i.e. that third party punishers are not only seen as formidable, they are also well liked.

4.6 Study 4: perceptions of a third-party are affected by their attempt at punishment and not its success

Study 3 found that third party punishers are judged to be more likable than individuals who engage in other aggressive behaviours, yet are seen as equally dominant as individuals who engage in other aggressive acts. Study 4 investigated whether observing punishment affects the perceived dominance rank of the individuals present in the interaction, i.e. if punishment can signal a dominant position relative to others. Study 4 also investigated what information observers are using to judge punishers, specifically whether judgements are affected by the success of the intervention and whether the level of threat an aggressor posed would further affect a participant's perceptions of the punisher. For this study, 'third party' is again used to describe actions of the punisher as they are 'disinterested'.

4.7 Method

4.7.1 Participants and materials

103 psychology undergraduate psychology students from the University of Exeter (85 females) successfully completed the study, with 12 participants either failing the manipulation checks or dropping out of the study before completion. Participants were recruited via email from the 1st year psychology cohort. As an incentive to take part, any participant who completed the survey was entered into a prize draw for a number of online-store vouchers worth £10 (about $13 US). The mean age of participants was 21. The study employed a between-subjects design with 3 experimental conditions and one control condition; participants followed an email link which randomly presented with one of four experimental vignettes, followed by a series of questions concerning the punisher in these vignettes. The study was conducted between October and December 2011.

4.7.2 Experimental Scenario

Participants were asked to imagine themselves seated alone in a local bar and told that they observed a group of men enter and occupy a table nearby. Participants were then told they observed an altercation between group members in which one member (the 'aggressor')[2] forced another (the 'victim') to relinquish his seat so the Aggressor could sit down. In Condition 1, the Successful condition, a third group member (The 'third party') successfully intervened and forced the aggressor to give back the seat. In Condition 2, the Unsuccessful condition, the third party intervened but failed to force the Aggressor to give back the seat. In Condition 3, the Increased Threat condition, participants were told they observed a successful act of punishment, but in this scenario the male characters were unknown to one another and not part of a self-contained group. Thus the Aggressor was a greater potential threat to the

[2] These labels are for clarity only; in the scenario itself the characters were identified by the colour of the shirts they were described as wearing.

Third Party and the participant/observer. This Increased Threat condition matched the Successful condition in all other respects. In Condition 4, the Control/No Action condition, participants were told they observed a Third Party become agitated but not intervene (for the full vignette, see Appendix B).

4.7.3 Likability and dominance

Participants were then asked to make a series of social judgements about the Third Party in the scenario. Firstly, participants were asked to rank the three characters in the story in terms of dominance (1 being most dominant and 3 being least dominant). All participants were then asked the five likability questions (α=0.88) and the five social dominance questions (α=0.85) as described in Study 3. As in Study 3 these items were collapsed into a single 'likability' and 'dominance' variable respectively for all future analyses.

4.7.4 Manipulation checks and demographic questions

Participants were then asked a comprehension/manipulation check question. They were asked to indicate, from a choice of "made the man-in-grey [the aggressor] move", "attempted but failed to make the man-in-grey move" or "did nothing" how the third party behaved in the scenario. Finally, participants indicated their age, sex and nationality.

4.8 Results

The study tested two distinct hypotheses: that there would be a relationship between how participants responded to a third party depending on their level of intervention (Successful vs. Unsuccessful vs. Control), and that there would be a difference in participant responses between the level of threat posed by the aggressor (Successful vs. Increased Threat). Data relating to these hypotheses were analysed separately.

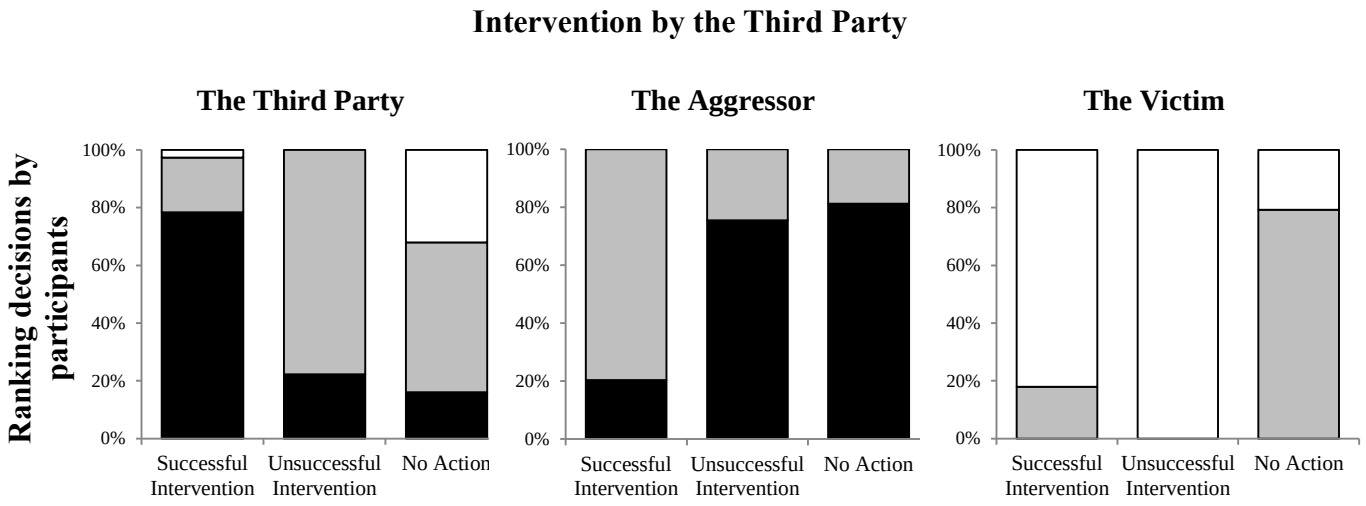


Figure 4.4: proportion of participants who, across conditions, ranked the Third Party, the Aggressor and the Victim as the most dominant character (Black bars), gave the character the middle rank (grey bars) and as the least dominant character (white bars).

4.8.1 Relative dominance rank of the third party

Participants ranked the third party to be most dominant when he successfully intervened, with fewer ranking him as most dominant when the intervention failed, and the fewest when he did not intervene. The victim was nearly always ranked as least dominant (See Figure 4.4). To investigate the relative difference between the characters, we considered which character was ranked as the most dominant by participants. Figure 4.5 shows that the third party was ranked as the most dominant when punishment was successful but not when intervention failed or

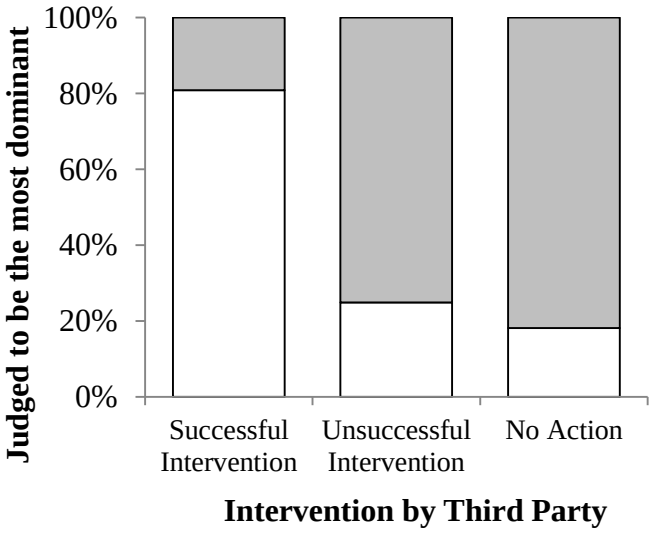


Figure 4.5: proportion of participants who perceived the Third Party (white) or the Aggressor (grey) to be the most dominant character in each condition.

when he took no action (χ^2_2=28.75, p<0.001).

4.8.2 Perceived dominance of the third party

As shown in Figure 4.5, the third party was judged to be more dominant when he attempted to intervene ($F_{2,77}$=7.88, p<0.001). It was assumed a priori that success would affect perception of dominance, however planned contrast analyses (Successful vs. Unsuccessful, $F_{1,78}$=1.65, p=0.20; Successful vs. No Action, $F_{1,78}$=14.30, p<0.001; Unsuccessful vs. No Action, $F_{1,78}$=4.06, p=0.047) demonstrated that the third party was seen as more dominant when he intervened, regardless of his success.

4.8.3 Likability of the third party

Figure 4.6 also shows that the Third Party was judged to be more likeable when he attempted to intervene, regardless of whether or not he was successful, than when he did not intervene ($F_{2,78}$=4.70, p=0.01). It was also assumed *a priori* that success would affect likability, however planned contrast analyses demonstrated that the third party was seen as more likable when he intervened, regardless of the success (Successful vs. Unsuccessful, $F_{1,78}$=0.15, p=0.70; Successful vs. No Action, $F_{1,78}$=7.27, p=0.009; Unsuccessful vs. No Action, $F_{1,78}$=6.40, p=0.01).

4.8.4 Judgements of the third party and the threat posed by the aggressor

As shown in Figure 4.7, the third party was judged to be more socially dominant when the threat posed by the Aggressor was increased ($F_{1,57}$=4.56, p=0.037). However, the level of threat did not affect how likable the Third Party was judged to be ($F_{1, 57}$=0.11, p=0.75).

4.9 Discussion

Here, the results concerning the judgements of dominance are unequivocal; successful intervention by the third party led participants to perceive him as most dominant, and unsuccessful intervention led to the aggressor being perceived as most dominant. This result

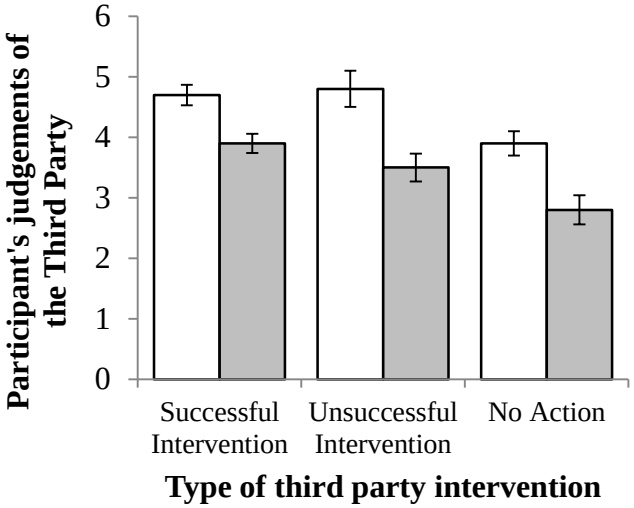


Figure 4.7: judgement of likeability (White) and dominance (grey) of the Third Party depending on the Third Party's response to an act of aggression. Bars = 1 Standard Error.

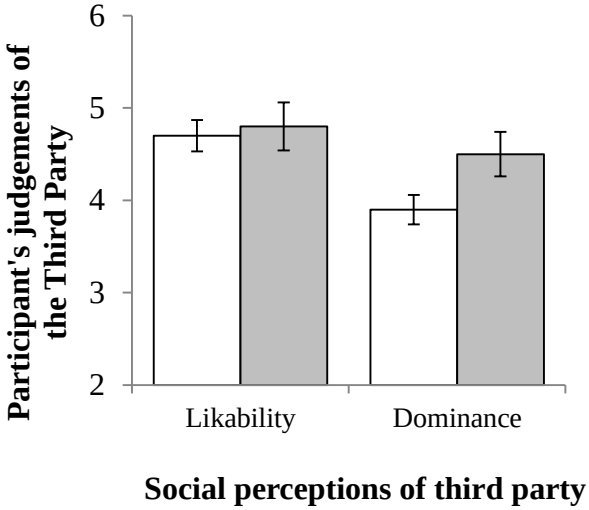


Figure 4.6: participants' judgements of the Third party between the Successful Punishment (white) and Increased Threat (grey) conditions. Bars= 1 Standard Error.

is no surprise as such, for example Jones et al (2011) found that humans eavesdrop on dyadic interactions to make dominance judgements and transitive reasoning is a core part of primate social cognition (for a review, see, Cummins, 1996a). However to the author's knowledge it is the first study to experimentally demonstrate that engaging in punishment directly affects the perceptions of an uninvolved observer with respect to the punisher in this manner.

What is surprising is that when asked to make social judgements about the punisher, participants judged him to be more likeable and dominant when he intervened, regardless of the success of the intervention. While previous studies, including Study 3 in this chapter, have demonstrated that punishment leads to reputational gains (Barclay, 2006; Nelissen, 2008), in these studies (as was the case in the Study 3 scenario), punishment has always been successful by design; that is to say the action of punishment always results in the actual imposition of costs on the target, something which is not guaranteed outside of the laboratory (Levine et al., 2011). The fact that perceived likeability and dominance remained even when the intervention was unsuccessful suggests that such ratings are not due to a halo effect of seeing an antisocial individual punished (de Quervain et al., 2004; Singer et al., 2006) or due

to the punisher being the recipient of indirect or strong reciprocity for carrying out a public function.

The results do however add further evidence to the suggestion that punishment can be seen as a costly signal of an altruistic nature. Due to the threat of retaliation (Nikiforakis, 2008; Rand et al., 2010), the mere act of engaging in costly punishment should provide an honest signal, as retaliatory costs will likely be present whether the intervention was successful or not (see Chapter 5). In Chapter 3, participants only responded to actual punishment behaviour, and the results of Study 4 imply that even if one fails to punish successfully, the attempt suggests to eavesdroppers that it could still be useful to form a punishing coalition with the failed punisher. That punishment is a costly signal was further highlighted by the higher dominance rating given to the punisher in the Increased Threat condition; the lack of any social information or social support from fellow group members made the risks even higher and thus the signal more reliable.

However, there was no corresponding increase in likability in the higher threat condition. Nelissen (2008) suggested that increased signal reliability should increase the positive attitude to the punisher and the lack of an effect here may suggest there is an upper limit to the positive attitude engaging in punishment generates. The motivations of punishers, might be questionable (Barclay, 2006; Ottone, 2008) and indeed, if punishment is a dominant behaviour then the motivations of punishers might not be as altruistic and trustworthy as they seem (for example, see Leibbrandt & López-Pérez, 2008; Nakamaru & Iwasa, 2006). If nothing else, in this study punishment was aggressive/violent and aggressive individuals are generally disliked (Hawley et al., 2008).

4.10 General Discussion

Study 3 demonstrated that while participant judgements of dominance can be accounted for by the use of aggression or winning a conflict, only individuals who engaged in punishment also gained a positive reputation. Study 4 demonstrated that success was less important than the attempt when observers made judgments about punishers. Participants perceived third parties who were unsuccessful in their intervention as being just as likeable and dominant as third parties who were successful. This overall finding could have a profound impact on the benefits that are available to someone who engaged in punishment.

4.10.1 Signalling dominance

Both studies 3 & 4 suggest that costly punishment is perceived as an indication of a dominant social position. The majority of antagonistic or confrontational behaviour is carried out by dominant individuals, or at least by those relatively dominant to the target (Silk, 2003), and we are adept at recognising the status of others (Anderson, Srivastava, et al., 2006; Sell, Cosmides, et al., 2009) and interpreting the outcomes of such confrontations (Jones et al., 2011). Indeed, as shown in the results of Study 4, while attempting punishment may have led a third party to be perceived as dominant, in terms of dominance rank the failed third party punisher was seen as being lower rank than the aggressor, i.e., while engaging in the act might raise the perceived dominance of an individual, the outcome is still relevant when making actual judgement about relative social rank. Equally, the fact that participants in Study 3 perceived a punisher to be just as dominant as an individual defending themselves or as someone involved in a random brawl suggests that the perception of the third party, in this domain at least, was of someone involved in an antagonistic conflict. The context (moralistic aggression) was not considered. This was further highlighted by the fact the strength of the third party was not taken into account for the judgements in Study 3; whether John was described as weak or strong did not affect the perceived dominance, likely because in all

instances he 'won' his aggressive encounter. Indeed, it was demonstrated in Chapter 2 that participants only paid attention to actual punishment behaviour regardless of any other social information.

Nevertheless, the fact that failed punishers were also judged to be dominant suggests that punishment could be one amongst a number of behaviours that act as a costly signal of personal prowess. Costly Signalling Theory suggests that individuals should engage in risky activity to signal something about themselves' and the aforementioned result suggests that punishment therefore be a mechanism to advertise one's formidability or dominance due to the costs it involves, specifically retaliation. Retaliation against punishment restricts its occurrence both inside and outside the laboratory (Nikiforakis, 2008; Tarling & Morris, 2010) and prevents the evolution of the behaviour in evolutionary models (Dreber & Rand, 2012). It is equally likely, if not more so, to occur when the punishment was unsuccessful, as, if punishment can be seen as a challenge, then the challenged must respond in order to save face and status (Clutton-Brock & Parker, 1995; Topalli et al., 2002). Therefore, any individual willing to punish must also be willing to bear these potential retaliation costs (see Levine et al., 2011). At the very least therefore costly punishment could be regarded as risky social behaviour that, by placing oneself in harm's way, signals to others that the punisher is not someone to be treated unfairly or challenged in the future (Barclay, 2006; Farthing, 2005).

Thus engaging in punishment might allow an individual to signal their dominance and status within a group. There are a great many benefits open from being seen as dominant and high status, for example in terms of access to resources (King, Douglas, Huchard, Isaac, & Cowlishaw, 2008), the willingness of others to acquiesce to your demands and threats (Dasgupta, 2011; Sell, Tooby, et al., 2009) and the ability to control and manipulate the social environment (Maner & Mead, 2010). In fact, individuals who are seen as dominant are seen

as more as attractive social and sexual partners (Bassett & Moss, 2004; Ellis, 1994, 1995). Thus, even if there were no further benefits to costly punishment, the ability to signal dominance and status could potentially net the punisher sufficient indirect benefits to make the behaviour viable.

4.10.2 Likeability

The studies in this chapter also demonstrated that punishers are indeed well liked. One interpretation of this result is that the positive regard (which others have shown does translate into actual reward, Nelissen, 2008) is a form of reciprocity directed to the punisher because they carried out a public function by removing a free-rider or social defector. From a Strong Reciprocity perspective, it can be seen as spontaneously rewarding an individual for acting for the good of the group. Alternatively, the desire to associate with a punisher might have more a selfish rationale behind it. One of the proximate motivations and effects of punishment is to reduce non-cooperation (for a study where participants were explicitly asked abou their reasoning, see Masclet, 2003) and therefore any group with a self-styled punisher will be more cooperative and efficient, and thus a more attractive prospect, even if only a single individual is punishing (O'Gorman et al., 2009).

This might seem to be disadvantageous to the punisher; after all the fact that everyone will benefit from the effects of their effort is one of the key stumbling blocks in the economics of punishment (Dreber et al., 2008). However this desire might also allow a punisher to recruit coalition partners and allies. Firstly, despite being less able to behave selfishly or unfairly, individuals prefer environments where punishment is possible (Rockenbach & Milinski, 2006) and are happy to cooperate within these environments with very little actual punishment being necessary. Secondly, and in an interesting interaction with dominance, when a punishing third party is present, low status individuals are far more likely to punish unfairness from higher ranking individuals (Kim et al., 1998). Thus, when someone has

engaged in an act of punishment, it would be beneficial to associate with them for protection from unfair individuals. It should be remembered that coalitions may have been vital to survival in our ancestral past (Gavrilets et al., 2008) and having a large number of allies as an effective way to deter aggression from others even today (Fessler & Holbrook, 2013; Sapouna et al., 2011). Therefore at the very least one benefit a punisher receives from this positive regard is the ability to recruit and retain social allies, with potential for this increase in social value to translate into greater rewards in the form of increased bargaining power (Barclay, 2013) and deferential displays from others (Henrich & Gil-White, 2001).

However, the above argument is based on the assumption that the would-be punisher is actually able to deter free-riding and the results of the previous studies demonstrated that the positive attitude to a punisher cannot be explained by their mere ability to punish. In Study 3, participants did not 'like' the individuals who foiled an attempted robbery on themselves, which would demonstrate the ability to punish, and in Study 4 participants liked the punisher who tried and failed to punish as much as successful punishers. Indeed, in both Study 3 and Study 4, the likability of the of an individual did not match their perceived dominance in key areas: in Study 3, only the third party punisher was judged to be both dominant and likable; and in Study 4 participants did not like the 'high threat' punisher any more than the lower threat punisher. This suggests that how likeable the punisher was perceived to be was not driven by the desire for a protector specifically or, more generally, for someone who would police the group.

Instead the present results suggest that third party punishment could act as a costly signal of the punishers trustworthiness and honesty. Punishment is costly, both in terms of the retaliation costs (see above) and in terms of the resources needed to produce it, and it has been shown that not only do individuals reward punishers based on this cost (Nelissen, 2008) but that in general those who punish are cooperative (Falk et al., 2005) and when punishment

is costly, only the altruistic punishment of non-cooperation occurs. Thus, punishment can be seen as a valid signal of an individual's altruistic nature which, as a result, can attract other 'altruists' even if these new partners are not punishers themselves. Indeed, this was apparent in Study 4 whereby an unsuccessful punisher was as well liked as a successful one, which would not be the case if it was the outcome – the act of public good in the removal of a defector – that was valued over the potential risk the attempted punishment represented.

Nevertheless, the result from Study 4 is problematic in the sense that an increase in signalling costs (the threat posed by the aggressor) should have led to an increase in signal strength, i.e. greater likability. This may be because while individuals do like risk-takers (heroic or otherwise, Bassett & Moss, 2004; Farthing, 2005), there is a limit to this: at some point the behaviours may seem reckless as opposed to brave (Farthing, 2007). In specific reference to the scenario in Study 4, while there may be advantages to associating with a punisher, these may be diminished if said punisher repeatedly start fights with strangers. Alternatively, if costly punishment is primarily a dominant behaviour, one that functions to signal and maintain this position, then there may be a point at which the benefits of being associated with a punisher are outweighed by the risk of this association (for example, of exploitation or coersion, Dasgupta, 2011; Hawley, 1999; Sell, Tooby, et al., 2009).

4.10.3 Indirect benefits of costly punishment

Panchanathan & Boyd (2004), Santos, et al (2010), and others have demonstrated that costly punishment can be evolutionarily stable if the punisher receives some sort of indirect benefit from their actions, and one indirect benefit is through reputation. Indeed, we know that punishment behaviour itself is strongly affected by the possibility of reputational gain (Kurzban et al., 2007; Rockenbach & Milinski, 2011) and the studies in this chapter demonstrated that punishment does alter the social perceptions of observers. Importantly,

punishers are seen as dominant but, unlike other confrontational behaviours, punishers were also well liked by these observers.

This has important ramifications for both proximate costly punishment and the ultimate stability of this behaviour, as, while there are advantages to being dominant or being seen as formidable, violent or aggressive individuals are disliked by others (Benard, 2013; Hawley et al., 2008). Indeed, in non-human primates short-term revolutionary coalitions will often form to depose a dominant individual, and while in this case the coalitions form for the entirely selfish reasons of supplanting the dominant individual, in our evolutionary history this coalitional psychology (Pietraszewski et al., 2014) was refined to curtail and contain overly dominant individuals (Boehm, 1997; Charlton, 1997; Gavrilets et al., 2008). In fact even if this was not possible, it *was* entirely possible for a number of individuals to simply leave the group of a despotic leader (Van Vugt et al., 2004).

Costly punishment, therefore, potentially provides a mechanism by which an individual could signal their own formidability without the negative consequences described above. In fact, this may be why the second-party punisher was not well liked; costly punishment demonstrates that one can punish free-riders and defectors, but it also demonstrates that force will be used only in a manner that conforms to social norms and attitudes of fairness, i.e., that by establishing one is a *moralistic* punisher an individual is also signalling they will not engage in spiteful or anti-social punishment or will engage in the more overtly coercive behaviours associated with dominance or high status. Thus, costly punishment can be seen as a sort of 'heroic helping' (Barclay, 2013), that allows an individual to demonstrate their formidability while at the same time signalling their pro-social and cooperative character.

However, a note of caution should be issued as these benefits rely on punishment being a costly signal; a punisher must be willing to spend resources on the act itself, willing to risk

the retaliatory actions of the target, and also (potentially) be able to continually demonstrate their reputation as a punisher. As faking or otherwise sending false signals can have severe repercussions (Anderson et al., 2008), the question remains as to whether these benefits are actually open to all individuals, and specifically whether the dominance and status of the punisher might set a barrier to entry for access to these benefits. This will be addressed in Chapter 5.

4.10.4 General conclusion

The studies in this chapter demonstrate that punishment can affect the reputation of a punisher in the eyes of observers, making punishers seem both likeable and formidable. Independently, these are both traits that would provide the punisher with long terms benefits for their actions, and together may increase an individual's overall 'worth' to other individuals by demonstrating the punisher is both useful to them and also capable of inflicting costs on them should anyone attempt to cheat or subvert the punisher. Thus, to paraphrase Machiavelli (1532/2003), while it may be better to be feared than loved, ideally a Prince should aspire to be both, and the results from Study 3 & 4 suggest one way to achieve this is to engage in costly punishment.

5 Chapter 5: dominance rank and observer perceptions of costly punishers

Chapter 5 investigates the role that a dominant position plays in the acquisition of the indirect benefits from punishment that were identified in Chapter 4, i.e., as suggested in 2.3.1 and 2.3.4, are these indirect benefits *only* accessible by dominant individuals? Study 5 investigates whether the dominance rank of a punisher affects how participants perceive both the likelihood of successful punishment occurring and the subsequent risk of retaliation from punishment. Study 6 investigates whether *only* dominant individuals are expected to punish, and how judgments of the occurrence of punishment, and the indirect benefits from punishment, are affected by the dominance ranks of the punisher and the target of punishment.

5.1 General introduction

Costly punishment can be evolutionarily stable if punishers can recoup the cost of punishment by indirect means (Panchanathan & Boyd, 2004; Santos et al., 2010). The studies in Chapter 4 demonstrated that punishment did signal something about the punisher to observers, that they were 'likable' and that they were also dominant individuals. Perhaps more importantly there was no similar effect seen in other confrontational behaviour. Chapter 4 also demonstrated that observers of punishment take the context of the altercation into account, with the risk to the punishment being factored into their judgements, which further suggests that punishment can function as a costly signal (Nelissen, 2008).

However, while these results suggest that the costs of punishment can be recuperated through indirect reciprocity, in order to continually access these gains the qualities signalled by punishment must be consistent over the long term (Számadó, 2011b), i.e. a punisher must be able to continually demonstrate their reputation as a punisher. This is important as research

on conventional signalling mechanism (those that function to induce a behaviour in conspecifics) has shown that their 'honesty' is continually tested and as a result 'false-signalling' can have severe repercussions (Molles & Vehrencamp, 2001; Számadó, 2011b; Tibbetts & Izzo, 2010). This effect also be seen in humans, with those acting 'above their station' facing significant social penalties (Anderson et al., 2008). Furthermore, for the signalling gains of punishment to be accessible, an individual must of course actually engage in punishment; a punisher must be willing to spend resources on the act itself and importantly be willing to risk the retaliatory actions of the target. The latter especially is perhaps the largest cost (Dreber & Rand, 2012) and the reason, as argued in the previous chapter, that punishment can be considered a conventional signal of cooperative intent and dominance.

Interestingly, numerous studies have shown that costly punishment is viable if *some* individuals can punish at a reduced cost compared to others (de Weerd & Verbrugge, 2011; Frank, 1996; Nikiforakis et al., 2009), i.e. if there is heterogeneity in the ability to punish. One source of this heterogeneity is dominance, or more specifically one's position in the social hierarchy. We spend our lives in both informal and formal hierarchies (Christakis & Fowler, 2010; Maestripieri, 2012) and differences in our social position can have profound effects on our behaviour (Fiddick & Cummins, 2007; Galinsky et al., 2003; Gambacorta & Ketelaar, 2013; Gregory & Webster, 1996; Maner & Mead, 2010). As will be detailed below, dominance can be seen as a biological factor that might provide heterogeneity in the costs of punishment. Chapter 4 suggested that punishment can signal dominance, and the present Chapter 5 will investigate whether costly punishment can signal dominance because *only* dominant individual can actually engage in it. Thus, Chapter 5 directly addresses the indirect benefits available to dominants as discussed in 2.3.4.

5.1.1 Dominance and costly punishment

Firstly, we may expect dominance to be associated with costly punishment because, if not an aggressive act per se, punishment is certainly a confrontational one: it is an antagonistic encounter between a punisher and an aggressor or social defector and dominance is strongly associated with antagonistic behaviour. Behaving aggressively can assert dominance and ensure that a dominant position is maintained (Silk, 2003), and conversely, reacting aggressively against being unfairly treated is vital to maintaining social status: examples of such reactions include redirected aggression in non-human primates (in vervet monkeys, Cheney & Seyfarth, 1989; and for an overview, see Kazem & Aureli, 2005), and conflicts in human societies, which are often explicitly driven by the desire to maintain status and to deter future antagonism (Mathew & Boyd, 2011; Topalli et al., 2002). Simply put, instigating antagonistic or confrontational interactions is characteristic of the dominant individual and, as was shown in Chapter 3, costly punishment was seen as an equally dominant behaviour as other *non-altruistic* aggressive actions.

In regard to punishment specifically, while there is little evidence of *altruistic* punishment in non-human animals, there is a great deal of evidence of intervention by dominant individuals across numerous taxa: examples include growth and reproductive policing (Cant et al., 2010; Wong et al., 2007), the disruption of conflicts between subordinates to curtail the winner effect (Jennings et al., 2011) and the disruption of affinitive behaviour between subordinates to prevent the formation of rival coalitions (De Waal, 1982/2007; Widdig et al., 2000). In all these cases intervention limits or prevents the rise of a social challenger, and directly comparable behaviour by dominant individuals can also be observed in humans (Maner & Mead, 2010).

Costly punishment of 'unfair' behaviour may therefore have an evolutionary origin in detecting – and responding to – potential social challengers (Brosnan, 2011; Cummins,

1996a): antagonistic/unfair behaviour may indicate a change in the social hierarchy, and it is in a dominant individual's best interest to recognise and respond to any such change. This may explain why dominant or high status individuals in general seem more willing to respond to perceived unfairness or the violation of social norms (Cummins, 1999; Lammers et al., 2010) and why the formidability of the violator is a strong predictor of 'outrage' (Jenson & Peterson, 2011).

5.1.2 Dominance and the cost of punishment

The above gives a theoretical rationale as to why there should be a relationship between dominance and punishment. In addition, dominance plays a proximate role in lowering the cost of punishment. Punishment can potentially be evolutionarily stable as long as some individuals have a greater amount of resources (Frank, 1996) or if some individuals can punish more effectively and more cheaply than others (de Weerd & Verbrugge, 2011; Roberts, 2013). A dominant position covers a number of attributes that would allow a dominant individual to punish more cheaply than others.

Firstly, dominant individuals do have access to a greater amount of resources. For example, their position gives them greater opportunities for reciprocity and cooperation (Jones & Rachlin, 2006) and their prominence means that others are willing to both tolerate asymmetries in reciprocity and to provide aid in conflicts in order to maintain a close relationship with the dominant individual (Barclay, 2013; Schino & Aureli, 2009). Dominant individuals also demand that their needs are met above others (Sell, Tooby, et al., 2009), can behave coercively in dyadic relationships to ensure this (Hawley, 1999), and are less likely to face punishment for behaving unfairly (Eckel et al., 2010; Kim et al., 1998). Thus, even if social status does not affect the absolute individual cost of punishment, the relative cost will be lower for dominant individuals.

Secondly, dominance may reduce the production cost of punishment by making it more effective. Effectiveness of punishment is important to its evolutionary stability (de Weerd & Verbrugge, 2011) and it has been shown that only effective punishment deters free-riding (for example, Nikiforakis & Normann, 2008). However, while the latter finding is consistent across the experimental costly punishment literature, so far little has been said as to how it would manifest outside of the laboratory, i.e. why would individuals be able to punish effectively? Dominant individuals can punish more effectively, insomuch as they can inflict a greater cost on the target physically (Sell, Tooby, et al., 2009) or use their social position to limit access to resources or information (Maner & Mead, 2010).

Furthermore, perhaps the most important cost to punishment is retaliation from the target (Dreber & Rand, 2012). Where retaliation to punishment is possible, punishment is reduced to the point that it no longer sustains cooperation or is evolutionarily stable (Nikiforakis, 2008; Rand et al., 2010); and, in everyday life, the threat of retaliation is a prime factor in preventing otherwise cost-free punishment behaviour such as reporting criminal activity (Tarling & Morris, 2010). Dominant individuals are, self-evidently, successful in dyadic conflicts and as previously stated, in essence punishment is a dyadic interaction between the punisher and the defector/norm-violator. Therefore dominant individuals may be able to engage in costly punishment without the risk of reprisals as the target will simply acquiesce to their demands. Indeed, when punishment occurs outside of the laboratory, it is carried out by formidable individuals (Huston et al., 1981) or by those with the support of allies (Mathew & Boyd, 2011); circumstances where the threat of retaliation would be reduced. This also suggests that retaliation could be a conventional cost to punishment that may make it a costly signal (Nelissen, 2008), as even if the production cost of punishment is low (for example, punishment by gossip, Bazzan & Dahmen, 2010; by ostracism, Bowles & Gintis, 2004; or by

condemnation, Masclet et al., 2003) the retaliatory cost may be severe for anyone in a subordinate position.

Finally, as dominant individuals can punish more effectively and face less risk from retaliation, it may be possible for them to lower the cost of punishment further, potentially to effectively zero, by establishing a credible threat of punishment (McNamara & Houston, 2002). Once a reputation for costly punishment has been established, an individual may never, or at least rarely, need to actually engage in punishment. In effect this can be seen as an extension of "don't mess with the enforcer" benefit to punishment (Barclay, 2006) to "don't mess with anyone in the vicinity of the enforcer".

5.1.3 The current studies

It has been suggested that punishment can act as a costly signal of an individual's 'altruistic' character and their commitment to the group (Barclay, 2006; Nelissen, 2008) and can also function as a signal of personal formidability (Barclay, 2006; Johnstone & Bshary, 2004). Chapter 3 demonstrated that punishers are judged by observers to be both likeable and dominant. However the direct costs of punishment provide a potential barrier of entry for access to these indirect benefits. So while costly punishment might, in principle, generate enough indirect or reputational benefits to be evolutionarily stable (Panchanathan & Boyd, 2004; Santos et al., 2010), only dominant individuals can access these benefits; subordinates may need to find less 'heroic' ways to generate a positive reputation (Barclay & Reeve, 2012). To test this, using a series of vignette studies, this chapter investigates how the dominance rank of a punisher affected judgements of success in a punishment situation, perceptions of the risk or retaliation, and whether status mediates the reputational benefits to engaging in punishment. As with Chapter 4, punishment is carried by 'third parties' as they are not affected directly or indirectly by the unfair behaviour being punished.

5.1.4 Operationalising dominance

While 'dominance' itself can be difficult to define in human (Lewis, 2002), and in non-human animals is often directly related to size and formidability (Clutton-Brock & Parker, 1995), here the term is used to cover a range of concepts such as formidability, status, prestige and power (Keltner et al., 2003; Lewis, 2002; Sidanius & Pratto, 2004). As defined in 2.1 dominance here implies that, through whatever mechanism, some individuals *"have priority of access to resources, especially reproductive resources"* (Cummins, 1996a, p. 467) or preferential access to *"any requisite that adds to the genetic fitness of the dominant individual"* (Wilson, 1980, p. 129). I.e. dominance is used here to identify an individual who, for whatever reason, has a 'strong position is a social hierarchy' which results in preferential access to these resources.

While there are likely to be differences between the different 'paths' to dominance (Cheng et al., 2013) in the effects they have on behaviour, we believe that the benefits of a 'strong social position' would be comparable whether this position was achieved through, for example, aggression or prestige (Henrich & Gil-White, 2001). To demonstrate this, in the current studies dominance was operationalised to mean a prestigious position rather than one based on formidability; if a prestigious individual is judged to be able to punish more cheaply because of biologically less tangible characteristics (e.g. 'seniority' or 'talent'), then it is likely a physically powerful individual would too. Doing so allows the current studies to directly test whether social power, as opposed to physical violence, is seen as a credible threat.

5.2 Study 5: can only dominant individuals enforce a credible threat of punishment?

Punishment is costly, but Chapter 3 demonstrated that punishers are seen as both likable and dominant and these are benefits that could allow a punisher to recuperate the cost of punishment. Equally, punishment can also be evolutionarily stable if the cost of punishment

is low (de Weerd & Verbrugge, 2011), and this can be achieved if a threat of punishment is credible (Cant & Johnstone, 2009) or through less aggressive punishment such as ostracism (Bowles & Gintis, 2004). The current study therefore addressed whether the status of a punisher affected the perception of their ability to make the threat of punishment credible, whether it affected the risk of retaliation they faced, and whether the type of intervention affected how punishers are judged. The current study also addressed how these factors affected any reputational gains generated from an act of punishment.

Also, informed by the results of Study 3 (Chapter 4), to ensure effects other than that of dominance and punishment-type were kept to a minimum, the scenario was altered to lower the 'risk' to participants from the aggressor: participants were described as being within the group and the targets for aggression were out-group members. These changes also allowed the information regarding status to be integrated more subtly into the vignette.

5.3 Method

5.3.1 Participants & Materials

108 psychology undergraduate students from the University of Exeter (86 females)[3] completed the study. Participants were not offered any incentive for taking part. The survey was administered in paper-form by a single researcher. Participants were approached around the Psychology building and those who agreed to take part were presented with a paper questionnaire containing one of four experimental vignettes and a series of questions concerning the punisher in the scenarios. Recruitment began on 1st October 2012 and concluded on 12th December 2012. All participants passed the manipulation checks (see 5.3.4).

[3] A pilot study (n=40) run prior to the current study using the same scenario and questions (with additional questions regarding the 'believability' of the scenario etc) intentionally achieved an equal sex ratio. There was no main effect of sex on the DVs recorded, nor were the study DVs affected by an interaction between sex and status/punishment type.

5.3.2 Experimental Scenarios

Participants were asked to imagine themselves as part of a local sports team, who, following an evening practice session, had retired to a local bar. The team had occupied a table but there were not enough seats for everyone. Therefore some members, including the participant, had to stand. Nearby, two strangers were sitting at another table and after a few minutes one of them headed to the bar to order drinks. Seeing this, one of the standing members of the team went over to the table and proceeded to take the now vacant chair, dismissing the objections of the still seated stranger. Upon their return with the chair, another member of the team confronted this person about their actions (for the full vignette, see Appendix C).

The study manipulated the status of the confronting team member – the third party - and how they carried out their confrontation. They were described as either "popular and the most skilled player" (dominant) or "unpopular and the least skilled player" (subordinate), and they either threatened to hit the other team member (aggressive punishment) or threatened to prevent them playing in all future matches (non-aggressive punishment), giving the study a 2x2 between-subjects design. Note that 'third party' is used here as the punisher can be considered to be 'disinterested' as the anti-social behaviour did not impact the group.

5.3.3 Social perception questions

Following the scenario, participants were asked a series of questions designed to investigate how credible the threats from the third party were. Participants were asked to indicate 'what happened next' from one of two choices; either the punishment was successful with the team member returning the chair, or unsuccessful and the team member kept the chair. They were also asked to indicate on a on a scale of 1-7 (1=not surprised, 7=very surprised), how surprised they were that the specific individual in the scenario intervened and, on a scale of 1-7 (1=very unlikely, 7=very likely), whether they believed the reprimanded individual would

retaliate against the punisher. All participants were then asked the five likability questions (current study, α=0.82) and the five social dominance questions (current study, α=0.85) as detailed in Study 1.

5.3.4 Manipulation checks and demographic questions

Participants were then asked two comprehension/manipulation check questions. They were asked to indicate, from a choice of "popular and skilled" or "unpopular and unskilled" how the third party was described in the scenario and to indicate, from a choice of "Threatened to hit them" and "Threatened to ensure they never played for the team again", how the third party intervened. Finally, participants indicated their age, sex and nationality.

5.4 Results

5.4.1 Credible threat of punishment

Participants were first asked whether they believed the aggressor would ignore or give in to the third party's demands. As shown in Figure 5.1 participants believed that the intervention by the dominant punisher would be more successful (Wald χ^2_1=147.53, p<0.001) and did not believe that the type of punishment alone would alter the outcome (Wald χ^2_1=0.51, p=0.48). Figure 5.1 shows that while participants believed the dominant punisher would be successful

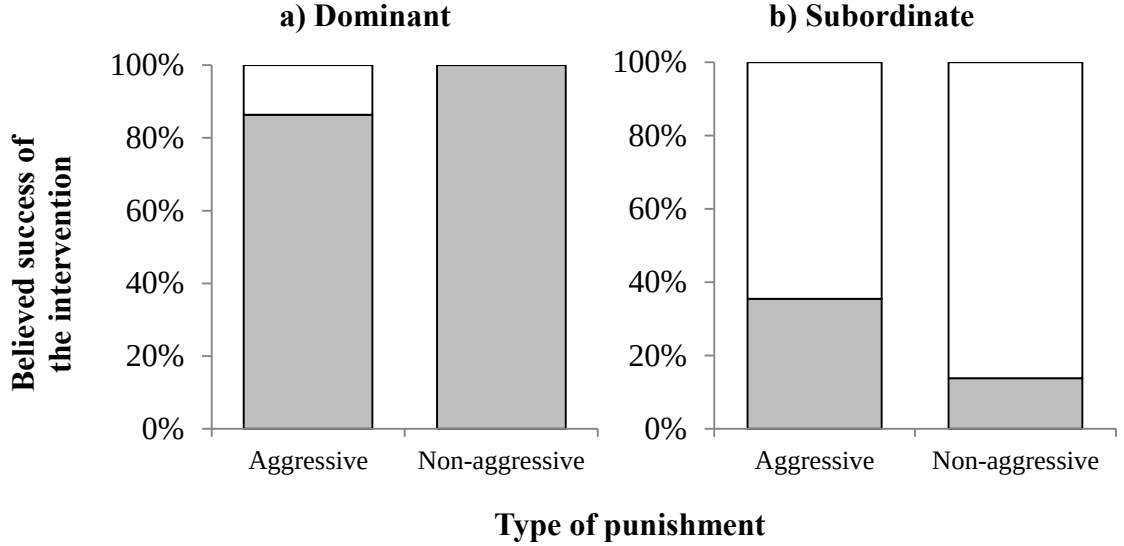


Figure 5.1: proportion of participants who believed the intervention by an a) dominant or b) subordinate punisher would be successful (grey) or unsuccessful (white).

regardless of punishment type, the subordinate punisher was thought to have even a modest chance of being successful only when being physically aggressive (Wald χ^2_2=9.80, p=0.002).

As shown in Figure 5.2, participants were far more surprised when the Subordinate member attempted punishment ($F_{1,105}$=128.16, p<0.001) and believed retaliation from this intervention was more likely to follow ($F_{1,105}$=6.70, p=0.011).

5.4.2 Perceived dominance and likability

The dominant third party was, as may be expected, perceived to be more dominant ($F_{1,105}$=111.76, p=0.001; dominant third party, M=5.5, SD=1.1; subordinate third party, M=3.6, SD=1.2) but there was no effect of status on how likable they were judged to be ($F_{1,105}$=0.48, p=0.49). Figure 5.3 shows that when the Third Party engaged in aggressive punishment they were seen as less likable ($F_{1,105}$=6.84, p=0.01): however, being more aggressive did not lead the punisher to be judged as more socially dominant ($F_{1,105}$=2.07, p=0.10, Figure 5.3). No interaction was found between either status and punishment for likability ($F_{1,105}$=0.83, p=0.77) or perceived dominance ($F_{1,105}$=0.43, p=0.51).

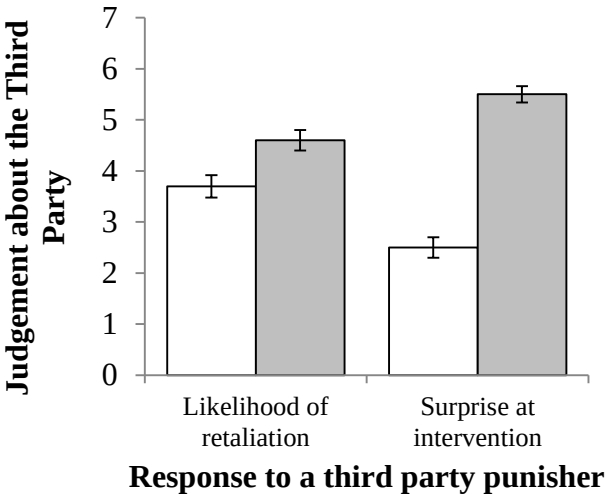


Figure 5.2: participants' reaction to the intervention for a dominant (white) or subordinate (grey) Third Party. Bars = 1 Standard Error.

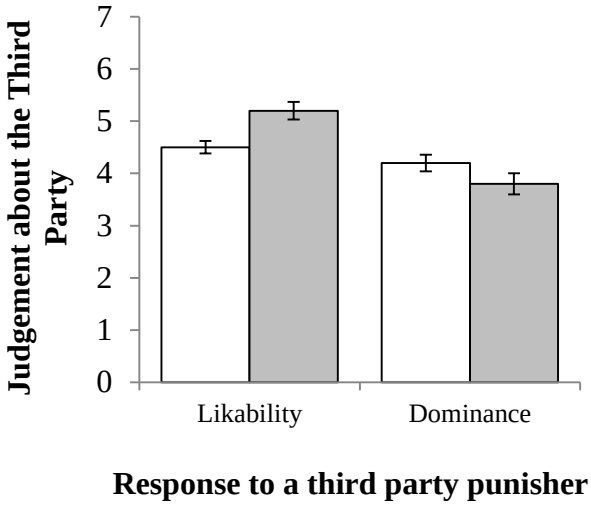


Figure 5.3: participants' perception a Third Party's likability and dominance when they engaged in Aggressive (white) or Non-aggressive (grey) punishment. Bars = 1 Standard Error.

5.4.3 Success, likeability and retaliation

Given the results regarding the insensitivity of participants to the success of punishment found in Study 4, a post-hoc analysis was carried out to see if there was any relationship between predicted success and likability; none was found (U=1308.5, N_1=59, N_2=48, p=0.5). However, there was a strong relationship between predicted success and retaliation, with participants believing unsuccessful punishers to be at greater risk from retaliation (U=856.5, N_1=59, N_2=48, p<0.001).

5.5 Discussion

These results clearly suggest that a dominant position can drastically lower the cost of punishment. Firstly, only dominant individuals were seen as being able to make a credible threat of punishment, that is to say participants believed the aggressor would back down when faced with a threat of punishment from a dominant third party. Thus, for dominant individuals, the realised costs of punishment can be effectively reduced or even removed completely by replacing physical action with a credible threat of punishment. Importantly, the credible threat imposed by the dominant group memeber was not based on the type of punishment employed; they were seen as equally likely to be successful whether the threat was aggressive (threat of physical violence) or non-aggressive (threat of ostracism from the group). In fact ostracism has previously been shown to facilitate group cooperation without coordinated punishment and at no cost to the punisher (Bowles & Gintis, 2004; Masclet, 2003). Such a threat therefore can be seen as highly credible and effective, but only if it comes from a dominant individual

Secondly, this study found that dominant individuals were judged to be at less risk of retaliation than subordinates. Chapter 3 established that individuals who attempt costly punishment are seen as more dominant, yet participants in the current study were both surprised at the intervention by the subordinate individual and believed they would be at

greater risk from retaliation. Even dominant individuals were judged to be at some risk (see Figure 4.3) and it may be the case that at least some risk of retaliation is required for any punishment to be a costly signal: while their threats may be credible, a dominant individual would need to prove on occasion they can actually enforce such threats. Indeed, while potentially costly punishment may be important in signalling one's dominant position, participants felt that a subordinate individual attempting to assert themselves in this way would be unsuccessful. In both human and non-human animals false-signalling is often responded to severely (Anderson et al., 2008; Számadó, 2011b; Tibbetts & Izzo, 2010) and in the current study participants believed that attempted punishment by a subordinate would lead to a greater risk of retaliation.

The study also suggested that the social benefits generated by engaging in punishment are significantly affected by dominance, specifically the ability to successfully use non-violent punishment. Participants disliked the third party who threatened physical violence and only the dominant punisher was perceived as being successful when non-violent punishment was threatened. Dominant individuals can therefore punish in a more socially acceptable way and as a result make greater reputational gains than subordinates.

Nevertheless, the dominant punisher in this study was only able to punish non-violently due to their authority in the groups and this leverage may not always exist in 'real life'. However, that such a restriction exists adds weight to the argument that dominance explains heterogeneity in the cost of punishment, because individuals will be more or less dominant depending on the circumstances and thus more or less able to punish cheaply (for how proximate costs influence punsihment behaviour, see Egas & Riedl, 2008; Nikiforakis et al., 2009). Still, while less liked, the violent stance by the dominant punisher was also predicted to be successful. In comparison to Chapter 3, where punishers were more well-liked in comparison to those engaging in other violent behaviour, the current study suggests that

while any punishment of anti-social behaviour is responded to positively by observers, there is a preference for less violent intervention.

The results contradict some of those from Chapter 3. There it was found that success was no predictor of 'likability', which does suggest that potentially both dominant and subordinate individuals could gain a reputational benefit from attempting punishment. However participants also believed that failure in punishment would invite retaliation so, for subordinates, the retaliatory cost of failure would likely outweigh any benefits from the attempt. Again, participants were very surprised at the intervention by a subordinate punisher, so while the vignette 'forced' a subordinate to punish, it is debatable whether in a real-life situation that a low status or subordinate individual would actually engage in any form of costly punishment.

Finally, the criticism above, that 'non-violent' punishment would only be available in a certain context actually add theoretical support to the suggestion that punishment is a costly signal, as individuals not in the context or condition to punish might not do so. While not within the purview of the current research, it would be interesting to see how domain-general different types of dominance are. Prestige (Henrich & Gil-White, 2001), for example, requires both reputational and culturally specific knowledge; without such knowledge a newcomer would not recognise that certain titles confer dominance (e.g. 'Lord'), whereas fighting ability is fairly easy to assess (Sell, Cosmides, et al., 2009). Thus, while a Prophet may never be respected in his own home town, the Ultimate Fighting Champion probably will be.

5.6 Study 6: Dominance rank, outcome, and observer perceptions of costly punishers

Study 5 demonstrated that dominance can lower the cost of punishment, both in terms of the production cost (i.e. one never has to engage in an actual physical confrontation for it to be

effective) and also reduces the risk of retaliation. However, in any conflict, the status and condition of both parties should contribute to the outcome (Maynard-Smith & Price, 1973). Study 6 therefore also manipulated the status of the aggressor in the scenario.

Also, Study 5 found that while dominance lowered the cost of punishment, the punisher was able to generate indirect benefits regardless of success. This implies that any group member could access some of the benefits of costly punishment even if they had not established a credible threat. One explanation for this is that the study scenario did not allow the third party to refuse to intervene: as discussed in more detail in Chapter 3, the attempt at punishment may at least signal a belief that one could win the confrontation regardless of status, whereas, in 'real-life' a subordinate group member might never actually intervene. The current study accounted for this by providing the participants with an additional 'does nothing' option when asked to predict the outcome of the confrontation.

5.7 Method

5.7.1 Participants

Participants were recruited from the University of Exeter via a university-wide web-based recruitment system (SONA). A total of 119 participants, 26 Males (M age = 24) and 93 females (M age = 20) with an overall age range of 18 – 46 completed the questionnaire. As an incentive, participants who completed the survey were entered into a prize-draw for one of several £10 shopping vouchers. Recruitment began on 26th September 2013 and concluded on 24th October 2013. No participants failed the manipulation checks (see 5.7.5).

5.7.2 Materials and procedure

The survey consisted of three sections. The first section presented participants with an experimental vignette and the second section collected participants' responses to these vignettes. The third section collected demographic information and contained the

manipulation check questions. The survey was conducted using the web-based application SurveyMonkey (www.surveymonkey.com) and was presented to participants in the order shown below (for the full vignette, see Appendix C).

5.7.3 Experimental vignettes

Initially, the scenario was identical to Study 5; participants were asked to imagine themselves with a group of team members in a local bar when a member of their team committed an anti-social act, which another member of the group noticed and was visibly angered by it. Unlike Study 5 however, the scenario ended there without describing how this third party responded to the norm violation. Again, 'third party' is used to denote proximate 'disinterest' (see 1.1).

Also, unlike Study 5, the current study manipulated the status of both the chair-taker (the 'aggressor') and the other team member (the 'third Party'). Depending on the condition, each was described as either "a popular and skilled player" (dominant) or "an unpopular and unskilled player" (subordinate), giving the study a 2x2 between-subjects design.

5.7.4 Social perception questions

Following the scenario, participants were asked to indicate 'what happened next' from one of three choices. They were asked to indicate whether they believed: the third party would intervene successfully, with the aggressor returning the chair; the third party would intervene unsuccessfully, with the aggressor keeping the chair; or the third party would not intervene at all. The former two options stated that "after a brief exchange, the chair taker…", that is to say it was not specified whether the intervention involved physical or social threats. All participants were then asked the five likability questions (for this study, α=0.87) and the five social dominance questions (for this study, α=0.86) as detailed in Study 1. Finally, participants were asked to indicate, on a scale of 1 (not likely at all) to 7 (extremely likely),

how likely it was that the aggressor would try and 'get even' with the third party then or at a later date.

5.7.5 Manipulation checks and demographic questions

Participants were then asked the two comprehension questions. They were asked to indicate, from a choice of "popular and skilled", "unpopular and unskilled" or "sort of popular and skilled" how the aggressor and the third party were described in the scenario. Finally, participants indicated their age, sex and nationality.

5.8 Results

5.8.1 Outcome

Participants were first asked to indicate "what happened next", whether the third party successfully intervened, unsuccessfully intervened or failed to intervene. As shown in Figure 5.4a, participants believed that for a dominant third party the most likely outcome was a successful intervention and that a subordinate third party was unlikely to intervene at all (Wald X^2_1= 18.33, p<0.001). As shown in Figure 5.4b, the rank of the aggressor also affected perceived outcome, with participants believing that a third party would be less likely to intervene when the aggressor was dominant (Wald χ^2_1=5.03, p=0.025). Perceived outcome was not affected by an interaction between the rank of the third party and the aggressor (Wald χ^2_1= 1.27, p=0.26). However Figure 5.5 does suggest that while the rank of the aggressor was important in the perceived outcomes, this was more clearly the case when the third party was subordinate.

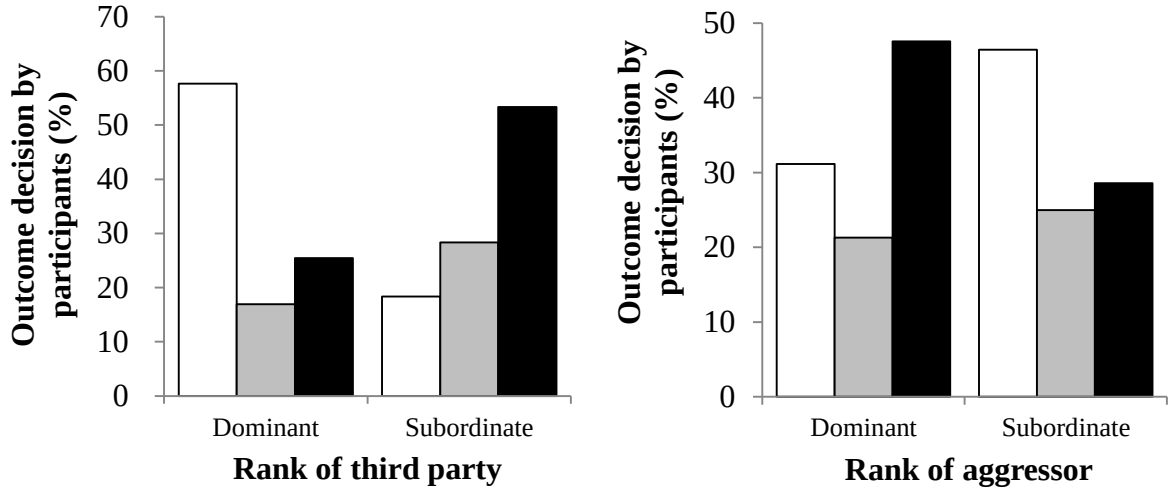


Figure 5.4: predicted outcome of third party punishment depending on a) the rank of the third party or b) the rank of the aggressor (White = successful intervention, grey =unsuccessful intervention, black= no intervention).

5.8.2 Likability

The rank of the third party did not affect their likability ($F_{1,115}$=2.57, p=0.11), however the third party was less well liked when the aggressor was dominant (M=4.6 SD=1.2) than when the aggressor was low ranked (M=5.0, SD=0.9; $F_{1,115}$=4.57, p=.0.035). The likability of the third party was not affected by an interaction between the ranks of the third party and the aggressor ($F_{1,115}$=0.98, p=0.75).

How participants predicted the outcome of the scenario had a strong effect on likability

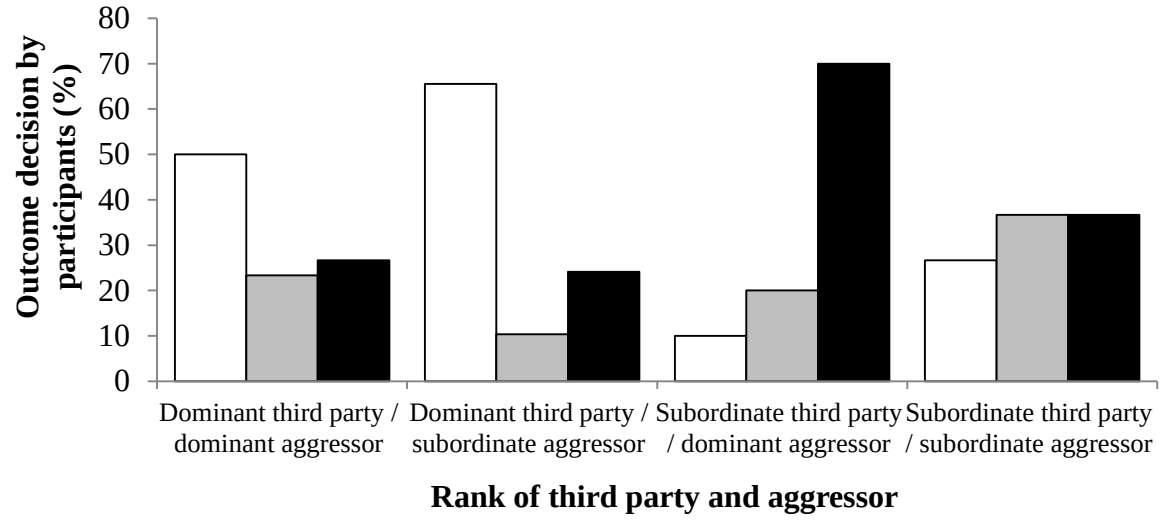


Figure 5.5: predicted outcome of third party punishment depending on the rank of the third party and the aggressor (White = successful intervention, grey =unsuccessful intervention, black= no intervention).

($F_{2,116}$=4.11, p=0.019), with participants liking the third party who was predicted to be successful in their intervention (M=5.1, SD=1.1) more than those predicted to be unsuccessful (M=4.7, SD=1.1) or predicted to not intervene (M=4.5, SD=0.9). This may explain why participants liked the third party who punished a subordinate aggressor more, as punishment was seen to be less successful when directed against a dominant aggressor. Therefore, a mediation analysis was conducted with 'outcome' as the mediating variable[4]. With the rank of the third party controlled for, the predicted outcome completely mediated the relationship between the rank of the aggressor and the likability of the third party (b=0.08, BCa CI_{95}=0.03, 0.24, on 5000 samples), with Aggressor Rank no longer effecting likeability (b=0.34, t=1.73, p=0.09). Thus, the likeability of the third party was dependent on whether their punishment was expected to achieve a successful outcome.

5.8.3 Dominance

As shown in Figure 5.6, unsurprisingly the third party was perceived to be more dominant when described as dominant as opposed to subordinate ($F_{1,115}$=16.18, p<0.001). The third

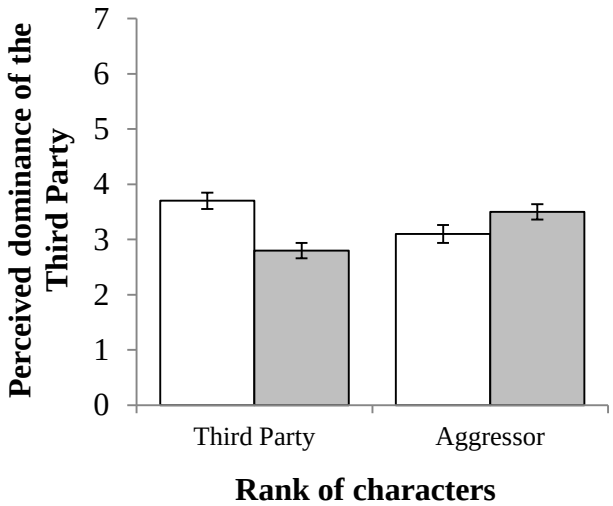


Figure 5.6: the perceived dominance of the third party depending on the rank of the third party and the aggressor (dominant = white, subordinate = grey). Bars = 1 Standard Error.

[4] Analyses carried out using linear regressions suggest that the 'outcome' categories produce a graded response and can therefore be considered as a 'scale of intervention', from no intervention to completely successful.

party was also perceived to be more dominant when the aggressor they faced was described as subordinate ($F_{1,115}$=3.64, p=0.059, Figure 5.6). The perceived dominance of the third party was not affected by an interaction between the ranks of the third party and the aggressor ($F_{1,115}$=0.24, p=0.63).

As with likeability, what participants predicted about the outcome had a strong effect on perceived dominance ($F_{2,116}$=9.89, p<0.001), with successful third parties being seen as more dominant (M=3.8, SD=1.2) than unsuccessful (M=3.0, SD=1.1) or non-intervening (M=2.8, SD=1.1) third parties. With the rank of the aggressor controlled for, the predicted outcome partially mediated the relationship between the rank of the third party and their perceived dominance (b=-0.23, BCa CI_{95}=-0.50, -0.08, on 5000 samples), although the direct relationship between the two was still present (b=-0.61, t=-2.75, p=0.007).

Interestingly, with the rank of the third party controlled for, the predicted outcome of the interaction fully mediated the relationship between the rank of the aggressor and the perceived dominance of the third party (b=0.11, BCa CI_{95}=0.01, 0.31, on 5000 samples), with aggressor's rank no longer effecting likeability (b=0.29, t=1.40, p=0.17). Therefore, while the third party's described rank affected the perceived dominance, the effect of the aggressor's rank on perceived dominance of the third party was entirely due to how the aggressor's rank affected the predicted outcome.

5.8.4 Retaliation

There was a trend towards the status of the punisher affecting the perceived risk of retaliation ($F_{1,115}$=3.31, p=0.07), with dominant third parties being perceived as being at less risk of retaliation (M=3.1, SD=1.9) than subordinate third parties (M=3.9, SD=2.1). The rank of the aggressor did not affect the perceived risk of retaliation ($F_{1,115}$=0.04, p=0.84), nor was the

risk of retaliation affected by an interaction between the ranks of the third party and the aggressor ($F_{1,115}$=0.67, p=0.412).

Outcome did not affect the risk of retaliation ($F_{2,116}$=0.53, p=0.59), and consequently a mediation analysis would not be appropriate. However, as an exploratory analysis, outcome was entered as a covariate; in this analysis the status of the third party did affect the perceived risk of retaliation ($F_{1,114}$=5.85, p=0.017).

It should be noted however that, as we assumed *a priori* that some participants would select the 'do nothing' outcome, the retaliation item asked participants to "assume the agitated person [the third party] did intervene, regardless of your initial decision". Therefore the analysis, with outcome as a covariate, was run again after removing participants who indicated that the third party would not intervene. As shown in Figure 5.7, while individually the rank of the third party ($F_{1,67}$=0.005, p=0.94) and the aggressor ($F_{1,67}$=0.008, p=0.93) did not affect the perceived risk of retaliation, retaliation was affected by an interaction between the two ($F_{1,67}$=4.08, p=0.047); participants who predicted the third party would intervene felt

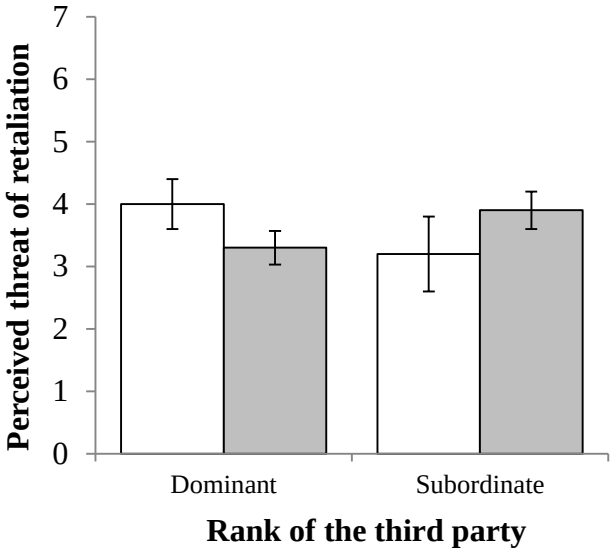


Figure 5.7: the perceived risk of retaliation against a successful or unsuccessful intervention depending on the rank of the third party and the aggressor (dominant aggressor = white, subordinate aggressor = grey). Bars = 1 Standard Error.

that retaliation was more likely when the third party confronted an aggressor of equal rank.

5.9 Discussion

Study 6 demonstrated that the rank of individuals involved in a conflict as considered important by observers making judgements about those involved, with the rank of the punisher being the most vital to a successful outcome. While this is not surprising as such, it does support the suggestion made in Study 5 that subordinates are not expected to intervene at all. Indeed, the pattern of predicted outcomes as shown in Figure 5.5 suggests that participants believed a punishing group member would either intervene successfully or not at all.

More interesting however was how the reputational benefits were affected by the interaction between the statuses of the belligerents and the predicted outcome of the intervention. Fundamentally, the likeability and perceived dominance of a third party was dependent on a successful outcome of the interaction. While a previous study (Study 5) failed to find such an effect, this is most likely explained by the lack of a 'failed to intervene' option in that study. In Study 5 for example, regardless of their other social judgements, participants were very surprised when a subordinate individual intervened and in Study 6 this surprise translated into participants predicting that subordinates would not intervene at all. It also provides a conceptual model of how status interacts with these benefits. While a successful outcome to punishment is the most important factor when observers make social judgements, that outcome only occurs if the punisher is dominant.

This point should be stressed as it highlights a wider issue with the economic research into punishment and the benefits thereof, as by design *all* punishment in economic experiments is successful, and this not what individuals always expect to occur in social conflicts. The results of Study 6 clearly show that whether punishment is successful has a strong impact on

any reputational benefits. While research has been concerned with downstream effects of successful punishment, for example on reputation (Nelissen, 2008), the behaviour of free-riders (Masclet, 2003) or group efficiency (Gächter et al., 2008), there has been no consideration that the attempt at punishment might fail. And both studies in this chapter have shown that punishment by dominants, in the perceptions of observers at least, does not fail. Thus the conventional experiments are, in effect, modelling the behaviour only of dominant individuals.

In terms of retaliation the initial analysis of the data suggested that, as was found in Study 5, dominant punishers were thought to be at less risk of retaliation, with the outcome of the interaction also exerting some effect. It should be stressed once again that retaliation-costs are possibly the main cost to punishment (Nikiforakis, 2008; Rand et al., 2010) and the most overlooked (Dreber & Rand, 2012), and these results suggest that the status of the punisher can significantly lower this cost. It was expected that retaliation risk would correspond to relative rank, i.e. that a dominant punisher would face lower risk from a subordinate aggressor than a dominant one, and that a subordinate punisher would face a greater risk from a dominant aggressor than a subordinate one. In the case of the latter however, in the second analysis (with the 'no intervention' data removed) the opposite was true; the risk of retaliation was perceived to be greater when the belligerents were of equal rank. As within dominance hierarchies conflict should occur more between those of similar ranks (Wilson, 1980, pp. 141-142), this results suggests that participants perceived the encounter as a dominance contest rather than as a (purely) moralistic act.

Thus, the effect of dominance on the outcome of punishment, and in turn the indirect reputational gains from punishment, suggests that dominance is an important variable in calculating the benefit from the behaviour. At the very least, the results suggest that costly punishment can be seen (and is perceived by those witnessing it) as a dyadic conflict between

individuals that is governed by the rules and motivations of such conflicts (Clutton-Brock & Parker, 1995; Johnstone & Bshary, 2004; Silk, 2003) as opposed to being a selfless act that requires idiosyncratic group-level explanations for its motivation and existence (Fehr & Gächter, 2002; Gintis, 2000).

5.10 General Discussion

The studies in this chapter suggest that dominance greatly lowers the cost of punishment and hence increases the likelihood that it will occurAs a result the clear reputational benefits that can be gained from punishment are only open to dominant individuals. The results suggest that only dominant individuals can lower the production costs via the effective use of non-violent and cost-free threats of punishment (Bowles & Gintis, 2004; de Weerd & Verbrugge, 2011) and only dominant individuals can punish with a reduced risk of retaliation. The costs of retaliation especially may stretch beyond the initial act of punishment. Humans are very good at reputation scoring (Nowak & Sigmund, 1998), and if punishers' are 'scored' in a similar fashion as altruistic individuals (Tennie, 2012), then a reputation for enforcing fairness, while potentially beneficial for attracting some cooperative partners, could act as an reputational badge that may invite aggression from others (Számadó, 2011b), akin to the sheriff in a Western or the eponymous character of a super-hero film; i.e., the person who needs to be 'taken out' to allow the exploitation of others. In this instance a reputation for enforcing fairness might work against a punisher, or at least one who could not resist such future actions.

The suggestion that only more dominant individuals are able to access the reputational benefits of costly punishment moves beyond the idea that punishment can signal fairness alone (Nelissen, 2008). Only dominant individuals are capable of giving this signal credibly, meaning that the reputational benefits from punishment are inextricably linked to dominance. At a further level, the results of the studies suggest that individual variation in dominance and

status may have played an important role in the evolution of costly punishment and that it is an important factor in any calculation of the economics of punishment. Punishment can be evolutionarily stable if there is heterogeneity in the cost of punishment (de Weerd & Verbrugge, 2011; Frank, 1996; Roberts, 2013) and we suggest that dominance causes sufficient heterogeneity in both the cost and rewards of punishment to make this behaviour evolutionarily stable.

5.10.1 Dominance and the origins of costly punishment

Pedersen et al (2013) recently suggested that any account of the evolution of 'moralistic' punishment in humans must be relatable to behaviour seen in non-human animals. As previously stated, punishment can be seen as an antagonistic dyadic interaction between a third party and the social defector/aggressor and such encounters are predominantly instigated and won by dominant individuals (Clutton-Brock & Parker, 1995; Silk, 2003): indeed, in Studies 5 & 6, participants believed a dominant punisher, and only a dominant individual, would be successful in their punishment. More directly in line with the assertion of Pederson et al, while there is very little evidence of altruistic punishment in non-human animals, dominance determines intervention and policing across numerous taxa (for example, Fallow deer, Jennings et al., 2011; Barbary macaques, Widdig et al., 2000; and in coral fish queues, Wong et al., 2007), and these interventions by a dominant individual limit or prevent the rise of a social challenger. Therefore punishment can potentially be seen as having an origin in recognising and responding to social challenges (Brosnan, 2011), with only dominant individuals possessing the freedom of action to act upon this recognition. This is important as, firstly, punishment as a tool to maintain social position provides an additional motivation for an individual to engage in the behaviour and, secondly, the benefits an act of punishment provides can be seen as independent from cooperation (Rand et al., 2010).

However, the data concerning likeability may suggest that the use of costly intervention purely to signal dominance may be where humans and other animals diverge. The non-human examples given above are from species with a relatively steep social hierarchy, yet the relatively flat social hierarchies of modern hunter-gatherer human societies, which likely reflect early human social systems, prevent one individual (or group of individuals) from becoming too dominant (Boehm, 1997; Gavrilets et al., 2008). While costly punishment may have an origin as purely a dominance behaviour, punishment in humans has become a more complex and strategic behaviour, in much the same way that human reciprocity and cooperation has an origin in the more limited cooperative behaviour of other animals (Melis & Semmann, 2010). As coalitions became more important to an individual's survival (Gavrilets et al., 2008), punishment allows an individual to attract cooperative partners by displaying both personal cooperative tendencies (Nelissen, 2008) and the ability and willingness to police free-riders (Rockenbach & Milinski, 2006), while at the same time preventing the rise of social challengers without losing positive attitude to the point where coalitions are mobilised against them or group members simply leave (Betzig, 2014). Thus, in terms of social bargaining (Sell, Tooby, et al., 2009), punishment allows an individual to signal they are both useful to others, and capable of inflicting costs upon others should the need arise.

5.10.2 Intervention or punishment?

A theoretical criticism that can be made of the studies in this chapter is that, in terms of the strict economic definition of punishment of 'paying a cost to inflict a cost on another', the studies and the scenarios contained within concern an *intervention* by a group member rather than a punishment per se (Kurzban, personal communication). However, while there may have not been any material costs inflicted, in both studies the described scenario would have resulted in costs to the aggressor in terms of social humiliation due to being publically

shamed for, and forced to retract, an 'unfair' behaviour (Barr, 2001). Equally, it should be noted that even within the economic literature, verbal or other non-monetary responses to unfairness or non-cooperation are considered to be 'punishment' (for example, Masclet et al., 2003; Ostrom et al., 1992).

Furthermore, we would argue that the imposition of actual material costs is an anticipated downstream effect of the subsequent action taken by the *target* of any punishment. As an illustrative example, if an individual came across someone smoking on public transport (illegal in the UK) and demanded they stop, this would still be an act of costly punishment in the classic Fehr (2004) sense (i.e. the desire and subsequent behaviour to uphold a social norm) even if the smoker apologised and snubbed out the cigarette with no further interaction taking place, because of the implied threat to escalate matters if this demand was not met. In such a situation there is only physical punishment if the 'intervention' is challenged. Indeed, Levine, Taylor, & Best (2011) showed that violence after the intervention by a third party only occurs after a series of escalating behaviours by the parties involved, each of which gives the opportunity for one party to back down. As is the case in animal dominance contests (Maynard-Smith & Parker, 1976; Maynard-Smith & Price, 1973).

Perhaps most importantly, the fact that potential punishment costs might not be realised is one of the core arguments as to how dominance affects the cost/benefit of costly punishment: essentially, a position of dominance, and its implied ability to inflict effective punishment on others, functions as a credible threat. Our smoker above would be well aware of the potential costs (further social embarrassment and/or a physical confrontation) and thus chose to acquiesce. We believe that people's understanding of this implication was demonstrated by both studies in this chapter, as when faced with a challenge from a dominant individual, the transgressor was predicted to back down rather than have cost of punishment realised.

5.10.3 General conclusion

The current studies strongly suggest that variations in dominance played an important role in the evolution of costly punishment. Dominant individuals are able to punish more effectively and at a lower cost than others and therefore dominant individuals can access the signalling or reciprocal benefits generated by punishment at a much cheaper rate. We suggest that taking dominance into account may help answer some of the questions and debates around the evolution of this behaviour, specifically in terms of how some individuals can overcome the costs of punishment. At the very least, these results demonstrate that social dominance is an important factor in overcoming the proximate costs of; however we also believe that these results point to human costly punishment having an evolutionary origin as dominance-based behaviour rather than having evolved to specifically promote cooperation.

6 Chapter 6: dominance and the behaviour of costly punishers

Chapter 6 investigates whether simulating some of the characteristics of a dominant position, specifically a greater amount of resources (1.1.2 and 2.3.2) and a disproportionate benefit from group cooperation (2.3.3), would result in more costly punishment by 'dominant' individuals. Study 7 manipulates the amount of additional resources certain participants received based on the group product. Study 8 manipulates both the source of the additional resources, i.e. whether they came at the direct expense of others in the group, and the stability of the 'dominant' position. Study 8 therefore also investigates whether dominants use punishment strategically to ensure continued benefit from the public good.

6.1 General introduction

The previous chapter found that varying the dominance of punishers can have a dramatic effect on observer expectations and perceptions of costly punishment behaviour. Dominant individuals were seen as being at less risk of retaliation from the target of punishment and, perhaps because of this, only dominant individuals were predicted to engage in punishment a) successfully and b) at all. In fact, the indirect reputational benefits highlighted in Chapter 4 were found to be dependent on a successful outcome to punishment in Chapter 5. As it has been repeatedly demonstrated by others that punishers do respond to the potential for a reputation to be established, either by increasing (Bering, 2008; Kurzban et al., 2007) or decreasing (Nikiforakis, 2008; Rockenbach & Milinski, 2011) punishment, it is safe to assume that those who punish are concerned about the way their actions are perceived and responded to by others.

Nevertheless, while these findings show that dominance plays an important part in shaping the judgements and expectations of observers, the question still remains as to whether individuals in a dominant position are actually more likely to engage in costly punishment. This chapter will attempt to simulate a dominant position by manipulating some of the advantages of dominant position suggested in 2.3.

6.1.1 Dominance and the cost of punishment

As detailed in Chapter 5 (5.1.1) and the literature review (2.3), dominant individuals can potentially experience a lower cost of punishment. Perhaps the most used method of lowering the cost of punishment in the economic and evolution literature is making punishment 'effective', where each unit of resources spent on punishment inflicts a greater amount of harm on the target (for a review see, Balliet et al., 2011; see also, Bowles & Gintis, 2004; Egas & Riedl, 2008; Fehr & Gächter, 2000; Gneezy & Rustichini, 2000a; Nikiforakis & Normann, 2008; Roberts, 2013). Those in a dominant position are able (and willing) to inflict meaningful costs upon others (Sell, Tooby, et al., 2009; Silk, 2003) and outside of the laboratory punishment is generally restricted to formidable individuals (Huston et al., 1981). Dominants are also in a position to inflict greater social penalties on others as their position can represent a bottleneck in connectivity between group members (Maner & Mead, 2010; Scott, 2007, pp. 86-87). Furthermore, as dominants have more social allies these can be employed to punish effectively: the presence of social allies lowers the threat potential foes are perceived to be (Fessler & Holbrook, 2013), lowers the likelihood of being a victim of aggression (Smith et al., 2004), and when punishment takes place in non-state societies, it does so after a sufficient number of allies have been gathered (Mathew & Boyd, 2011). Indeed, spreading the cost of punishment between individuals has been suggested as an alternative way for punishment to evolve (Boyd et al., 2010), and while Chapter 2 suggested

that a 'cheap' verbal signal from a conspecific does not coordinate punishment, a dominant position might allow for more direct and overt coordination of group activity (Gillet et al., 2010; Van Vugt, 2006).

Because of the factors summarised above, dominant individuals can monopolise resources as they can behave coercively (Clutton-Brock & Parker, 1995; Hawley, 1999) or because others are willing to tolerate asymmetries in reciprocity to maintain a close relationship with them (Barclay, 2013; Schino & Aureli, 2009). The resource-controlling ability of dominant individuals (Hawley, 1999) therefore allows them to punish at a lower net cost than subordinates, which alone might allow punishment to be evolutionarily stable (de Weerd & Verbrugge, 2011; Frank, 1996). The fact that individuals are less inclined to retaliate against dominants for behaving 'unfairly' in dyadic interactions (see also, Eckel et al., 2010; Kim et al., 1998) might also lower the cost to a dominant individual from retaliation/counter-punishment. While counter punishment is one of the main costs to punishment (Dreber & Rand, 2012; Nikiforakis, 2008), a subordinate would be as unlikely to respond antagonistically to punishment as they are to other forms of aggression from a dominant individual (Silk, 2003). This was demonstrated in Study 6.

6.1.2 Dominance and the direct benefit of punishment

Nevertheless, even if dominant individuals experience a lower cost of punishment because they can punish more effectively, at a lower relative resource cost, and can avoid the cost of retaliation, it is argued that over evolutionarily time those who punish will still be outcompeted by second-order free riders, those who cooperate but do not pay the cost of punishment[5] (Dreber et al., 2008; Yamagishi, 1988), however slight it might potentially be.

[5] It should be noted here that, as discussed in Chapter 5, a further cost-mitigating advantage of a dominant position is the ability to enforce a 'credible threat' of punishment. However, it is unlikely that such a credible threat could be established without at least some punishment (Barclay, 2006). In principle this can be seen in the

Indeed, this is a fundamental argument as to why costly punishment might require an idiosyncratic group-level section mechanism in order to evolve (Fehr & Gächter, 2002; Gintis, 2000).

However, this conclusion is based on two premises that need not be true. The first is that individuals are homogeneous in the cost of punishment, which as described above is not the case when dominance is considered. The second, and perhaps more important, premise is that individuals are homogeneous in the benefits derived from punishment[6]. This also need not be the case. Heterogeneity in either the cost or benefit might make punishment evolutionarily stable (de Weerd & Verbrugge, 2011; Gavrilets & Fortunato, 2014; Roberts, 2013). We cannot manipulate a true dominance hierarchy directly in an economic experiment, although the next chapter (Chapter 7) attempted to circumvent this by making use of naturally-occurring group position. However it is possible to incorporate the advantages of a dominant position into game mechanics, for example the ability to punish or preventing retaliation. The studies in this chapter focused on another much less experimentally studied method of lowering the cost of punishment that we believe also represents an advantage of a dominant position: allowing certain participants to have disproportional access to resources. The current chapter investigated how this aspect of a dominant position, and specifically how heterogeneity in marginal benefit from group cooperation, might influence the willingness to punish.

In non-human primates and other social animals, dominant individuals monopolise resources such as food or reproduction, and while one might hesitate to compare the following

---

pattern of punishment in many of the public goods games mentioned above, i.e. when punishment is effective, there is initial free-riding and corresponding punishment, however as the game progresses punishment decreases as cooperation increases in response to the punishment.

[6] For the moment I am ignoring indirect benefits open only to the punisher, for example reputation (Nelissen, 2008; Panchanathan & Boyd, 2004, and see Chapters 3 & 4 of this thesis), preventing future exploitation (Barclay, 2006; Jensen, 2010) or the benefits of behaving spitefully (Gardner & West, 2004b; Jensen, 2010).

examples to human 'norm enforcement' (but see, Cummins, 2005), dominant individuals certainly benefit from the punishment of non-dominant breeders (Cant et al., 2010; Clutton-Brock & Parker, 1995), 'cheating' conspecifics (Raihani et al., 2010), the prevention of association between conspecifics (Widdig et al., 2000) and growth policing (Wong et al., 2007). The history of human society has been characterised by disparities in dominance, status and in the distribution of resources (see, for example, Betzig, 2014; Turchin et al., 2013; Turchin & Gavrilets, 2009). One need only to look at the now infamous graphs demonstrating the disparity between general increases in productivity and the pay gap between CEOs and the average worker to see this is still very much the case (Anderson, Benjamin, Cavanagh, & Collins, 2006; Berman, 2013).

Nevertheless, it is traditionally assumed that public goods cannot be divided insomuch as the products of the public good cannot be actively kept from non-contributors (Davis, 1993). However, as suggested by Reuben & Riedl (2013), while everyone in a community might benefit to some extent from public good activities such as dams and irrigation systems, the downstream benefits from the increase in productivity such a project may cause will vary amongst the population, either based on relative distance to that project or from the land one already holds. A good example from history, due to the informality of the hierarchy, might be found in the notorious Pirates of the Caribbean whereby those in positions of authority and control received an additional share of any loot taken (Cordingly, 2006). While the crew must cooperate to achieve an outcome (e.g. the capture of a Spanish treasure ship), the flow of resources from that activity can be divided. A further example would be work on flood or territorial defence; again everyone benefits from a lack of flooding or invasion, but some individuals have a larger amount of resources to lose to a disaster.

Thus, heterogeneity in the marginal benefit (or indeed, loss) from group-level activities may alone make investing in punishment an effective solution for more dominant individuals. However, few studies have explicitly examined how heterogeneity in resources affects punishment and cooperative behaviour, and those that have demonstrate conflicting results. Using heterogeneity in endowment, as opposed to marginal returns (i.e. additional benefit from cooperation) Burns & Visser (2006) found that participants who received lower endowments contributed more to the public good, whereas Reuben & Riedl (2013) found that who received higher endowments contributed more. The same is also true for studies that provided different marginal returns from group cooperation; Reuben & Riedl (2013) and Tan (2008) found no difference in contributions between high and low earners, while Nikos Nikiforakis, Noussair, & Wilkening, (2012) and Reuben & Riedl (2009) found that high earners contribute more, but only when punishment was not possible.

Interestingly however, of the aforementioned studies all but Tan (2008) did not report any differences in punishment behaviour between those who received higher or lower marginal benefits from group-level cooperation. This is likely because punishment was effective, spanning from a lower effectiveness of a 1:3 ratio (Reuben & Riedl, 2009) up to a 1:5 ratio (Burns & Visser, 2006), so punishment was still cheap even for participants who received less resources. Indeed, in the latter study, low earning participants proved very willing to spend their resources to reduce the income of higher earners regardless of the latter's contributions to the public good (see also, Leibbrandt & López-Pérez, 2008).

6.2 Study 7: additional benefit from group cooperation increases punishment behaviour

Study 7 addressed whether receiving additional resources as a result of group success would be enough on its own to encourage punishment, and therefore cooperative, behaviour. To test the effect of marginal return from group production alone, the study set the parameters of a

public goods game to ensure that punishment was as unlikely as possible to occur without the presence of these additional resources for punishers; the groups were organised using the stranger method, i.e. they were randomly reorganised after each round, and the punishment ratio was set to 1:1. This also meant that punishment could not be strategic, i.e. punishment could not be used with the expectation of future gains from group cooperation through changing the behaviour of free-riders (Masclet, 2003; Shinada et al., 2004).

The study mechanism also allowed only a single individual in a group to punish (Baldassarri & Grossman, 2011; O'Gorman et al., 2009). On the one hand, this is because, as mentioned previously, dominant individuals are able to monopolise resources, and specifically resources derived from group cooperation, so the current simulation of dominance required a single punisher. Equally, limiting punishment to one individual, allowed the punisher the freedom of action dominant individuals have (Galinsky et al., 2003; Gavrilets & Fortunato, 2014; Keltner et al., 2003; Van Vugt, 2006), including removing the prospect of being altruistically, spitefully or counter-punished.

It was predicted that punishers who received higher marginal returns from group cooperation would engage in more severe and more frequent punishment than those who did not. Because punishment severity and frequency were predicted to increase with marginal benefit for the punishers, it was also predicted that cooperation should increase when punishers receive additional resources, as participants would respond to the more severe punishment meted out by a well-resourced punisher. However, as a 1:1 punishment ratio does not normally increase cooperation (Nikiforakis & Normann, 2008), we predict there should only be a difference in cooperation between groups with the most well-resourced punisher and groups where punishers were given no additional resources.

6.3 Method

6.3.1 Participants

112 undergraduate psychology students (87 female) were recruited through the University of Exeter's internal email system. Due to a technical problem, age of participant was not recorded, however based on previous studies mean ages were likely to be between 18-21 years. In total, eight experimental sessions were conducted with a range of 8-20 participants in each session. The mean payment received by participants was £5.14 and the mean duration of each session was 45 minutes. All participants passed the comprehension check included with the experiment instruction sheet.

6.3.2 Experimental design

The study used a modified version of a Public Goods Game with Punishment (PGG+P) and used the stranger method. The study contained two phases, and participants played both one after the other. Participants played each phase for 8 rounds, but to avoid any end-round effect, were told there would be between 5-10 rounds in each phase. Phase 1 proceeded like a standard PGG: in each round, the participants were randomly divided into groups of 4 and given an allocation of 20 points, which they could then contribute to the 'group pot' or keep for themselves. Participants had 20 seconds to make this decision (and all subsequent decisions) before a conspicuous warning began to flash. At the end of the round the amount in the group pot was doubled and then divided equally between the group members. Participants were then presented with a list of the contributions made by other group members (who were arbitrarily labelled in each round as 'Player 1', 'Player 2', or 'Player 3'), the group pot total, and their individual earnings for that round.

In Phase 2, (PGG+P), after making contribution decisions identical to those in Phase 1, one individual in each group was randomly selected to have the ability to punish other

participants in their group at a ratio of 1:1, i.e. every point they removed from another player would cost them one of their own points. This selection took place after contributions had been made, and the selection was random each round, so participants did not know whether they had been selected for this role when making their initial contribution decisions. Participants played in 1 of 4 conditions: in the 0% bonus condition (N=20) the punisher received no additional resources; in the 10% bonus condition (N=32), the punisher received additional points worth 10% of the value of the Group Pot (after it has been doubled), with corresponding consequences for the 25% (N=36) and 50% (N=24) bonus conditions. These were additional points, i.e. points were not removed from other participants to fund the 'bonus'. Therefore, unlike other studies that have varied marginal benefits, participants not in the punisher role received the same amount as they would have in a standard PGG. This ensured that any pattern of contribution from the conditions in Phase 2 could be directly compared to both Phase 1 and studies that use a homogeneous benefit, i.e. behaviour would only be affected by the resources potentially available to a punisher.

Those selected to be the Punisher were informed of all group members' contributions and their own total earnings for the round. Punishers were told this total included the bonus, but to avoid any anchoring around this value, they were not specifically told how much of the total the bonus represented. While nothing would have directly stopped participants working this out, they were (as stated above) under some time pressure. The study had a mixed-model design with participants acting as their own control group; all participants played Phase 1 before Phase 2 as it has been previously demonstrated that the order of presentation does not affect behaviour in the punishment rounds in this situation (Fehr & Gächter, 2002). The study was run on a collection of networked PCs using zTree (Fischbacher, 2007) and at the end of the experiment, points were converted at a ratio of £1 for every 100 points. For a full version of the instructions, see Appendix D.

6.3.3 Procedure

Participants were each seated in an experimental cubicle (walled to a height of 1.5 meters on three sides) that contained a computer terminal and an instruction sheet that gave a description of both Phase 1 and Phase 2. After 10 minutes these instructions were read out verbatim by the experimenter, and participants were asked to raise their hand if there was anything they did not understand. After the instructions had been read out, participants were presented with a series of questions about contributions and payoff mechanisms to ensure they understood the mechanics of the experiment. All participants answered these questions correctly. Before the study proper began, participants played practice rounds consisting of 4 rounds without punishment and 4 rounds with punishment appropriate to their experimental condition. The practice rounds were included as early decisions made in economic games can be seen as 'mistakes' made by participants due to unfamiliarity with the situation (Anderson, Goeree, & Holt, 1998; Houser & Kurzban, 2002). In total therefore participants played 24 rounds of the Public Goods Game.

6.3.4 Statistical Analysis

The analysis used Generalised Estimating Equations (G.E.E.) modelling in SPSS 20. This approach allows for the potential non-independence of data that could arise because behaviour in rounds will be influenced by the previous one. It also allows for the dependent variables being non-normally distributed as contributions and punishment decisions were skewed towards zero. Non-independence was handled by using an auto-regressive correlation matrix and distribution was corrected for using a Tweedie model. Unless otherwise stated, Round and Phase were entered into the model as within-subject factors, and Bonus as the between-subject factor.

6.4 Results

6.4.1 Punishment severity

Only Phase 2 contributed data for the following analyses, since no punishment took place in Phase 1. Out of 194 opportunities to punish, punishment occurred in 91 (47%) cases with participants spending a mean of 2.4 points on punishment (See Table 6.1). The mean contribution difference between punisher and punished was 4.3 points (SD=5.4). Of the 91 cases of punishment, in 15 (16%) the punisher contributed less than the target; however, in 4 of these cases both punisher and target were very low contributors (for example, the target contributed 2 points, and the punisher 1 point), and in these and 4 further cases the punished participants was the next-lowest contributor in the group. Thus 7 cases of punishment (8% of total punishment and 3% of total possible punishment opportunities) can be considered anti-social punishment. Overall, the majority of punishment in the current study can be considered *altruistic*.

As shown in Figure 6.1, participants spent more on punishment in the conditions where they received a greater bonus (Wald χ^2_3=13.25, p=0.004). With the 0% condition as a comparison, punishers in the 50% (β=3.36, s.e.=1.6, p=0.043) and the 25% (β=1.56 s.e.=0.89, p=0.053) conditions punished more severely. Compared to the 50% condition, only those in the 0% condition spent less on punishment. As we assumed *a priori* that the bonus mechanism would affect punishment severity, a series of pair-wise contrast analyses were performed and found a significant difference in severity of punishment between the 0% condition and 25% (p=0.004) and 50% (p=0.007) conditions. There were also significant differences between the 10% condition and the 25% (p=0.029) and 50% (p=0.028) conditions. As shown in Table 6.1, Rounds marginally affected punishment spending, (Wald χ^2_7=12.83, p=0.08) with less

Table 6.1: mean contributions and spending on punishment across conditions.

	Phase 1 (without punishment)	All rounds	Round 1	Rounds 2	Rounds 3	Rounds 4	Round 5	Round 6	Round 7	Rounds 8
Contributions	0%	3.8 (4.9)	6.5 (6.3)	5.8 (6.4)	3.2 (3.4)	3.9 (4.2)	3.8 (4.2)	2.6 (3.6)	3.1 (5.5)	2 (3.3)
	10%	4.5 (4.6)	7.9 (5.3)	6.0 (5.5)	4.7 (4.2)	5.2 (4.7)	4.2 (4.6)	3.0 (3.3)	2.2 (3.3)	2.8 (3.1)
	25%	5.4 (5.2)	7.9 (4.9)	6.4 (5.7)	5.6 (4.6)	6.0 (5.5)	5.1 (4.5)	4.9 (4.3)	3.6 (4.1)	3.7 (5.2)
	50%	6.3 (4.7)	7.3 (4.2)	7.6 (4.2)	7.6 (5.2)	6.0 (4.5)	6.1 (5.4)	6.9 (5.1)	5.3 (4.7)	3.7 (3.7)
	Phase 2 (with punishment)									
Contributions	0%	3.1 (4.3)	5.8 (5.8)	4.0 (5.1)	3.9 (3.6)	2.3 (2.9)	2.4 (4.1)	1.9 (3.2)	2.3 (4.5)	1.9 (3.5)
	10%	4.8 (3.9)	6.2 (4.1)	6.0 (3.9)	4.9 (3.8)	5.4 (4.2)	4.6 (3.9)	4.0 (3.3)	3.3 (3.3)	3.8 (3.6)
	25%	5.8 (4.5)	6.4 (4.6)	6.9 (6.1)	6.4 (4.6)	6.4 (3.9)	6.0 (4.3)	5.0 (4.4)	4.7 (3.4)	4.6 (3.5)
	50%	7.2 (5.0)	7.0 (4.3)	6.4 (4.4)	7.4 (5.3)	8.7 (5.3)	7.6 (5.9)	6.3 (5.5)	7.0 (4.8)	7.2 (4.8)
Punishment	0%	0.8 (1.9)	1.4 (1.9)	2.8 (4.3)	0.4 (0.8)	0.2 (0.4)	1.0 (1.3)	1.0 (2.2)	0.2 (0.4)	0.2 (0.4)
	10%	1.2 (1.4)	1.6 (1.9)	2.5 (1.6)	1.5 (1.5)	1.4 (1.5)	0.0 (0.0)	0.8 (0.8)	0.7 (1.1)	0.8 (1.2)
	25%	2.2 (3.9)	2.0 (4.7)	1.8 (2.9)	1.8 (2.1)	1.8 (2.4)	0.8 (1.7)	5.0 (8.6)	1.2 (.7)	3.2 (4.3)
	50%	2.6 (3.9)	2.5 (2.5)	3.1 (4.0)	4.8 (6.2)	1.3 (1.6)	2.5 (3.8)	4.0 (5.7)	1.8 (4.0)	1.0 (2.0)

Means were calculated for all rounds and for individual rounds. Standard deviations are in the parenthesis.

punishment occurring in the later rounds. Punishment spending was not affected by an interaction of Round and Bonus (see Table 6.2 for full model summery).

6.4.2 Punishment frequency

As shown in Figure 6.2, participants who received any Bonus also punished more frequently than punishers in the 0% condition (Wald χ^2_3=9.14, *p*=0.027). Compared to the 0% condition, those in the 10% (β=1.45, s.e.=0.53, *p*=0.006), 25% (β=1.15, s.e.=0.51, *p*=0.024) and 50% (β=1.50, s.e.=0.58, *p*=0.01) conditions all punished more frequently. There were no differences between the Bonus conditions. Rounds did not affect punishment frequency, nor was frequency affected by an interaction between Round and Bonus (see Table 6.2).

6.4.3 Contributions (phases 1 & 2)

Table 6.1 shows the mean contributions made by participants across the different conditions. Bonus strongly affected contributions (Wald χ^2_3=11.89, p=0.008), with those in the 50% condition contributing more than those in the 25%, 10% and 0% conditions. Regression coefficients derived from the GEE model show that all bonus conditions differed significantly from the 0% condition (50%, β=1.36, s.e=0.44, p=0.002; 25%, β=0.9, s.e.=0.45, p=0.038;

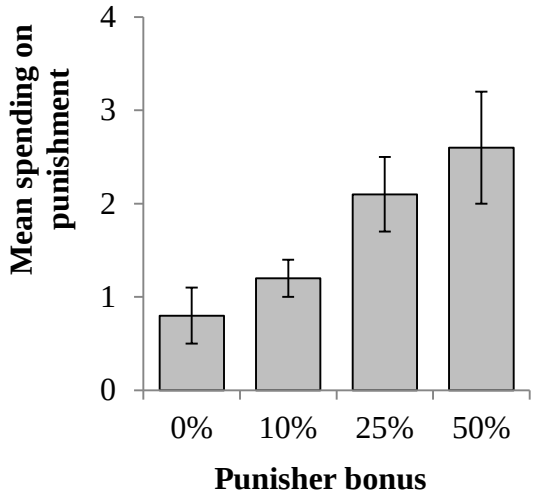


Figure 6.1: punishment severity across punisher bonus conditions. Bars = 1 Standard Error.

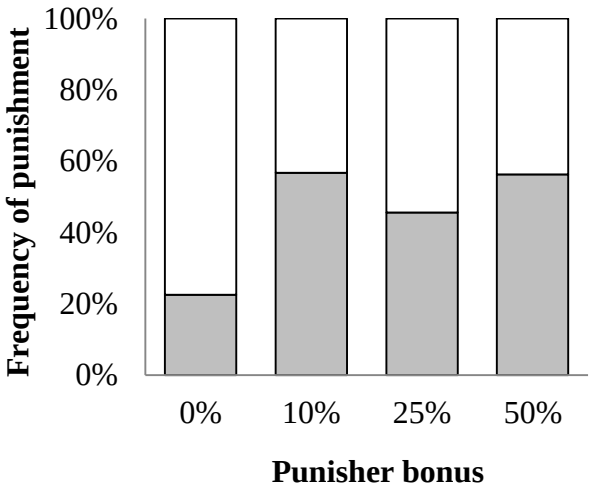


Figure 6.2: percentage of punishment opportunities taken by punishers. Grey=punishment occurred, blank=no punishment occurred.

10%, β=0.72, s.e.=0.45, p=0.01), and compared to the 50% condition, participants contributed less in the 10% (β=-0.42, s.e.=22, p=0.049) and 25% condition (β=-0.64, s.e.=0.21, p=0.03). As is expected in public goods games, cooperation at the beginning of an experimental phase was greater than at the end (Wald χ^2_7=112.21, p<0.001; Round 1 vs Round 8, β=-1.14, s.e.=0.40, p=0.005). However the phase participants were playing did not affect average contributions (See Table 6.2).

As shown in Figure 6.3, contributions were affected by an interaction between Bonus and Phase (Wald χ^2_3=13.11, p=0.004). All bonus conditions showed a difference in contributions between Phase 1 and 2 compared to the 0% condition (50%, β=0.31, s.e.=0.11, p=0.01; 25%, β=0.24, s.e.=0.09, p=0.008; 10%, β=0.27, s.e.=0.09, p=0.005). Compared to the 50% condition the difference in contributions between Phase 1 and 2 for the 0% and 10% was less (0%, β=-0.82, s.e.=0.25, p=0.001; 10%, β=-0.37, s.e.=0.15, p=0.017), with the 25% condition showing a smaller non-significant difference between (β=-0.22, s.e.=0.15, p=0.1).

As shown in Figure 6.4 an interaction of Rounds and Phases also affected contributions (Wald χ^2_7=19.58, p<0.007). Compared to Phase 2 (where punishment was possible)

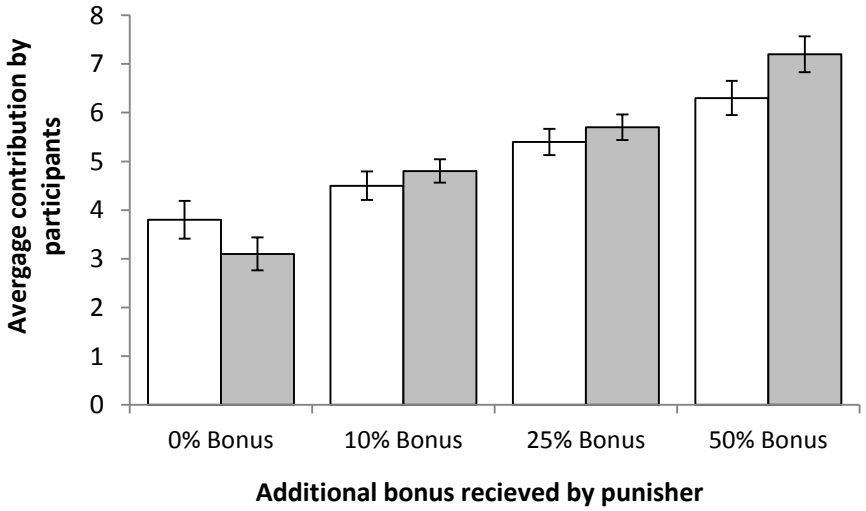


Figure 6.3: mean contributions by participants to the public pot in Phase 1 (blank) and Phase 2 (grey). Bars = 1 Standard Error.

Table 6.2: model summaries

Punishment severity	Factors	Wald $\chi 2$	df	*p*
	Bonus	13.25	3	0.004**
	Rounds	12.83	7	0.08
	Bonus*Rounds	14.24	21	0.86
Punishment frequency				
	Bonus	9.14	3	0.027*
	Rounds	9.23	7	0.24
	Bonus*Rounds	9.24	21	0.98
Contributions				
	Bonus	11.89	3	0.008**
	Phase	0.40	1	0.53
	Rounds	112.21	7	<0.001***
	Bonus*Phase	13.11	3	0.004**
	Bonus*Rounds	73.939	21	<0.001***
	Phase*Rounds	19.58	7	<0.007**
	Bonus*Phase*Rounds	44.32	21	0.02
Phase 1 contributions				
	Bonus	7.61	3	0.06
	Rounds	121.40	7	<0.001***
	Bonus*Rounds	70.411	21	<0.001***
Phase 2 contributions				
	Bonus	14.78	3	0.002**
	Rounds	42.89	7	<0.001***
	Bonus*Rounds	48.77	21	0.001***

*p<0.05, **p<0.01, ***p<0.001

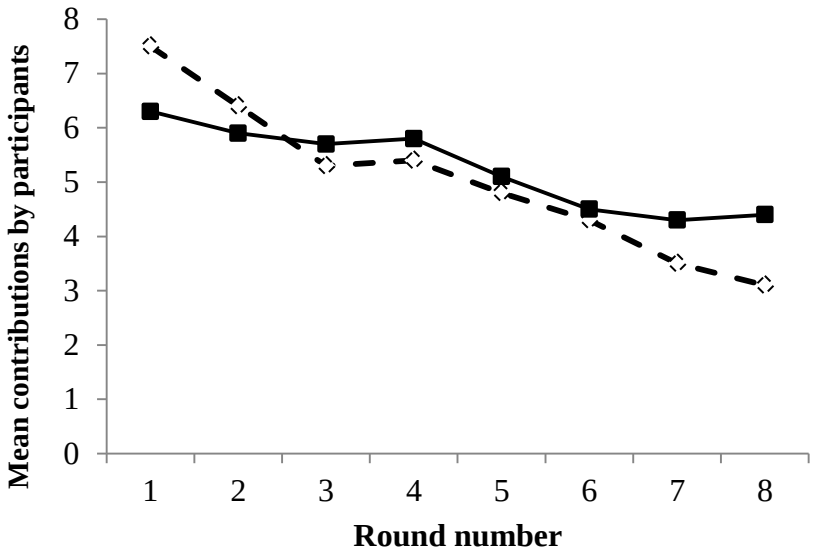


Figure 6.4: mean contributions over time in Phase 1 (without punishment - dashed lines) and Phase 2 (with punishment - solid lines).

contributions started higher in Phase 1 but also dropped off more steeply. As shown in Figure 6.5, contributions were also affected by an interaction between Bonus and Rounds (Wald χ^2_{21}=73.94, p<0.001), with contributions remaining higher in the 50% bonus condition compared to the other conditions. There was a significant three-way interaction between Phases, Rounds and Condition (Wald χ^2_{21}=44.32, p=0.002). As shown in Figure 6.5 while participants in the 50% condition, and to a lesser extent the 25% condition, were more cooperative overall, the level of contributions in these conditions was more stable across rounds in Phase 2.

6.4.4 Phase 2 Contributions

As shown in 6.4.3 there was no main effect of Phase on contributions. Given the interaction between Phase and the other variables, and where the differences lie within these analyses, this is likely because contributions were not maintained in the 0% and 10% conditions, but increased in the 25% and 50% conditions. Also, there may have been some effect of the practice rounds played by participants prior to the experimental rounds (See 6.3.1). Since the study was primarily interested in whether the additional resources changed punishment and

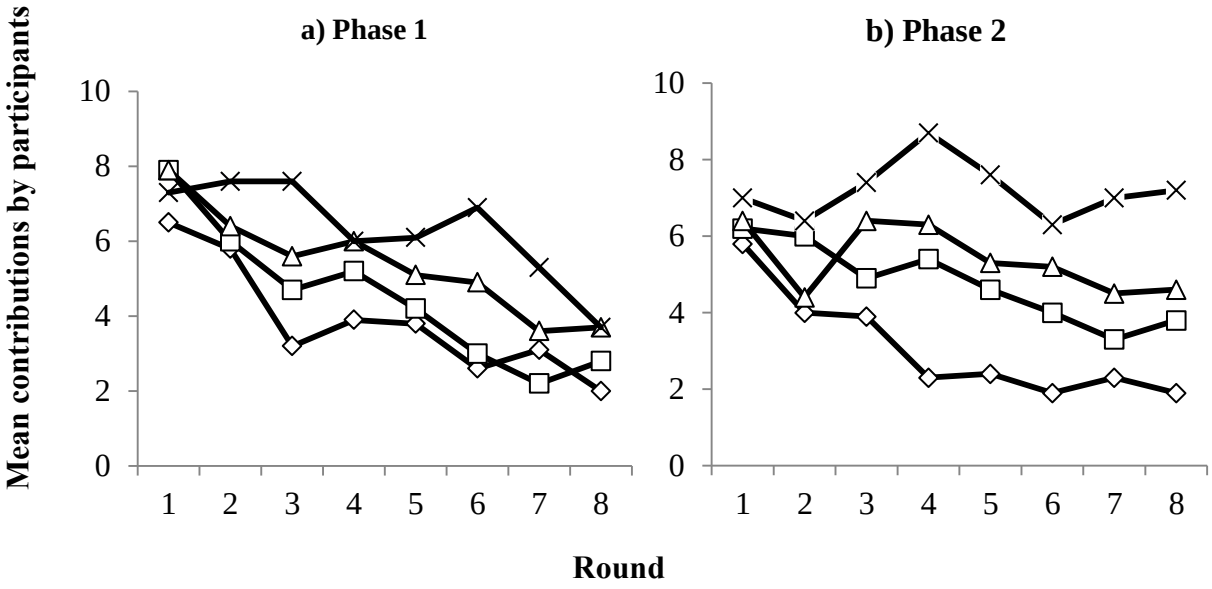


Figure 6.5: contributions over time in Phase 1 (no punishment) and Phase 2 (with punishment) by bonus conditions: diamond=0% Bonus, squares=10% Bonus, triangles = 25% bonus, crosses = 50% bonus.

contribution behaviour, rather than the difference between the potential for punishment and no punishment, the results for Phase 2 were also analysed independently (See 6.4.5 for Phase 1).

As shown in Table 6.1 the bonus that punishers received affected overall contributions with participants in the 0% condition contributing the least and those in the 50% condition the most (Wald χ^2_3=14.79, p=0.002). Compared to the 0% condition, participants in the 25% (β=0.89, s.e.=0.44, p=0.046) and the 50%, (β=1.36, s.e.=0.44, p=0.002) bonus conditions contributed more. Participants in all other bonus conditions contributed less when compared to the 50% condition, (10%, β=-0.64, s.e.=0.21, p=0.003; 25%, β=-0.48, s.e.=0.20, p=0.016). There was also a significant main effect of Rounds on contributions (Wald χ^2_7=42.89, p<0.001) with cooperation decreasing as the number of rounds increased (See Table 1). As shown previously in Figure 6.5b, contributions decreased more in the 0% bonus condition than in the other conditions, with contributions remaining steady in the 50% condition, however this interaction was not significant.

6.4.5 Phase 1 contribution data

While theoretically there should be no effect of condition on the Phase 1 data, as the actual mechanisms in Phase 1 were identical between conditions, the previous analyses suggests that cooperative behaviour might have been influenced by being exposed to differing mechanism in the practice rounds; participants completed practice rounds for both Phase 1 and Phase 2 before beginning the study proper. Therefore the Phase 1 contribution data were also analysed separately (See Table 6.1 for descriptive data).

Condition showed a trend towards Bonus affecting contributions (Wald χ^2_3=7.61, p=0.06) with participants in the 50% condition contributing more compared to the 0% condition (β=0.64, s.e.=0.39, p=0.09). Rounds strongly affected contributions (Wald χ^2_7=121.40, p<0.001) with contributions decreasing as the number of rounds increased (Rounds 1 Vs 8, β=0.69, s.e.=0.17, p<0.001; Round 2 Vs 8, β=0.72, s.e.=0.19, p<0.001; Round 3 Vs 8, β=0.72, s.e.=0.14, p<0.001; Round 4 Vs 8, β=0.47, s.e.=0.16, p<0.004). However, a series of Bonferroni-corrected pair-wise comparisons found no differences between any of the conditions.

Contributions were affected by an interaction between Bonus and Round (Wald χ^2_{21}=70.41, p<0.001). As shown in Figure 6.5a, contributions decreased more in the 0% bonus condition than in the other conditions, with contributions decreasing the least over time in the 50% condition.

6.5 Discussion

6.5.1 Punishment

Dominant individuals have greater access to resources compared to others, and are also in a position to benefit to a greater degree from group cooperation. This can be either because

they can actively monopolise and/or control how others access the results of individual or group activity (Clutton-Brock & Parker, 1995; Cummins, 2005) or because their level of pre-existing resources provides a higher marginal return from group cooperation (Reuben & Riedl, 2013). As a result, individuals in such a position experience a lower net cost in punishment. As the cost of punishment is the key predictor of its occurrence (McCullough et al., 2013; Nikiforakis & Normann, 2008), the current study investigated whether individuals who received an additional benefit from group cooperation would be more willing to engage in costly punishment. This was found to be the case; those who benefitted more from group cooperation punished more severely and more frequently than those who did not.

By providing punishers with more resources, the study made punishment cheaper for them, and as a result these individuals were more willing to punish free-riding within their group. The majority of previous research has made punishment cheap by making it effective, for all (Egas & Riedl, 2008; Falk et al., 2005; Nikiforakis & Normann, 2008), or for certain members of a group (Nikiforakis et al., 2009). The results of the current study can therefore be seen as the opposite side of the coin. That is to say, it has been shown that within and between group variations in effectiveness alters punishment behaviour, and the current study found that variation in resources derived from group-cooperation had a similar effect.

Interestingly however, the finding of the current study contradicts other studies that found that individuals punished at comparable levels regardless of the benefit they received (Reuben & Riedl, 2013). More likely, additional resources did not affect punishment in those studies punishment was effective and was thus cheap even for those who possessed fewer resources. Nevertheless, Reuben & Riedl (2013) did suggest that, regardless of resource allocation, individuals still *agree* on what sort of behaviour *should* be punished. This is evident to an extent in the current study, as all participants who received additional resources

tended to punish only non-co-operators (there were only 7 cases out of 194 where punishment was not directed towards the lowest contributing group-member). This supports the more general finding that when punishment is ineffective, it is almost always directed towards free-riders (Egas & Riedl, 2008; Falk et al., 2005; Sigmund, 2007). Thus, providing a punisher with additional resources did not affect this aspect of an ineffective punishment ratio.

The results therefore support the suggestion that a dominant position can lower the cost of punishment. Dominance plays an important role in individuals' absolute resource levels (Ellis, 1995) and the ability of individuals to access resources or sequester/monopolise any group resources (Cummins, 2005; Gavrilets & Fortunato, 2014; Hawley, 1999). Here it was demonstrated that one way to encourage punishment is to provide a punisher with such additional resources, specifically resources generated from increased marginal returns from group cooperation. Thus, the results of the current study also provide behavioural-data support to the observer/perception results gathered in the Chapter 4. There, dominant individuals were perceived as being more likely to punish social defection and, in the current study, individuals who varied on one dimension of a dominant position (greater marginal returns from group cooperation) did punish more. At the very least, Study 7 demonstrated that one property of a dominant position, access to additional resources, it facilitated a behaviour that all individuals would engage in if they could afford to.

It was also suggested that when disproportionate benefits are derived from group success, this should act as an additional motivation to punish non-cooperation. At this point however, it is not possible to deduce such a motivation. Because group composition and the punisher position were randomised, anyone assigned that role did not have any strategic incentive to punish non-cooperation; they would not benefit from any change in behaviour of the individual they punished because a) they might never interact with this person again, and b)

they might never be in the punisher/bonus role again. Therefore, punishment could have been entirely altruistic (Fehr & Fischbacher, 2003; Fehr & Gächter, 2002); individuals do feel the desire to punish (Crockett et al., 2010; Falk et al., 2005), and punishment still occurs when there is no possibility of the punishment affecting future behaviour (Fehr & Gächter, 2000). Thus, giving punishers a bonus lowered the cost of punishment enough so that individuals were able and willing to act on this altruistic motive.

However, it should be remembered that participants in a public goods-style games are not disinterested observers of the outcome of group cooperation, and that the current study linked the additional resources available to punishers to the result of group cooperation. So while the random group structure meant punishment could not have been used strategically to alter behaviour (Masclet, 2003), non-cooperative behaviour did impact the punisher. An alternative motivation is therefore less pro-social; participants may have been motivated to punish out of spite (Jensen, 2010; Leibbrandt & López-Pérez, 2011) because free-riders had explicitly cost them additional points in that round. The primary purpose of spite is to harm the target rather than to directly benefit the punisher (Jensen, 2010; West et al., 2007) and therefore it would not be affected by the lack of strategic benefit from punishing. Therefore punishment in random-group settings (for example, Fehr & Fischbacher, 2003) might not be a result of anger caused the violation of cooperation or fairness norms per se, but because such violations directly cost the punisher resources. This would explain why punishment frequency was equal across the benefit conditions - all lost more resources compared to the 0% Bonus condition - and why severity scaled with the benefit; punishers in the higher bonus conditions could inflict spiteful punishment at a lower cost.

Finally the pattern of results also partially rebuts an alternative, demand characteristics, explanation. It is possible that, when given additional benefits and additional responsibilities,

participants assumed they were expected to use the former for the latter. Such an explanation may account for some of the observed behaviour; when alternative options are available to participants, for example the ability to reward co-operators (Rand, Dreber, Ellingsen, Fudenberg, & Nowak, 2009), punishment spending does drop. However if this had been the case here it would be logical to expect both spending and frequency of punishment to increase with the value of the bonus. In practice, whereas punishment severity did scale with the value of the bonus (i.e. the relative cost of punishment), punishment frequency did not. All those who received additional resources punished at a similar frequency, i.e. punishment decisions were not simply the result of a lack of other behavioural options.

6.5.2 Cooperation

While not the focus of the study, it was predicted that because a well-resourced punisher would punish more severely, their presence would also increase cooperation. This proved to be partially the case; average contributions were higher when the punisher received any bonus, and contributions over rounds were more stable, especially in the 25% and 50% bonus conditions. Thus, while ineffective-punishment (1:1 ratio) generally does not maintain contributions over time (Egas & Riedl, 2008; Nikiforakis & Normann, 2008), in the current study contributions were greater (relative to the no-bonus condition) when the punishers received additional resources. This might suggest that participants did believe the well-resourced punishers would punish free-riding in their groups, i.e. that individuals with the means to punish would punish more severely – as indeed was the case (See 6.4.2.)

This result supports the suggestion introduced in previous chapters that individuals will not want to provoke a response from a dominant group member, when they might be willing to do so from a subordinate or weak one. In Chapter 5, participants predicted that individuals would withdraw their attempt at anti-social behaviour when a dominant group member

indicated an interest. Here, the increased level of contributions to the public good does suggest that individuals were responding to the likelihood of being punished for non-cooperation by a more dominant individual. In this case, what made the punisher dominant was the amount of resources they held, rather than their ability to punish effectively.

Providing punishers with an additional benefit from group cooperation resulted in more sustained group cooperation. It should be noted however that while cooperation did not decay in the higher resource conditions, it did not increase either. In reference to Nikiforakis & Normann (2008, Figure 2) specifically, the pattern of results resembles the contributions seen when punishment was moderately effective (ratio of 1:2). Therefore, punishment that was cheap-relative-to-resources did not result in the level of cooperation seen for punishment that is cheap due to its effectiveness, but the former did maintain cooperation compared to moderately effective punishment (see also, Egas & Riedl, 2008). This comparison is important partly because very little research has investigated how heterogeneity in resources alone affects behaviour: studies that have varied heterogeneity in marginal benefit have done so using highly or moderately effective punishment (for example, Burns & Visser, 2006; Nikiforakis et al., 2012). The current study does suggest that when the cost of punishment is lowered via accessible resources alone, ineffective punishment can at least moderately promote cooperation.

Nevertheless, there are a few issues with the cooperation data that must be mentioned before any firm conclusions can be drawn. A punisher's additional benefit was predicted to have some effect on cooperation, but in the overall model there was no difference in contributions between Phase 1 and 2. The lack of a difference may partly be explained, as shown in Figure 5.3, by the fact that contributions in Phase 1 started at a higher level, but there is some evidence that contributions were affected by way the study was introduced to participants. It

is likely that experiencing the Phase 2 practice round prior to the Phase 1 experimental rounds affected behaviour in the latter. Without such an effect, the Phase 1 rounds would not have demonstrated a difference between conditions, because there was no punishment. Yet, as shown in Figure 6.4, the pattern of cooperation in Phase 1, if not the actual amount, is consistently arranged by the Bonus value that would be available in Phase 2; i.e. those in the 50% condition contributing the most, followed by 25% etc. This does suggest that giving the participants the full instructions for both Phase 1 and 2 at the start of the experiment had some influence on their behaviour in Phase 1.

In Phase 2, there could be a separate, more economic, explanation for the variation in cooperative behaviour between conditions. Public goods games are a measure of cooperation insomuch as, in a 4-player game, for each point an individual contributes to the group, they lose 0.5 while all others gain 0.5. Hence, it is beneficial to free-ride on the cooperation of others. However, the mechanism used in Phase 2 of the current study altered this payoff ratio. An individual in the 50% condition, for example, had a ¼ chance of becoming the punisher, and thus receiving 50% of the group pot, equivalent to a 33% return on every point they themselves contributed to the group pot, regardless of the activity of other group members. E.g. had they contributed 10 points and everyone else 0, the punisher would still have received 15 points. Those in the 25% condition had ¼ chance of not losing any points regardless of their contributions, and those in the 10% had a ¼ chance of losing only 0.3 points per contribution. It has been suggested that individuals might choose to invest in the public good if it is the only way for them to protect or increase their resources (Barker, Barclay, & Reeve, 2013; Burns & Visser, 2006), and the bonus was an opportunity for participants to gamble to receive an additional reward. And while it is unlikely participants did the specific calculations for risk and rewards of contributing to the group pot, the

recognition of some potential gain may still have increased contributions to the levels observed.

Finally, perhaps cooperation was affected by punisher bonus because the bonus itself was an inducement to cooperate, as opposed to the fear of punishment by a well-resourced punisher. Even if this was the case, it has been demonstrated that experiencing a benefit from the public good is itself a possible reason why dominant individuals might be more cooperative (Gavrilets & Fortunato, 2014). However, it should be noted that Reuban & Riedl (2013) found that individuals who received a similar 50% bonus did not contribute more than others, despite a fixed group structure that guaranteed them a greater return per investment in the public good (see also, Tan, 2008). Therefore, if individuals who were guaranteed a higher return on their contributions, and in the presence of effective punishment which could be directed towards them, did not increase their contributions, it is questionable whether such an effect would occur in the current study, when the bonus was uncertain and the punishment ineffective.

6.5.3 Conclusion

Heterogeneity in both the effectiveness of punishment and the net cost of punishment has been shown to be important for the evolution of costly punishment (de Weerd & Verbrugge, 2011; Frank, 1996). Such heterogeneity can be caused naturalistically by differences in dominance within a social hierarchy. Dominant individuals are certainly able to inflict effective punishment on others through physical prowess (Clutton-Brock & Parker, 1995; Huston et al., 1981), number of social allies (Gavrilets et al., 2008; Mathew & Boyd, 2011), or prestige/political power (Henrich & Gil-White, 2001). Dominant individuals are also in a position to access resources (Jones & Rachlin, 2006), to receive them as part of social bargaining (Barclay, 2013), and to acquire them coercively (Hawley, 1999; Sell, Tooby, et

al., 2009). Yet while effective punishment has been well studied over the years, the effect of resources on punishment has received substantially less attention. The current study found that, when provided with a greater marginal return from group-level cooperation, and thereby given one of the attributes of a dominant position, participants were more willing to engage in punishment, and spent more on punishment, as the amount of addition resources increased. Making punishment relatively cheap for some participants made them more likely to punish.

This may be because the lowering of the cost of punishment allowed individuals to act altruistically by punishing free-riders because they behaved contrary to social norms of fairness and cooperation. Alternatively, marginal benefit may have encouraged punishment out of spite because, by free-riding, other group members deprived the punisher of a good amount of additional resources. Both these mechanisms are consistent with dominance playing a role in this behaviour.

6.6 Study 8: Private gain and the public good - monopolisation of group resources by punishers' increases spending on punishment

Study 7 simulated a dominant position by providing punishers with greater marginal return from group cooperation. It found that providing punishers with such an additional benefit increased the frequency and severity of punishment. However, the motivation for this remains in doubt; were punishers acting in such a way because their resource-level simply allowed them to punish more cheaply, or because they had a higher stake in the outcome of any group cooperation? This is important to answer as, if it is the latter, then this would suggest that dominant individuals punish not simply because they have the resources to, but because a dominant position gives individuals an strategic motivation to punish non-cooperation and 'unfairness' that is not present in subordinates. This would show a stronger direct link

between costly punishment and dominance, and would support the suggestion that the punishment of anti-social behaviour can be seen as a fundamentally selfish behaviour tied to dominance, rather than to altruistic and other-regarding preferences. Study 8 investigated this by varying the strategic potential of punishment; as well as giving punishers a greater return from punishment, punisher role and groups were also fixed in some conditions.

Study 8 also added an additional benefit mechanism. While punishers in Study 7 received a greater marginal return on cooperation, they did not do so at the expense of other group members. This is not the case in real-life dominance hierarchies, where dominant individuals tend to monopolise resources (Cummins, 2005) or otherwise attract them over others (Barclay, 2013; Henrich & Gil-White, 2001); obviously, by taking additional resources, dominant individuals are depriving those below them in the hierarchy of them. Thus Study 8 sought to simulate this by making it clear to all participants that the additional resources available to the punisher were at the expense of the non-punishers. This is referred to below as the 'monopoly' condition.

It was predicted that participants would punish with greater frequency and severity when they received either a bonus (as they were in Study 7) or a monopoly-bonus as opposed to no additional resources. No predictions were made concerning the difference between the two types of bonus mechanism. It was also predicted that when groups were fixed, punishment would be more severe and more frequent than when groups were random (as in Study 7), with punishment being the most severe and frequent when the punisher also received an additional benefit from group cooperation. Again, no predictions were made about any differences between the two types of benefit.

In terms of cooperative behaviour it was predicted that participants in the fixed groups should be more cooperative than those in random groups (see, Fehr & Gächter, 2000). Also, in

reference to fixed groups specifically, in response to questions raised in Study 7 it was also predicted that punishers who were received the greater benefits from group-level cooperative behaviour would contribute more than those who did not. Also in regards to Study 7, it was predicted that punishers who received any form of bonus in the fixed groups would contribute the most.

6.7 Method

6.7.1 Participants

144 undergraduate students (99 female) were recruited through the University of Exeter's paid-participant recruitment website. Mean age of participants was 20. In total, seven experimental sessions were conducted with a range of 12-24 participants in each session. The mean payment received by participants was £5.52 and the mean duration of each session was 40 minutes. All participants passed the comprehension check included with the experiment instruction sheet. Prior to analysis, two additional groups were removed because, despite warnings on the study advertisement, members of these groups were discovered to have taken part in a previous session of the study. These groups are not included in the data given above.

6.7.2 Experimental design

The experiment consisted of a modified version of a public goods game with punishment using the zTree software (Fischbacher, 2007). At the start of each round, participants were randomly sorted into groups of 4. They were given an allocation of 20 points and had the option to contribute between 0-20 points to a group pot. This total was then doubled and divided amongst all group members. Once this had occurred, all participants were shown the contribution decisions of others in their group.

Following this, one participant in each group was randomly assigned the ability to punish, referred to as being able to "assign deduction points" in the instructions. They could do so at a ratio of 1:1; every point they removed from another player would cost them one of their own points. The selected group member was presented with a list of the contributions made by other members and their own total earnings from the round (see below), and could choose to spend between 0-20 points on punishing a single group member of their choice. Once this decision had been made the punished group member was informed they had been punished and by how much, but not who punished them. At this point the round was over and the next began.

The study manipulated both the 'Stability' (Stranger or Partner protocol) of the groups and the mechanism by which the punisher received their additional resources, henceforth 'Benefit' (25% bonus, 25% monopoly-bonus, or no-bonus - described below), giving the study a 2x3 design. At the end of the experiment, the total points earned by each participant were converted to pounds at a ratio £1:100 points.

6.7.3 Benefit mechanisms[7]

Participants played in one of three conditions which manipulated how the punisher's additional resources were generated. In the 'bonus' condition, the punisher received additional points to the value of 25% of the group pot total, i.e. of the total was 80 points, the punisher would receive 20 points and the 80 points were divided equally between all four group members. As in Study 7, these points were not deducted from any other member of the group.

[7] Strictly speaking, the conditions represent different marginal gains between participants from group cooperation. However, the conditions are described below as they were described to participants. Also, keeping these descriptions throughout emphasises the specific differences in how resources were divided.

In the 'monopoly' condition, punishers also received additional points to the value of 25% of the group pot total. However this value was taken from the group pot prior to its division, i.e. if the total pot was 80, 20 points would be removed and given to the punisher, and the remaining 60 points were divided equally between all four group members.

In the 'no-bonus' (control) condition, the punisher did not receive any additional resources. i.e. if the total was 80 points, the punisher would receive 25% of these, 20 points, with other group members each receiving 20 points.

At the punishment screen punishers were shown their total earnings for the round so far and told this total included the bonus, if any. However, to avoid any anchoring effect, they were not told how many points were represented by the additional benefits, if any, they received.

The 25% value was chosen for four reasons a) in Study 7 this value produced similar results as the 50% bonus, b) in terms of the punisher's own contributions, the return on investment is zero, c) individuals in standard 4-player public goods games receive 25% of the group pot, and d) it was felt that 50% provided such a difference in resources that it might have overshadowed any differences between the experimental conditions.

6.7.4 Stability

Participants played in either 'Random' groups (groups were randomly organised each round) or 'Fixed' groups, (groups comprised of the same individuals each round). This corresponds to the 'stranger' and 'partner' protocols respectively in standard terminology (Fehr & Gächter, 2000). These terms are not used here due to the punisher-selection mechanism employed in the current study. In the case of the fixed groups, the punisher remained constant throughout the game, however in the random groups the punisher was randomly selected each round. For a full version of the instructions, see Appendix D.

6.7.5 Procedure

Participants were each seated in an experimental cubicle (walled to a height of 1.5 meters on three sides) that contained a computer terminal and an instruction sheet that gave a description of Stage 1. After 10 minutes these instructions were read out verbatim by the experimenter, and participants were asked to raise their hand if there was anything they did not understand. After the instructions had been read out, participants were presented with a series of questions about contributions, payoffs and the group structure to ensure they understood the mechanics of the experiment. All participants answered these questions correctly. Participants played for 10 rounds and, to avoid any end-round effect, were told there would be between 6-12 rounds the game.

It has been suggested that many decisions in economic games are due to the novelty of the situation or a lack of understanding of the game rules by participants (Andreoni, 1995). Therefore, before the study proper began, participants played practice rounds consisting of 4 rounds in their experimental condition. In the Partner protocol, participants were told that while the punisher role **would be fixed in the actual game** [emphasis as in instructions], in the practice rounds it would be randomly assigned. This was to ensure that, potentially, every group member had a change to practice-play the punisher role. In total therefore participants played 14 rounds of the Public Goods Game.

6.7.6 Statistical analysis

The analysis used Generalised Estimating Equations (G.E.E.) modelling in SPSS 20. This approach allows for the potential non-independence of data that could arise because behaviour in rounds will be influenced by the previous one. It also allows for the dependent variables being non-normally distributed as contributions and punishment decisions were skewed towards zero. Unless otherwise stated, non-independence was allowed for by using

an auto-regressive correlation matrix and distribution was corrected for by using a Tweedie model.

Due to the nature of the data, analysis was conducted at the group level (N=36). In the stranger conditions groups and punisher were random, and in the partner conditions groups and punishers were fixed. Analysing behaviour at the group level therefore can be seen as comparing group-level cooperative and punishment behaviour between two extremes, completely *random* groups (random punishers and groups) where there could be no strategic motivation to punish, and completely *fixed* groups (fixed punisher and groups), where there was a strong strategic motivation. To reduce the effects of random fluctuations, Rounds were condensed into pairs, i.e. Round 1-2, Rounds 3-4 etc for the analysis. Unless otherwise stated Round was entered into the model as a within-subject factor, with Bonus and Stability entered as between-subject factors.

6.8 Results

6.8.1 Punishment severity

Out of 360 opportunities to punish, punishment occurred in 139 (39%) cases with participants spending a mean of 4.4 points on punishment (See Table 6.3). The mean contribution difference between punisher and punished was 5 points (SD=7), and of the 139 cases of punishment in 22 (16%) the punisher had contributed less than the target; however, in 12 cases both punisher and target were very low contributors (for example, the target contributed 2 points, and the punisher 1 point), and the punished participant was the next-lowest contributor in the group.

Thus 10 cases of punishment (7% of total punishment and 2% of total possible punishment opportunities) can be considered anti-social punishment. As with Study 7, overall the punishment behaviour of participants in the current study can be considered *altruistic*.

As shown in Figure 6.6, punishment severity was affected by the benefit received by the group punisher (Wald χ^2_2=21.21, p<0.001). Compared to the 0% benefit condition, only those in the 25% monopoly condition spent significantly more on punishment (β=2.29, s.e.=0.7, p=0.01). As we hypothesised *a priori* that the bonus mechanism would affect punishment behaviour, a series of pair-wise contrast analyses were performed and found a significant difference in punishment spending between the No-bonus and Monopoly groups (p=0.001) and between the no-bonus and bonus conditions (p=0.002). There was also a marginal difference in punishment severity between the bonus and monopoly conditions (p=0.07).

However there was no significant difference between the bonus and monopoly-bonus conditions (p=0.1). Participants in the fixed groups did not spend any more on punishment overall than those in the random groups (See Table 6.4 for full model summery).

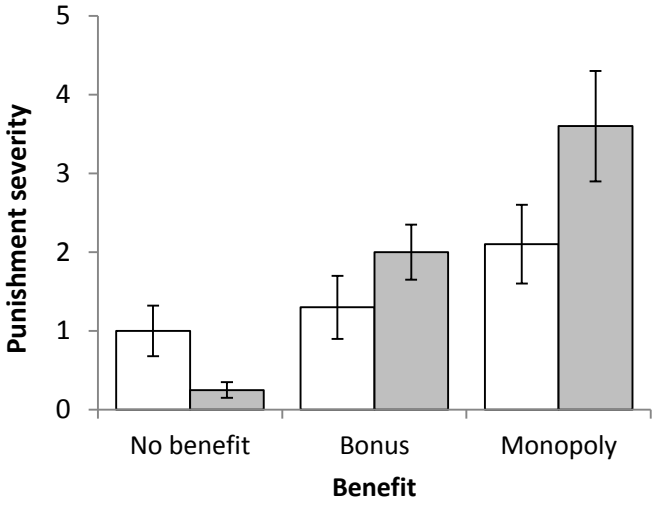


Figure 6.6: severity of punishment across the different types of benefit available to punishers, when groups were random (blank) or fixed (grey). Bars = 1 Standard Error.

Table 6.3: mean of group-level contributions and spending on punishment across conditions.

	Contributions	All rounds	Rounds 1-2	Rounds 3-4	Rounds 5-6	Rounds 7-8	Round 9-10
Random groups	No bonus	7.7 (3.8)	11.3 (3.9)	10.0 (4.3)	8.9 (4.0)	9.1 (3.8)	8.6 (4.8)
	25% bonus	7.7 (3.2)	11.0 (3.6)	10.1 (3.5)	8.5 (3.1)	8.3 (3.8)	7.8 (4.1)
	25% monopoly bonus	7.9 (2.8)	11.9 (2.8)	10.7 (3.4)	10.1 (3.5)	8.4 (2.9)	7.8 (3.8)
Fixed groups	No bonus	11.1 (3.9)	12.5 (3.5)	11.5 (4.0)	9.4 (4.6)	11.2 (2.6)	10.8 (4.8)
	25% bonus	10.2 (3.8)	12.0 (4.1)	11.0 (4.3)	9.4 (2.4)	9.5 (3.2)	9.3 (4.5)
	25% monopoly bonus	12.0 (3.1)	14.0 (1.9)	12.9 (2.3)	12.6 (2.6)	10.5 (2.5)	9.8 (4.0)
	Punishment						
Random groups	No bonus	1.0 (2.3)	0.9 (1.6)	0.8 (2.3)	0.2 (1.1)	0.4 (1.1)	0.6 (2.2)
	25% bonus	1.3 (3.1)	1.9 (1.9)	2.9 (5.4)	1.6 (2.7)	1.2 (2.2)	1.1 (2.3)
	25% monopoly bonus	2.1 (4.4)	4.0 (6.0)	2.0 (3.9)	1.7 (3.1)	4.3 (5.9)	2.0 (3.1)
Fixed groups	No bonus	0.3 (0.8)	0.8 (0.3)	0.1 (0.2)	0.0 (0.0)	0.1 (0.2)	0.3 (0.6)
	25% bonus	2.0 (3.1)	2.2 (2.0)	3.6 (5.2)	1.8 (2.0)	1.8 (2.7)	0.9 (2.6)
	25% monopoly bonus	3.6 (5.1)	5.0 (7.4)	3.2 (4.8)	2.3 (3.2)	5.3 (6.0)	2.4 (3.1)

Means were calculated for all rounds and for individual rounds. Standard deviations are in the parenthesis.

As also shown in Figure 6.6, punishment severity was affected by an interaction between Stability and Benefit (Wald χ^2_2=10.16, p=0.006). Specifically, compared to the No-benefit condition there was a marginally significant difference in severity between random and fixed groups for those in the Monopoly condition (β=1.82, s.e.=1.0, p=0.08). Punishment severity was not affected by an interaction between Stability and Rounds, nor was it affected by an interaction between Benefit and Rounds. There was also no three-way interaction effect of Stability, Benefit and Round on punishment severity (see Table 6.4).

6.8.2 Punishment frequency

We also investigated whether Stability and Benefit affected the frequency of punishment i.e. whether punishers within the groups chose to spend any amount above zero punishing. The 'Frequency' with which an individual in a group/round did or did not punish was entered into the G.E.E. using a binary logistic model. Due to the number of comparisons and the nature of the data the three-way interaction resulted in instability in the model, therefore it was excluded from the analysis.

As shown in Figure 6.7, punishers who received any sort of bonus punished more frequently compared to those in the no-bonus groups (Wald χ^2_2=18.37, p<0.001). Compared to the no-bonus condition, punishers in the bonus (β=1.56, s.e.=0.78, p=0.06) and monopoly (β=2.54, s.e.=0.96, p=0.033) conditions punished significantly more often. Monopoly and bonus conditions did not differ significantly in the frequency of punishment (β=-0.98, s.e.=0.83, p=0.24). Punishment was more frequent in Rounds 1-2 (52% of opportunities) than at Rounds 9-10 (33%; Wald χ^2_4=11.63, p=0.02), however the regression coefficients generated by the G.E.E. showed no specific differences between Rounds 9-10 and other Round pairs. Fixed groups did not punish more often than random groups.

Table 6.4: model summaries

Punishment severity	Factors	Wald $\chi 2$	df	p
	Benefit	21.21	2	<0.001***
	Stability	0.49	1	0.48
	Rounds	7.87	4	0.09
	Benefit*Stability	10.16	2	0.006
	Benefit*Rounds	10.77	8	0.22
	Stability*Rounds	3.62	4	0.46
	Stability*Benefit*Rounds	9.84	8	0.20
Punishment frequency				
	Benefit	18.37	2	0.001***
	Stability	0.24	1	0.63
	Rounds	11.63	4	0.02*
	Benefit*Stability	3.27	2	0.15
	Benefit*Rounds	3.70	8	0.16
	Stability*Rounds	1.02	4	0.22
Contributions				
	Benefit	1.88	2	0.28
	Stability	30.39	1	<0.001***
	Rounds	46.81	4	<0.001***
	Benefit*Stability	0.85	2	0.65
	Benefit*Rounds	3.54	8	0.90
	Stability*Rounds	6.43	4	0.17
	Stability*Benefit*Rounds	5.80	8	0.67
Punisher contributions				
	Benefit	6.95	2	0.031*
	Stability	8.62	1	0.003**
	Rounds	15.55	4	0.004**
	Benefit*Stability	2.92	2	0.23
	Benefit*Rounds	7.22	8	0.18
	Stability*Rounds	5.97	4	0.51
	Stability*Benefit*Rounds	17.72	8	0.023*
Fixed Punishers contributions				
	Benefit	8.30	2	0.016*
	Rounds	13.36	4	0.01**
	Benefit*Rounds	7.14	2	0.52

*p<0.05, **p<0.01, ***p<0.001

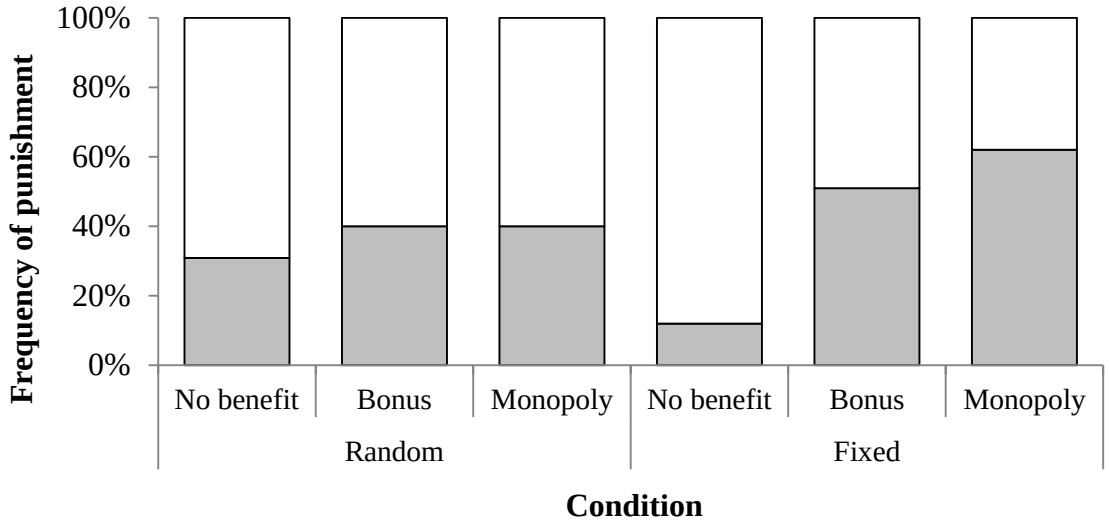


Figure 6.7: percentage of punishment opportunities taken by punishers across benefits conditions when groups were random or fixed. Grey =punishment occurred, Blank = punishment did not occur.

The decision to punish was not affected by an interaction between Stability and Bonus (Wald χ^2_4=3.27, p=0.15). However, as shown in Figure 6.7, the punishment frequency did show a trend towards a significant difference between unstable and stable conditions in the Monopoly condition compared to the no-bonus condition (β=1.78, s.e.=1.0, p=0.09). Rounds did not have any effect on the decision to punish, nor did Rounds interact with Benefit or Stability. In both these cases there were effect sizes that approached significance but did not reach it.

6.8.3 Contributions

The stability of the group affected overall contributions, with random groups behaving less cooperatively (M=7.8, SD=3.2) than fixed groups (M=11.0, SD=3.7; Wald χ^2_1=30.39, p<0.001; β=-0.45, s.e.=0.2, p=0.036). As shown in Table 6.3, as is often the case in public goods games, contributions decreased as the game progressed (Wald χ^2_4=46.81, p<0.001), with participants being more cooperative at the beginning of the game (Round 1 Vs Round 10, (β=0.36, s.e.=0.18, p=0.041). Benefit did not affect overall cooperation. Contributions were not affected by any second or third order interactions (See Table 6.4).

6.8.4 Punisher behaviour

A question raised in Study 7 was whether individuals were strategically raising their contributions in the hopes of receiving additional resources if they were assigned the punisher role. To test this initially, contributions by participants assigned as Punishers were entered as a dependent variable into the model. While those in a random group could not have known they would receive the additional resources, by comparing the randomly selected punisher to those who *knew* they would receive the benefit, it is possible to see whether at a group level behaviour differed.

As might be expected given the group-contribution data (see 6.8.3), punishers in the fixed conditions (M=11.03, SD=6.3) contributed more than those in the random conditions (M=8.6, SD=5.8; Wald χ^2_1=8.62, p=0.003; β=0.4, s.e.=0.32, p=0.1). Contributions were also higher at Round 1-2 (M=11.94, SD=5.0) than Rounds 9-10 (M=8.43, SD=6.7; Wald χ^2_4=15.55, p=0.004; β=0.24, s.e.=0.1, p=0.1). Interestingly, as shown in Figure 6.8 there was also an effect of Benefit (Wald χ^2_2=6.95, p=0.033), with those in the Bonus condition contributing

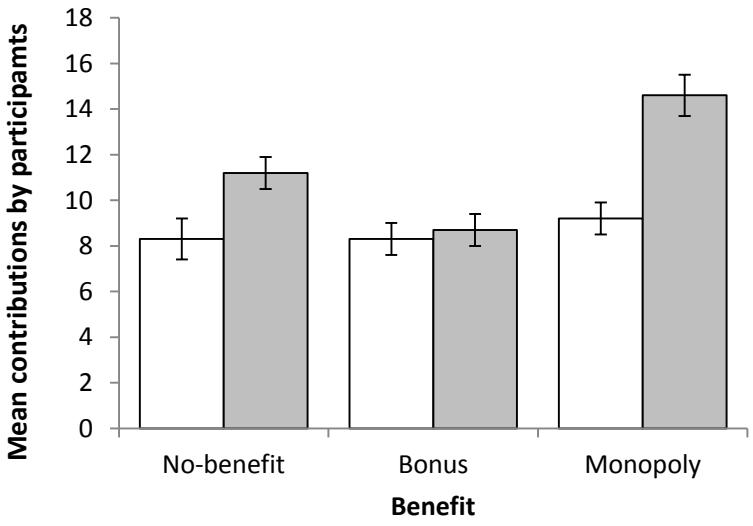


Figure 6.8: mean contributions by punishers across benefit conditions, when groups were Random (blank) or fixed (grey). Bars = 1 Standard Error.

less than those in Monopoly condition (β=-0.67, s.e.=0.34, p=0.05).

Given the lack of a main effect of Benefit on overall contributions (see *6.8.1*), this suggests that the effect in the punisher-only data was being driven by behaviour in the Stable condition, however while Figure 6.8 suggests a trend in this direction, there was no interaction between Stability and Benefit on punisher contributions. There was also no interaction effect of Stability and Rounds (Wald χ^2_2=5.97, p=0.23) or Benefit and Rounds. As shown in Figure 6.9, there was however a three-way interaction (Wald χ^2_8=17.72, p=0.023), with a clearer difference seen between the benefit conditions when the groups were stable.

Thus, the effect of Benefit might be driven by the fixed group. To test this, a separate analysis was run on this subset of the data (N=19). To control for the general cooperation-enhancing effect of stability, a new variable 'deviation from group mean' was created by subtracting the punisher's contribution from the mean group-level contribution of the three other participants in that group/round; thus a positive figure suggests punishers contributed less than the group as a whole, and a negative figure that they contributed more. Due to the normal distribution

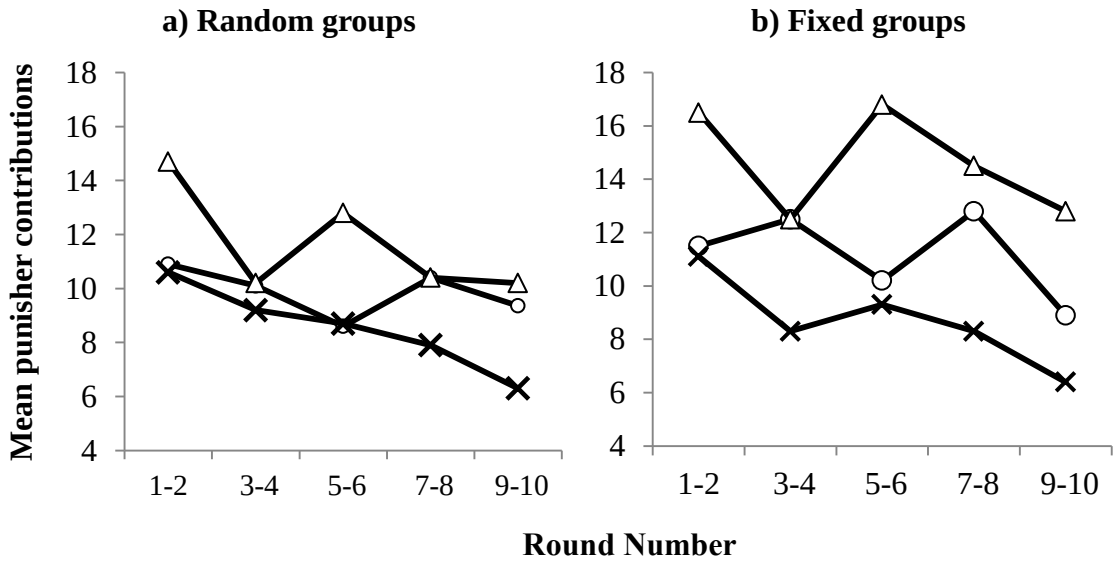


Figure 6.9: contributions by punishers over time when groups were random (a) or fixed (b): Circles = no-benefit condition, crosses = bonus condition, triangles = monopoly condition.

of the deviation data, a linear model was used.

As shown in Figure 6.10, Benefit affected the difference in punisher contributions (Wald χ^2_3=8.30, p=0.016). Punishers in the Monopoly condition contributed more compared to the group mean than those in the Bonus (β=5.72, s.e.=1.4, p<0.001) and No-benefit conditions (β=4.95, s.e.=1.31, p<0.001). Punisher contributions matched the group-mean more at the start of the game (M=0.01, SD=3.5) than at the end (M=1.0, SD=4.6; Wald χ^2_4=8.30, p=0.016). However the regression coefficients did not demonstrate individual differences between Rounds 9-10 and other rounds. An interaction between Benefit and Rounds did not affect relative difference.

6.9 Results 2: group level effects

The current study (and the current chapter overall) are primarily concerned with the effects that additional resources and the method of their accumulation would have on punishment and cooperative behaviour. Thus, group-level efficiency was not a concern per se. However, punishment does strongly affect group efficiency (Gächter et al., 2008), and this is the

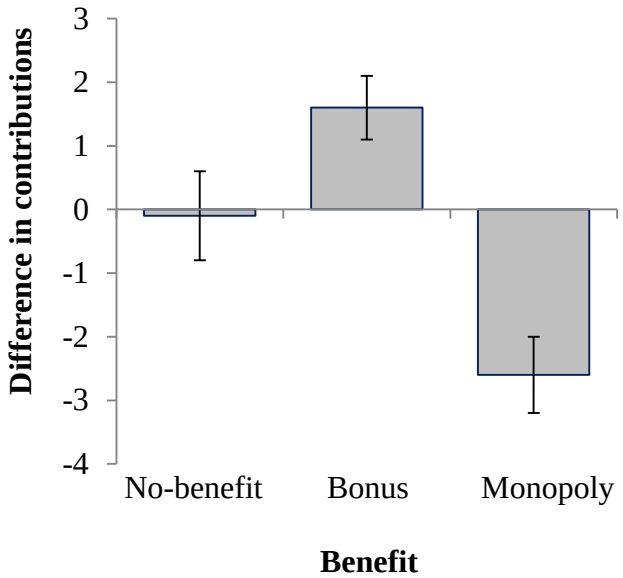


Figure 6.10: difference between punisher contributions and mean of other group-members between benefit conditions. A negative value suggests the punisher contributed more than the mean of non-punishers. Bars = 1 Standard Error.

mechanism by which group selection is suggested to act (Bowles & Gintis, 2004; Denga et al., 2011). Equally, there has been debate as to the role of leadership in coordinating collective action and in the cost/benefits of such a position (Gillet et al., 2010; Maner & Mead, 2010). Therefore, an additional set of analyses investigated the group efficiency and actual earnings of participants in the current study.

6.9.1 Group efficiency

Group efficiency, i.e. a measure of the social product of the group was calculated by subtracting the amount of punishment (given and received) from the group pot. This was carried out as punishment can be considered a 'waste' of group resources. These data were normally distributed and accordingly a linear model was used in the analysis.

As shown in Figure 6.11a, Stability affected group efficiency with fixed groups being more efficient (Wald χ^2_1=18.47. p<0.001; β=29.02, s.e.=14.01, p=0.038). As also shown in Figure 6.11a, Benefit affected group efficiency (Wald χ^2_2=5.53. p=0.06), although the regression coefficients did not show any significant differences between conditions. Round also affected group efficiency, with efficiency being greater in the early rounds compared to the later rounds. (Wald χ^2_4=34.35. p<0.001; Rounds 1-2 vs Rounds 9-10, β=33.83, s.e.=15.28, p=0.027). There was also a marginally significant interaction between Benefit and Stability (Wald χ^2_2=4.91. p=0.08). Group efficiency was not affected by interactions between Benefit and Rounds, Stability and Rounds, nor was efficiency affected by a three-way interactions between Benefit, Stability and Rounds (See Table 6.5 for model summery).

6.9.2 Group punisher earnings

Group-punisher earnings were calculated by adding the additional return from the benefit mechanism to their share of the group pot, minus punishment. As with 6.8.4, while those in a

random group could not have known they would receive the additional resources, by comparing the randomly selected punisher to those who *knew* they would receive the benefit, it is possible to see whether at a group level, the study IVs affected earnings. These data were normally distributed and accordingly a linear model was used in the analysis.

As shown in Figure 6.11b, Stability affected punisher pay-off, with punishers in fixed groups earning more (Wald χ^2_1=17.53. p<0.001; β=12.92, s.e.=6.23, p=0.04). Unsurprisingly, Benefit affected punisher pay-offs (Wald χ^2_2=69.50. p<0.001), with punishers in the monopoly condition earning marginally more than those in the no-bonus condition (β=12.34, s.e.=6.90, p=0.07). Rounds also affected punisher earnings (Wald χ^2_4=43.23. p<0.001), with efficiency being greater in the early rounds compared to the later rounds (Rounds 1-2 vs Rounds 9-10, β=16.02, s.e.=5.95, p=0.007).

As also shown in Figure 6.11b, punisher earnings were marginally affected by an interaction between Stability and Benefit (Wald χ^2_2=5.63. p=0.06). Punisher earnings were affected by an interaction between Rounds and Benefit (Wald χ^2_8=18.16. p=0.02), the earnings of punishers in the bonus conditions reduced between Round 1-2 (Bonus, M=39.79, SD=11.56; Monopoly, M=42.08, SD=10.58) and Rounds 9-10 (Bonus, M=28.21, SD=14.99; Monopoly, M=26.93, SD=13.38), while those in the no-bonus condition remained consistent (Round 1-2, M=22.52, SD=7.89; Rounds 9-10, M=17.25, SD=9.66). An interaction between Stability and Rounds did not affect Punisher earnings, nor were Punisher earnings affected by a three-way interaction between benefit, Stability and Rounds.

Table 6.5: group level analysis - model summaries

Group efficiency	Factors	Wald $\chi 2$	df	p
	Benefit	5.53	2	0.06
	Stability	18.47	1	<0.001***
	Rounds	34.35	4	<0.001***
	Benefit*Stability	4.91	2	0.08
	Benefit*Rounds	6.10	8	0.64
	Stability*Rounds	4.60	4	0.33
	Stability*Benefit*Rounds	11.69	8	0.17
Punisher earnings				
	Benefit	69.50	2	<0.001***
	Stability	17.53	1	0.04*
	Rounds	43.23	4	<0.001***
	Benefit*Stability	5.63	2	0.06
	Benefit*Rounds	18.16	8	0.02*
	Stability*Rounds	2.81	4	0.59
	Stability*Benefit*Rounds	8.20	8	0.41
Non-punisher earnings				
	Benefit	7.42	2	0.02*
	Stability	15.86	1	0.04*
	Rounds	7.42	4	0.02*
	Benefit*Stability	2.76	2	0.25
	Benefit*Rounds	4.37	8	0.82
	Stability*Rounds	4.96	4	0.29
	Stability*Benefit*Rounds	12.56	8	0.13

*p<0.05, **p<0.01, ***p<0.001

6.9.3 Mean non-punisher earnings

Mean non-punisher earnings were calculated by dividing the group-pot by 3, once the punisher share (and, for the monopoly condition, the punisher bonus) was removed, as was the value of any punishment. As above while those in a random group could not have known they would receive the additional resources, by comparing the randomly selected non-punishers to those who *knew* they would not be selected, it is possible to see whether at a group level, the study IVs affected earnings. These data were normally distributed and as such a linear model was used in the analysis.

As shown in Figure 6.11c, Stability affected non-punisher earnings, with those in the fixed groups earning more (Wald χ^2_1=15.86. p<0.001; β=5.54, s.e.=2.69, p=0.04). As also shown in Figure 6.11c, Benefit also affected non-punisher earnings (Wald χ^2_2=7.42. p=0.02), however there were no specific differences between conditions. Rounds also affected non-punisher earnings (Wald χ^2_2=7.42. p=0.02), with earnings being higher in the early Rounds (Round 1-2 Vs Round 9-10, β=6.86, s.e.=2.69, p=0.007). Non-punisher earnings were not affected by interactions between Stability and Benefit, Benefit and Rounds, Stability and Rounds, or a three way interaction between Stability, Benefit and Rounds (See Table 6.5 for model summery).

6.10 Discussion

6.10.1 Punishment

Dominant individuals are in a position to benefit to a greater degree from group cooperation (Gavrilets & Fortunato, 2014). This can be either because they can actively monopolise and/or control how others access the results of individual or group activity (Clutton-Brock & Parker, 1995; Cummins, 2005) or because their level of pre-existing resources provides a higher marginal return from group cooperation (Reuben & Riedl, 2013). Because of this

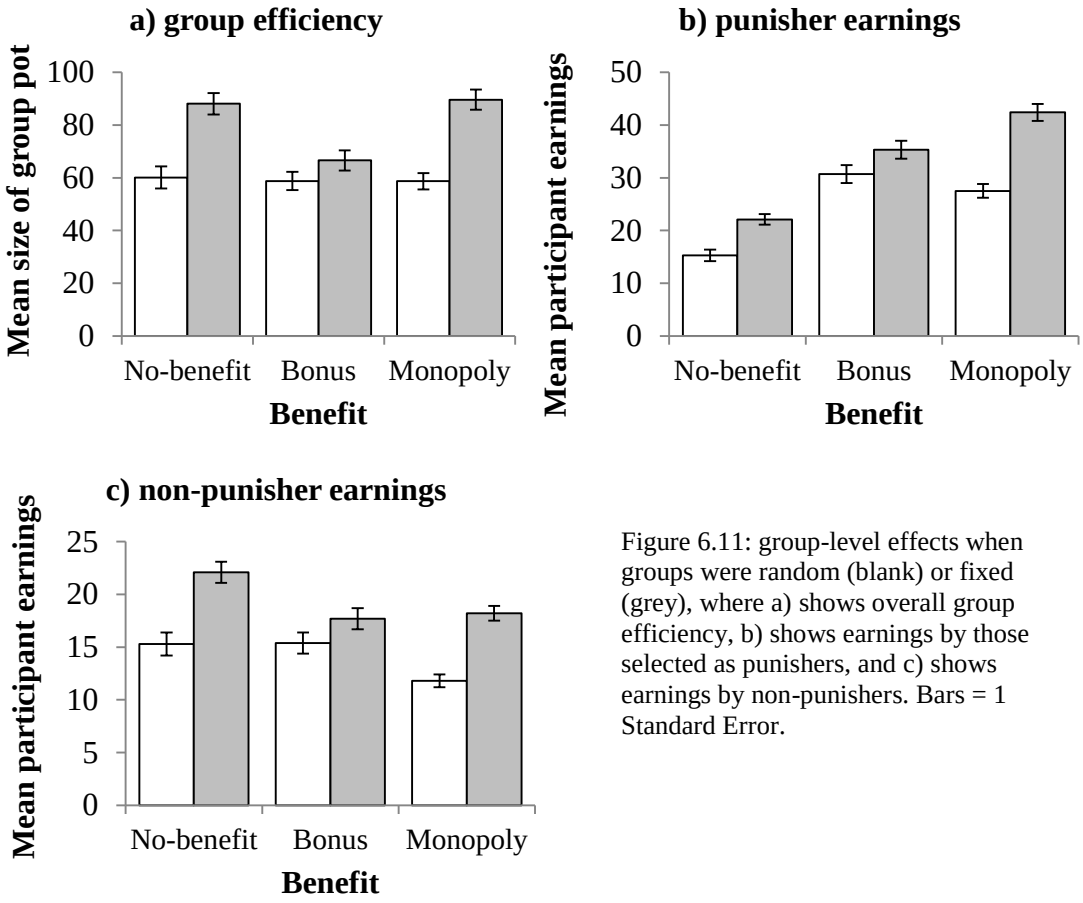


Figure 6.11: group-level effects when groups were random (blank) or fixed (grey), where a) shows overall group efficiency, b) shows earnings by those selected as punishers, and c) shows earnings by non-punishers. Bars = 1 Standard Error.

greater stake in the outcome of any collective activity, dominant individuals may be more motivated to punish non-cooperation. The current study investigated this by providing some participants with one of the attributes of dominance, additional resources, and by manipulating how strategically beneficial punishing non-cooperation would be for them.

Study 8 successfully replicated the results of Study 7; participants who received additional resources punished more severely and with greater frequency than those who did not. Individuals are more likely to punish when the cost is low (McCullough et al., 2013) and by providing punishers with more resources, the studies in this chapter lowered the net cost of punishment for some individuals and, as result, those punishers were more willing to punish free-riding within their group. As with Study 7, this result suggests that one simulated aspect of a dominant position, access to additional resources from group cooperation, at the very

least facilitated a behaviour that all individuals would engage in if they could, due to other-regarding (Camerer & Fehr, 2006) or spiteful (Leibbrandt & López-Pérez, 2011) motives.

However, the results demonstrated a strategic motivation behind punishment behaviour; participants punished more severely and more frequently when they received additional resources *and* when groups were fixed. One of the main proximate purposes of punishment is to change the behaviour of free-riders (Masclet, 2003; Shinada et al., 2004) or otherwise exclude them (Bowles & Gintis, 2004), and when groups were fixed punishers were in a position to continually reap the benefits from any subsequent pro-social behavioural change in the target of punishment. When punishers were guaranteed to repeatedly benefit from group cooperation, they were more willing to punish non-cooperation. This demonstration of strategic motivation complements other research which has shown that individuals will punish non-cooperation when faced with inter-group competition (Abbink et al., 2010), and that dominant individuals specifically will vary their behaviour in response to personal gain (Gavrilets & Fortunato, 2014; Maner & Mead, 2010). Thus, because of this greater stake in the outcome of any group-beneficial activity, dominant individuals may be more motivated to punish behaviour that prevents such activity.

Furthermore, the current results also contradicts the suggestion made regarding the results of Study 7, that punishers only act out of anger and spite due to the heightened loss of resources *for that round* that free-riding represented. This is not to say such motivation could not account for some punishment, as punishers do often show malicious behaviour and/or motivation (Cinyabuguma et al., 2006; Leibbrandt & López-Pérez, 2008; Ostrom et al., 1992; Zizzo & Oswald, 2001). Again, the interaction between Benefit and Stability suggests that, rather than being driven by altruistic or spiteful motivations per se, punishment has a strategic motivation: individuals will engage in costly punishment if it is both cheap and beneficial for them to do so.

A position of dominance therefore provides self-serving motivation for encouraging group-level cooperation. Indeed, a dominant position eliminates the problem of second-order free-riding, as it does not matter whether others benefit from punishment, as long as the punisher does so to a greater degree. The results suggests that a) lowering the relative cost of punishment increases punishment, and b) punishment increases when it can be used strategically. Both these conclusions are consistent with dominance playing a role in punishment behaviour.

6.10.2 Cooperation

Perhaps surprisingly, increases in punishment frequency and severity in response to the available resources did not affect group-level cooperation. In fact, cooperation consistently fell across rounds in the same way as when punishment is not possible (Camerer, 2003), or indeed when an ineffective punishment mechanism is employed (Egas & Riedl, 2008; Nikiforakis & Normann, 2008). Even with the increase in punishment seen in the Benefit conditions, the overall amount spent on punishment, and therefore its impact on the target, was still low (see Table 5.3). As a result participants did not respond as if punishment itself constituted enough of a threat to their own earnings to force them to contribute. That punishers did not increase their absolute spending on punishment proportionally to their bonus could be partly explained by loss aversion and endowment effect (Kahneman, Knetsch, & Thaler, 1990); individuals would rather spend 1 point to punish by three points, than spend three points to punish by three (as shown by Egas & Riedl, 2008)), even if they had three times the available resources of non-benefit punishers – which was not the case here. Even though the cost of punishment was lowered to the point that participants were willing to punish, the cost was still relatively monetarily, and psychologically, high compared to effective punishment.

However, in Study 7, and specifically in the 50% bonus condition, participants who were faced with the prospect of a punisher who received a large amount of additional resources did contribute consistently across rounds. Thus, while a well-resourced punisher can be considered enough of a threat to encourage cooperation, the additional resources available clearly have to be sufficiently high for this to be the case. Some models have suggested that heterogeneity in resources is the most important factor in punishment behaviour (de Weerd & Verbrugge, 2011; Frank, 1996; Kahneman et al., 1990), but the results from Study 8 suggest that punishment still needs to be effective in order to have any proximate effect on cooperation. This suggests that the threat of punishment has to feel credible, as well as be possible. Still, we would argue that dominant individuals can also punish more effectively, insomuch as they are, for example, physically formidable (Sell, Tooby, et al., 2009) or have priority access to information (Maner & Mead, 2010). Furthermore, in Chapter 5, the threats by a dominant individual were seen as credible insomuch as participants felt their intervention would alter the behaviour of a social defector. Thus, while in the present study, cooperation was not maintained by an individual in a 'dominant position' punishing ineffectively, it is unlikely that someone in an actual position of dominance could not punish effectively.

6.10.3 Group efficiency and group-member earnings

While not a direct aim of the study, group efficiency and member earnings provided some interesting data. Perhaps unsurprisingly, results of Study 8 clearly show that punishers did earn more than subordinate group members even though they spent resources on punishment. What is surprising however is that the highest earners were those in the monopoly condition, even though mathematically they earned less per unit of contribution than those in the bonus condition; those in the latter condition should have had a greater incentive to invest in group cooperation (Gavrilets & Fortunato, 2014). This result was likely driven by higher

contributions by the monopolypunishers themselves (see Figure 6.10, and for a more detailed discussion, see 6.10.3). This is surprising. In the stable condition, as with Study 7, punishers who received a bonus would not lose any points regardless of the investment of others, i.e. if they invested 1 point into the group pot, they were guaranteed at least 1 point back regardless of the behaviour of others. However bonus-punishers contributed less than those in the monopoly condition even though the latter received a return of 0.87 on each point they invested. Indeed, given the difference between returns from investment, we may have expected those in the Bonus condition to punish more readily as they received a flat bonus of 0.25 points for every point others in the group invested. However, this was also not the case (see 6.10.1). It seems as if the monopoly-punishers were uniquely motivated to spend resources on the public good (both contributions and punishment), and this resulted in them subsequently gaining greater rewards, but also benefitted other group members in the process.

The Strong Reciprocity theory of punishment is based on group-level selection (Gintis, 2000) that is, self-evidently, dependent on groups out-competing one another, either directly or due to different levels of survival when faced with environmental disasters (Bowles & Gintis, 2004; Boyd et al., 2003). Punishing groups do indeed have a competitive advantage over non-punishing groups (Gächter et al., 2008; Gürerk et al., 2006), and Strong Reciprocity theory argues that, as a result, *altruistic* punishment was selected for even though punishing free-riders is disadvantageous for the individual punisher. What is interesting about the present results (see Figure 6.11a) is that they show a competitive advantage for punishing groups without the need for altruism. Despite the unequal disruption of resources within the group, the monopoly and no-benefit groups had the same level of group efficiency, i.e. we found that a selfishly motivated dominant individual can maintain a group's competitive advantage even

while sequestering resources for themselves. While some have suggested that such a coordination/'leadership' role might be disadvantageous (Gillet et al., 2010), this need not be the case as long as the 'leader' can monopolise the product of any group-level cooperation. Because dominants have both the means to mitigate the immediate costs of punishment and a selfish motive to engage in punishment, no group-level mechanism is needed to explain why they would punish to increase group efficiency (see also, Gavrilets & Fortunato, 2014; Powers & Lehmann, 2014).

6.10.4 Conclusion

As well as having additional overall resources compared to others, dominant individuals also benefit disproportionally from group success. This can be either because they can actively monopolise and/or control how others access the results of individual or group activity (Clutton-Brock & Parker, 1995; Cummins, 2005) or because their level of pre-existing resources provides a higher marginal return from group efficiency (Reuben & Riedl, 2013). Because of this greater stake in the outcome of any collective activity, dominant individuals may be more motivated to punish. In Study 8 it was demonstrated that giving some participants such a greater stake in group-cooperation increased their tendency to punish more severely and frequently. Together, these results suggests that a) lowering the relative cost of punishment increases punishment, b) punishment can be an individually-beneficial strategic act, and c) it is a position of dominance that allows both the former points to occur.

6.11 General discussion

6.11.1 Dominance and costly punishment

Punishment is costly to the punisher, yet as long as there is heterogeneity in the cost of punishment, either through reduced cost relative to resources (Frank, 1996), through the effectiveness of punishment (Roberts, 2013) or in the likelihood of retaliation (Rand et al.,

2010), then punishment can be stable at the individual level. Equally, if there is heterogeneity in the benefit individuals derive from group cooperation, this will overcome the remaining second-order free-rider problem (Dreber et al., 2008; Yamagishi, 1988), as it does not matter whether others benefit from punishment, as long as the punisher does so to a greater degree. We believe an individual's position in a dominance hierarchy can furnish the various heterogeneities in cost and benefit described above. The current chapter addressed the role of greater resources in punishment decision making.

By definition, a dominant position reflects that an individual has *"...priority of access to resources"* (Cummins, 1996a, p. 467) or preferential access to *"any requisite that adds to the genetic fitness of the dominant individual"* (Wilson, 1980, p. 129). Fundamentally, whether because dominant individuals can monopolise resources as they can behave coercively (Clutton-Brock & Parker, 1995; Hawley, 1999) or because others are willing to tolerate asymmetries in reciprocity to maintain a close relationship with them (Barclay, 2013; Henrich & Gil-White, 2001; Schino & Aureli, 2009), dominants have a high resource-controlling ability (Hawley, 1999). The resource-controlling ability of dominant individuals also means they gain an increased marginal return on any group activity should invest in the public good more than others (Gavrilets & Fortunato, 2014), including investing more in the punishment of non-cooperation (which can be considered a public good, Nikiforakis & Normann, 2008).

Both studies in this chapter found that when a dominant position was simulated by providing some individuals with additional resources, they punished non-cooperation more frequently and more severely. While the motives of punishers in Study 7 might be up for debate (see 5.5.1), the results of Study 8 clearly show punishment to be a strategic behaviour; participants punished more severely and more frequently when groups (and their position) were stable,

i.e. when it was both cheap and beneficial for them to do so. Thus, a dominant position not only provides the means to punish, it also provides a motivation. Importantly, this additional benefit provides a means to overcome the second-order free-rider problem, which is the crux of the debate around the evolution of punishment: it does not matter, to a degree at least, what the cost to the punisher is as long as they benefit disproportionally from their actions compared to non-punishers.

6.11.2 A single (ineffective) punisher

As has been discussed throughout this thesis, a dominant position provides an individual with multiple paths by which the cost of punishment can be reduced. Effectiveness of punishment is a key mechanism that reduces the cost of punishment (Balliet et al., 2011; Egas & Riedl, 2008), and a dominant position allows individuals to punish effectively insomuch as they are, for example, physically formidable (Sell, Tooby, et al., 2009) or have priority access to information (Maner & Mead, 2010). Chapter 5 demonstrated that participants believed the threats from dominant individuals to be credible and successful.

This is why the current studies kept punishment ineffective, and this is likely why, unlike other studies that investigated punishment when there was heterogeneity in marginal benefit (Reuben & Riedl, 2009, 2013; Tan, 2008), heterogeneities in resources did lead to more punishment in Studies 7 & 8. Effective punishment lowers the cost of punishment and as a result even those who have relatively low resources can punish cheaply. Effective punishment usually produces a great deal of punishment, altruistic or otherwise (Cinyabuguma et al., 2006; Nikiforakis & Normann, 2008) and according to some, punishment far in excess of that which occurs in everyday life (Guala, 2012). It is no surprising therefore that when this ‘dominant’ ability was removed in the current studies that the effect of heterogeneity in resources on punishment was revealed.

More importantly, the fact that individuals are less inclined to retaliate against dominants for behaving 'unfairly' in dyadic interactions (see also, Eckel et al., 2010; Kim et al., 1998) might also lower the cost to a dominant individual from retaliation/counter-punishment. Such counter punishment is one of the main costs to punishment (Dreber & Rand, 2012; Nikiforakis, 2008) yet a target who receives punishment from a dominant individual would be expected to be as disinclined to respond antagonistically against dominants as they would be in other dyadic interactions. To implicitly simulate this, a single-punisher mechanism was employed (Baldassarri & Grossman, 2011; O'Gorman et al., 2009), as this would prevent the punisher from being altruistically, spitefully or counter- punished (see, Ostrom et al., 1992; Ottone, 2008); i.e. the punisher had the freedom of action open to dominant individuals (Van Vugt, 2006), including whether or not to contribute (see 5.10.3).

Additionally, previous studies have found that a single punisher (with effective punishment) can sustain cooperation (Baldassarri & Grossman, 2011; O'Gorman et al., 2009) and there have been questions raised as to whether an informal peer-sanctioning mechanism represents actual group dynamics. It has been suggested that groups of individuals tend to self-organise towards some individuals having discretionary/leadership roles (see, Baldassarri & Grossman, 2011), and even in egalitarian societies dominant individuals do have a louder voice in group decisions (Henrich & Gil-White, 2001). Traulsen et al. (2012) found that individuals prefer an environment of pool punishment, where a single authority dispenses punishment, even at the expense of group efficiency. Therefore the current chapters mechanism, whereby resources were concentrated in one dominant/despotic individual (as opposed to the automated central authority of Traulsen et al., 2012), could be seen as a stepping stone between the peer-punishment of non-state societies and the formalised policing of state societies.

Still, why non-dominant group members tolerate such sequestering of both resources and power is another matter. It has been suggested that inequality in resource distribution might be a fair price for free-riding on norm enforcement (see, Roberts, 2013), especially if there are limited outside options (Gavrilets, 2012; McNamara & Houston, 2002). In fact a recent model of the shift from an egalitarian social structure to a more familiar hierarchical one in humans depends on the latter especially (Powers & Lehmann, 2014). It could be suggested, if speculatively, that such an acquiescence to the punishing power to dominant individuals could be compared to other forms of asymmetry between dominants and subordinates that the latter just have to tolerate (Barclay, 2013; Schino & Aureli, 2009).

6.11.3 Beyond dominance: a case of Noblesse Oblige?

We have suggested that dominance might play an important role in the evolution of costly punishment; from a certain perspective dominance does relate to various factors that have been shown to encourage costly punishment in the literature, and specifically in Study 7 & 8 in how resources are gained from group cooperation encourage punishment. Study 8 found that punishers in stable groups *who benefitted directly at the expense of the group* (i.e. the monopoly condition) behaved in the most pro-social way. These punishers contributed to the public good by both punishing free-riders more frequently and severely than those in other groups, and also contributed more than other group members. This is interesting as while (spiteful) punishment of higher earners (Burns & Visser, 2006; Van De Ven et al., 2010; Zizzo & Oswald, 2001) might drive some cooperation, here there was no opportunity for this to happen here because only the high earner could punish.

An explanation for this is that a dominant position can be precarious, especially in human groups. While a strong individual could demand a greater share of any resource (Sell, Tooby, et al., 2009; Zak et al., 2009), humans respond negatively to being on the disadvantaged side

of resource distribution (Leibbrandt & López-Pérez, 2008, 2011; Van't Wout et al., 2006; Zizzo & Oswald, 2001) and the prevention of such exploitation likely played a strong role in the evolution of our social/coalitional psychology (Cummins, 1996a; Gavrilets et al., 2008). Dominant behaviour therefore might lead to either a withdrawal of cooperation by individuals, the dissolution of a group, or rebellion by subordinates (Brandt, Hauert, & Sigmund, 2006; Hirschman, 1970; Van Vugt et al., 2004)

In fact the threat from revolutionary coalitions is main reason for the relatively flat dominance hierarchy of pre-state tribes (Boehm, 1997). To circumvent the threat of a coalition forming, dominant individuals might therefore have to behave in a pro-social and generous way (Fiddick & Cummins, 2007). This is supported by the results of Chapters 5 & 6, which demonstrated that that punishment is not only an activity that dominant individuals can do, but also one they are expected to take part in. By placing punishers in a position where their success was at the expense of others (the monopoly condition), the study activated the psychology designed to ward off the negative reactions of subordinates, even though no reaction was actually possible. While the data supporting this suggestion is exploratory (see 6.5), punishers who *knew* they were benefiting at the expense of other group members behaved in the most pro-social way; they both contributed and punished more than other group members or punishers in different conditions. This does suggest on some level they were acting with a sense of *noblesse oblige.*

It should be remembered that individuals do actively prefer an environment where punishment is possible (Gürerk et al., 2006) and that having a single individual 'in charge' can lead to more efficient use of punishment to promote cooperation (Baldassarri & Grossman, 2011; O'Gorman et al., 2009), and can improve general decision-making (Gillet et al., 2010; King et al., 2009). Because such a dominant position can be exploited (Maner &

Mead, 2010), it is likely that a lot of the 'moral' emotions surrounding fairness evolved to ensure we and our allies were not exploited by others and to keep track of any potential social challengers or threats (Boehm, 1997; Brosnan, 2011; Byrne & Whiten, 1997; Cummins, 1996a; Jenson & Peterson, 2011; Peterson, 2012). Thus, the inter-dependent group structure of our evolutionary past (Boehm, 1997; Charlton, 1997), rather than producing a psychology of group-level egalitarian motives (Fehr & Schmidt, 1999), has instead more finely honed the ability to form coalitions for our own gains and to watch out for competitors (Cummins, 1996a, 2005). This would certainly explain the emergence of despotism once the 'Mexican stand-off' of our pre-state /pre-agricultural existence was broken (Powers & Lehmann, 2014; Turchin et al., 2013).

Taken together, the above arguments suggest that engaging in costly punishment could be seen as the *price* of a dominant position, i.e. anyone who is determined to extract more than an equal share of resources from individuals or the group as a whole is tolerated as long as some of these resources are used for the public good. In fact, it has been suggested that subordinates will tolerate inequality for the opportunity to free-ride on public good spending and norm enforcement (Gavrilets & Fortunato, 2014; Roberts, 2013), and it is probably no coincidence that the reputation gained from an act of punishment (trustworthy, group-focused; Barclay, 2006) is similar to the set of traits demanded in a leader (Hogg et al., 2012). Costly punishment can be seen as a strategic tool signal and ensure a dominant position without the negative consequences of being seen as too despotic and therefore risking the dissolution of any group or coalition. There are clear advantages to being in environments where punishment is possible (Fischbacher et al., 2001; Gürerk et al., 2006), and, to state the argument from a Hobbesian perspective, individuals are willing to sacrifice freedom of action for benevolent protection or, in a more contemporary fashion, no one ever voted for 'soft on crime'.

6.11.4 General conclusion

Chapter 4 found that those who engage in punishment are thought of as likeable and dominant, and Chapter 5 found that dominant individuals are expected to punish and that the aforementioned reputational benefits are dependent on this punishment being successful. Nevertheless, those results did not demonstrate that dominants would *actually* punish. The studies in the present chapter aimed to investigate whether dominant individuals, or rather, participants experiencing one advantage of a dominant position, would actually punish. The studies simulated an advantage of a dominant position by a) providing punishers with additional resources (Study 7) and, b) by varying the strategic motive for punishing (Study 8). The studies found that gaining additional resources from group cooperation motivated individuals to engage in more severe and more frequent punishment, and that this behaviour further increased when there was a strategic benefit for doing so.

These results support the suggestion that a dominant position leads to greater punishment and represents the individual heterogeneity in the cost of punishment that might allow it to be evolutionarily stable. Furthermore the current study demonstrated that a dominant position provides a selfish incentive to maintain group cooperation that is not available to subordinates. Finally, Study 8 specifically, suggests that dominant costly punishers are not simply motivated by the direct benefits from group cooperation they receive alone but, potentially at least, are acting in an altruistic way in order to maintain their position in a social hierarchy, i.e., investing in the public good is the price one has to pay for a dominant position.

7 Chapter 7: dominance and behaviour 2 - naturally occurring dominance

Previous chapters have demonstrated that dominance is potentially an important factor in costly punishment. Chapter 7 therefore does not manipulate dominance experimentally, but investigates whether those in a 'real life' dominant social position will engage in a greater amount of punishment.

7.1 General introduction

The previous chapters have demonstrated that there are individual-level reputational and material advantages to engaging in costly punishment, and that these advantages are inextricably linked to a dominant position. Chapter 4 suggested that costly punishment makes one appear dominant and likeable in the eyes of observers, Chapter 5 demonstrated that dominant punishers are expected to punish and that their position can lower the cost of punishment, and Chapter 6 demonstrated that those in a position to benefit disproportionally from group success will punish to enforce group cooperation. However, the studies in Chapter 6 did not manipulate dominance as such, but rather experimentally manipulated the payoffs of an economic game in order to simulate an aspect of dominance, access to higher resources. The question still remains as to whether those actually in a dominant social position will engage in a greater amount of costly punishment. Therefore, unlike previous chapters that have tested a specific characteristic of a dominant position, the final empirical chapter will address the question of whether a real-world position of dominance affects punishment behaviour.

7.1.1 Cooperation and dominance

Social connections have an important impact on general cooperative behaviour, with individuals behaving far more altruistically towards friends than non-friends (Brañas-Garza et

al., 2010; Jones & Rachlin, 2006). Participants will, for example, endure discomfort for close reciprocal friends (Harrison et al., 2011) and will work hard to maintain relationships with them (Roberts & Dunbar, 2010). While we are clearly generous to those closest to us, status in a group may decrease an individual's level of cooperation and altruism when dealing with other members; specifically dominant individuals often behave less ethically and less fairly (Lammers et al., 2010; Piff et al., 2012).

Dominant individuals might behave less fairly as they do not fear the punishment associated with unfairness. High status individuals have access to greater resources (Ellis, 1994) and a greater number of social allies (Von Rueden et al., 2008), and this may help guard them from any antagonism their behaviour invites. For example being central to a group is a strong negative predictor of victimisation and exploitation (Figueredo et al., 2001). While selfish behaviour itself is cheap, if others are willing to respond aggressively to that behaviour the retaliatory cost may be very high. Yet when faced with an uncooperative partner, participants are less likely to punish those perceived as being of high status (Eckel et al., 2010; Kim et al., 1998). Also, dominant individuals are far more willing than others to respond aggressively to any action against them (Griskevicius et al., 2009; Silk, 2003). Dominant individuals are therefore more free to behave in either a cooperative or coercive manner towards other members of the group as needed, and can use their dominant position to manipulate group interactions to ensure a favourable outcome for themselves (Dasgupta, 2011; Maner & Mead, 2010).

Finally, dominant individuals may be too important, due to their connections or resources, to alienate. Ostracism has been suggested as a mechanism to deter non-cooperation without resorting to costly punishment (Bowles & Gintis, 2004; Masclet, 2003; Rand, Ohtsuki, et al., 2009), but such studies have assumed there is no cost associated with removing someone from a social circle. It may be in an individual's best interest to stay in proximity to a

dominant individual regardless of their behaviour. This may be especially true if uncooperative dominant individuals have attained their position due to a useful skill set (Henrich & Gil-White, 2001; Petersen et al., 2012), but it would also apply if proximity to a physically dominant individual ensures safety (for example, Snyder et al., 2011). Thus lower status individuals may have to simply endure an unbalanced relationship with more dominant peers, as is the case in non-human primate societies (Schino & Aureli, 2009; Watts, 2002).

7.1.2 Punishment and social status

The desire to inflict retribution on those who behave unfairly seems to be an automatic response (Crockett et al., 2010). Yet there is great variation in the willingness to actually punish others for their behaviour, and dominance may play a role in this decision making. As mentioned, social position gives high status or dominant individuals an advantage in dyadic interactions and dominant individuals are very sensitive to any unfairness directed towards them (Brosnan, 2011; Burnham, 2007) as this may indicate a challenge to their position. Equally, while it is advantageous for all members of a social group to track the dominance relationships between conspecifics (Cummins, 1996a), it might be advantageous for dominant individuals to be especially sensitive to these interactions as, again, they could represent the rise of a social challenger. This may explain why socially dominant individuals seem more willing to respond to perceived unfairness or the violation of social norms (Cummins, 1999; Lammers et al., 2010).

Yet, because of their ability to monopolise resources, dominant individuals do benefit disproportionally from group success. They should therefore be motivated to encourage group cooperation. This may be due to capital return, because their position means, for example, they have more land or property and therefore any group benefit from an action is multiplied by this fact (Reuben & Riedl, 2013), or, in a less formal environment, because their central position in a social group allows dominants to take advantage of the flow of

resources or information that closer ties brings (Harrison et al., 2011; Jones & Rachlin, 2006). A series of recent models (Gavrilets & Fortunato, 2014; Roberts, 2013) have suggested that, because of the additional benefits from group efficiency that dominance provides, dominant individuals should always invest in both cooperation and punishment. This prediction was empirically tested in Study 8 and supported by that study's results.

Finally, as well as possessing additional motivation to punish, as has been discussed at length in previous chapters a dominant individual experiences a lowers the cost of punishment. Both the additional resources available to dominant individuals and the ability to punish effectively lower the cost of punishment, which in turn makes in more likely to occur. Furthermore, as with the reduced risk of a response to unfair behaviour, a dominant position may lower the threat to an individual from retaliation, because their social allies, physical prowess, or general willingness to respond to a confrontation would make retaliation from the target of punishment less likely.

7.2 Study 9: cooperation and punishment in an informal social network

Thus a dominant position would both allow and motivate an individual to engage in costly punishment. While the arguments above are empirically based, few studies have explicitly examined the role of actual status within a group; the effects of a dominant position have been simulated through specific experimental manipulations (e.g. Chapter 6). The current study therefore investigated directly what effect an individual's status within a group had on cooperative and punishment behaviour. Rather than experimentally manipulating the advantages of being dominant by, for example, by varying resource levels (Reuben & Riedl, 2013) or the effectiveness of punishment (Nikiforakis et al., 2009), the current chapter measured actual status by investigating social relationships in a closed social network.

Social network analysis (see Krause et al., 2007, and; Scott, 2007 for review) can offer one means of measuring status by identifying how central an individual is in a group. A simple measure of centrality is *degree centrality,* which indicates the amount of connections an individual receives (InDegree) and sends out (OutDegree). Of particular interest is InDegree as this indicates prestige/dominance (Wasserman, 1994): high InDegree suggests that an individual is being watched and approached by others; behaviour strongly connected to high status and dominance (Hawley, 1999; Rege, 2008). Because of this the current study focused on InDegree alone.

Being in a dominant position affects how environmental and interpersonal events are perceived. For example, dominance changes belief about the distribution of resources (Cummins, 2005): specifically, dominant individuals feel entitled to a greater share of resource and are more willing to endorse the use of force both interpersonally and internationally in order to achieve this (Sell, Tooby, et al., 2009). The dominance of a competitor influences both our own responses to their behaviour (Jenson & Peterson, 2011) and the way we behave towards them (Gambacorta & Ketelaar, 2013). Yet, interestingly, formidable individuals (Watkins et al., 2010) those who are primed to feel dominant (Watkins & Jones, 2012) no longer respond to cues of formidability in potential competitors, and those who are surrounded by allies perceive potential foes to be less threatening (Fessler & Holbrook, 2013). Finally, it is important to note that we are very aware of our status within a social group (Anderson, Srivastava, et al., 2006; Cummins, 1996a), with social penalties for being perceived to have stepped beyond our status (Anderson et al., 2008).

These effects of a dominant position may be due to environmental feedback, i.e. learning what one can get away with, may reflect proximate social norms on how dominants should behave (for example, Fiddick & Cummins, 2007), or they may result from instinctive condition-dependent evolved behaviours. For whatever reason, a position of dominance

fundamentally changes the cost/benefit analyses of our behaviour. Behaviour in economic games is expected to reflect individual differences (for example Gunnthorsdottir et al., 2002) and differences in the social environment from which participants are drawn (For example Henrich et al., 2010), and also to reflect real world behaviour (Rustagi, Engel, & Kosfeld, 2010). Therefore, even without any direct motivations (for example, reputational gain, see Chapter 4; or benefit from group success, Chapter 6), individuals who occupy a dominant/prestigious position within a group should be expected to act within an economic game in accordance with that position.

In the present experiment, therefore, individuals who were members of a closed social network took part in a third-party punishment game. It was predicted that an individual's behaviour would be related to their social position, with dominant individuals behaving more unfairly and being more likely to engage in costly (in this instance 'third party') punishment than subordinate individuals. It was also predicted that the effectiveness of punishment would interact with status in participants' decision making; dominant individuals would behave less pro-socially under high punishment effectiveness, and would punish more when effectiveness was low.

7.3 Method

7.3.1 Participants and research context

Participants were 2nd year undergraduate students from the University of Exeter who were attending a week-long field course at a Field Study Council site in South Wales, UK. In total, 29 students (23 females, 6 males; mean age=20) attended the course and completed the social network questionnaire. Twenty-one of the students (17 females, 4 males; mean age=20) took part in a Third Party Punishment Game (TPPG). Although the sample size was small, the situation represented a rare opportunity to study a closed-network of individuals where there

was no existing formal or semi-formal hierarchy, as there might be, for example, in a sports team due to ability or playing position. Data collection took place in April 2011.

7.3.2 Design

In a standard TPPG, participants play in one of three roles, that of the Proposer, Receiver or Third Party. Participants in the Proposer role are given an allocation of 20 points and can chose to send between 0-20 of these points to the participant in the Receiver role. The participant in the Third Party role is given 10 points and observes this interaction; they then decide whether to spend their own points to punish the Proposer. The receiver plays no active part in the game and earns only the points sent to them. The Proposer and Third Party keep any points they don't send to the Receiver or spend on punishment.

Due to the small sample size, a pen-and-paper strategy method version of the TPPG was used (Fehr, 2004) with participants being asked to make proposal and punishment decisions simultaneously. Also, when making proposal decisions, participants could only transfer between 0 to 10 points to the Receiver, rather than the full 20. This limit was implemented as individuals rarely send more than half their points during such decisions (Fehr, 2004) and to limit the number of strategic decisions participants had to make in the Third Party role.

7.3.3 Procedure

The study took place on the 6th day of the field course. Participants were informed that participation was entirely voluntary and that they had the potential to earn up to £30 by taking part. Once all participants were seated in the main teaching room, the instruction sheet was handed out. This sheet explained the various roles and how payment would be generated (see 7.3.4 Matching & Payment Procedure). Instructions were read out verbatim, and participants were instructed to raise their hand if they were unclear about any details. None did.

Participants were asked to make proposal and punishment decisions for two levels of punishment effectiveness: An "effective" condition whereby one point assigned by a Third Party removed three of a Proposer's points (1:3 ratio), and an "ineffective" condition where one point assigned by a Third Party removed one point of a Proposer's points (1:1 ratio). The Receiver had no active role in the game but participants were made aware that part of their payoff for the experiment would depend on what others had sent to them. Thus, the number of points (and therefore, money) participants earned depended on a) their proposal decision, b) points they received from a Proposer, c) their behaviour AS the Third Party, and d) the punishment received FROM a Third Party.

Participants were then randomly assigned into two groups, with one half being led to another teaching room nearby. Following this, no further explanations were given by the researchers. One of the groups was given the Proposer decision card first, and the other the Third-Party decision card first. The order of presentation had no effect on Prosper and Third Party decisions, and therefore was not included in any analyses.

The Proposer card asked participants to indicate how many points they wished to send to the Receiver "If each deduction point assigned to me by a Third Party removes one of my points" or "... three of my points". The Third Party card asked participants to indicate "The amount of points I would assign to the Proposer" for each effectiveness, with boxes available for each possible offer the Proposer could make (0-10). Once all participants had completed their first decision card, it was collected and the other was handed out. The experimental session lasted approximately 25 minutes.

7.3.4 Matching & Payment Procedure

After the experiment, participants were randomly matched into triads (containing a Proposer, a Third Party, and a Receiver). Twenty-one triads were generated as each participant was

assigned once in each role and acted in each role in a different triad. Furthermore, each triad was randomly assigned a punishment condition, i.e. whether that triad would use the strategic decisions participants had made for effective (1:3) or ineffective (1:1) punishment. Based on the decisions made in each triad, a point total was generated for each participant. Half of the participants were randomly selected to receive payment and their points were converted at a ratio of 5:1 (5 points = £1 pound sterling). Two participants were randomly selected to receive payment on a 1:1 ratio of points to £. Payment was made privately to those selected the following day and those selected to receive their payment at a ratio of 5:1 earned a mean amount of £4.40. Participants selected to receive payment at a ratio of 1:1 earned a mean amount of £18.00.

7.3.5 Generating the social network

On the 7thday of the field course, participants were asked to complete the social network questionnaire. After being informed that participation was entirely voluntary they were handed the booklet containing the questionnaire and a list of those attending the course. As with the TPPG, instructions on the sheet were also read out verbatim, and participants were instructed to raise their hand if they were unclear about any details. The questionnaire asked participants to indicate who they believed to be the most influential individuals on the field course, who they socialised with on the field course, who (of those attending the course) they socialised with at home (i.e., in Exeter), and who they thought of as close friends. There was no limit on the number of individuals participants could identify for each question. These questions were chosen as an attempt to capture the different relationships between participants, for instance a few individuals may be recognised as the most influential in a group, but friendship groups may be more dispersed. The individual network questions, "Close Friends", "Most Influence", "Socialise at Home", "Socialise on Trip", correlated on a scale with high reliability (α=0.851), therefore the additional measure "Total Network" was

created by summing the ties indicated by participants in the network questions. This last measure was a 'weighted/valued network' (Scott, 2007, p. 65) and represented the overall strength of associations between group members. For the full questionnaire, see Appendix E.

7.3.6 The Trait Dominance-Submissiveness Scale (TDS)

At the end of the social network questionnaire, participants were presented with the Trait Dominance-Submissiveness Scale (Mehrabian, 1994). The TDS is a 26-item scale designed to measure trait dominance independent of arousal or extraversion and contains questions such as "When I am with someone else, I usually make the decisions". The version used here asked participants to indicate their agreement with such statements on a 9-point scale (1= Very strong disagreement, 9= Very strong agreement) with a '1' response subsequently being scored as -4 and a '9' response scored +4. This measure was used as an alternative way of measuring dominance. The alpha reliability for this measure was 0.92.

7.3.7 Statistical analysis

The software package UCINET (Borgatti, Everett, & Freeman, 2002) was used to generate the network variables. As social network data violate assumptions of independence (Lusher, Robins, & Kremer, 2010), it was not possible to use standard software such as SPSS for the analysis. Therefore UCINET was also used to analyse the effects of the network variables on behaviour, as the program takes into account the non-independence of the network data when conducting statistical tests. Due to the limited range of analytical tests available within the package, we were unable to perform a direct analysis of the relationship between the social network data and punishment effectiveness on participant's behaviour. Therefore, in order to conduct the mathematical equivalent of an analysis of covariance, additional calculations were performed on the data prior to correlational tests being carried out within UCINET.

7.4 Results

Table 7.1 shows the relationship between the TDS and Network matrices measurements of dominance. Dominant personality scores correlated with the social network metrics, suggesting that the use of InDegree can be seen as an accurate measure of dominance.

Table 7.1: the relationship between trait dominance and social network position

	r (n=29)	p
Friends	0.46	0.019*
Influence	0.5	0.01**
Socialise on Trip	0.32	0.08¥
Socialise at Home	0.34	0.07¥
Total Network	0.48	0.014*
(*p<0.05, **p<0.01, ¥p<0.09)		

7.4.1 Cooperation and social status

The overall mean Proposer offer was 4.8 points (SD=3.2). The mean proposer decision under effective punishment was 5.7 points (SD=3.4) and under ineffective punishment was 3.8 points (SD=2.8). When making proposer decisions under effective punishment, two individuals made offers of zero, and under ineffective punishment four participants (including the previous two) sent zero offers to the Receiver.

In order to investigate whether there was a relationship between overall Proposer behaviour and the social network variables, the sum of participant offers across both efficiencies was calculated. The correlation of this sum with the network-position variables was used to indicate how much variance in overall Proposer behaviour could be explained by the covariates (network position). The

Table 7.2: the relationship between social network position and Proposer offers under different punishment effectiveness.

	Aggregate Proposer Offers		Difference between Proposer Offers under effective and ineffective punishment	
	r (n=21)	p	r (n=21)	p
Friends	-0.40	0.074¥	0.20	0.394
Influence	-0.54	0.011*	0.19	0.392
Socialise on Trip	-0.53	0.014*	0.49	0.025*
Socialise at Home	-0.42	0.059¥	0.45	0.041*
Total Network	-0.58	0.006**	0.37	0.102
(*p<0.05, **p<0.01, ¥p<0.09)				

correlations are shown in the left two columns of Table 7.2. They show that the network variables 'Influence', 'Socialise on Trip' and the 'Total Network' score accounted for a significant amount of variance in Proposer behaviour; those with a stronger network position made lower overall offers than lower ranking members of the network. The variables 'Close Friends' and 'Socialise at Home' also showed similar trends that approached significance.

To investigate whether social network position differentially affected Proposer offers under different levels of punishment effectiveness, the difference between participants' offers under ineffective and effective punishment was calculated by subtracting the former from the latter. Again these data were correlated with the network variables. The network variables 'Socialise on Trip' and 'Socialise at Home' explained a significant amount of the variance in the interaction between Proposer offers and punishment effectiveness (see figures 7.1 & 7.2). This interaction is further highlighted by the correlation data (see Table 7.3): individuals with low network status gave significantly higher offers than high prestige individuals only when punishment was effective, i.e., the risk from punishment was high.

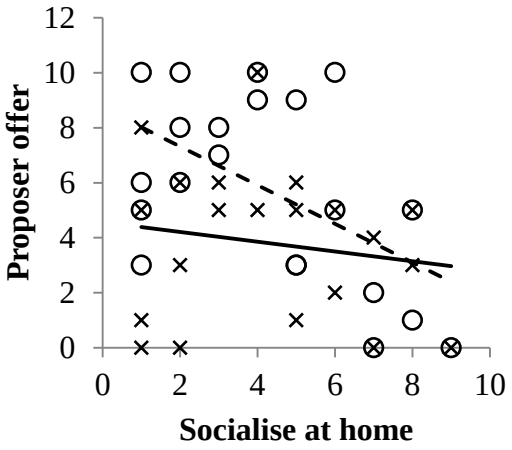


Figure 7.2: relationship between 'Socialise at home' InDegree network position and Proposer offers when punishment was effective (Circles, dashed line) or ineffective (Crosses, solid line).

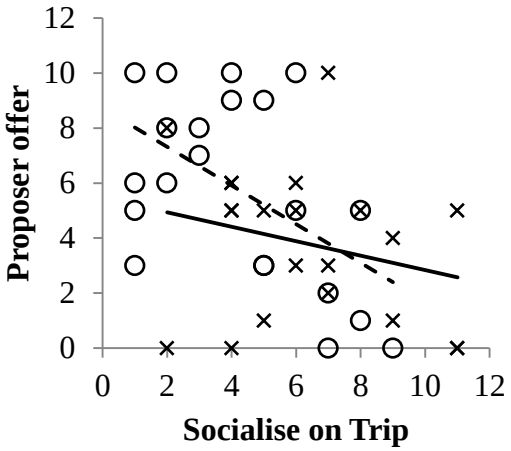


Figure 7.1: relationship between 'Socialise on Trip' InDegree network position and Proposer offers when punishment was effective (Circles, dashed line) or ineffective (Crosses, solid line).

7.4.2 Punishment and social status

The mean number of points allocated across the possible punishment decisions was 2.6 (SD=1.7). When punishment was effective, the mean number of points was 2.2 (SD=1.4) and when ineffective, 3.0 (SD=2.0). All participants indicated they would punish at least one possible proposer-offer when punishment was effective, but two participants did not punish a single possible proposer-offer when punishment was ineffective.

To investigate whether a participant's social network position explained any of the variance in punishment behaviour in response to the potential offers made by Proposers, the gradients of the regression lines relating to punishment responses to each potential offer were calculated for both ineffective and effective punishment. The correlations between these gradients and the social network variables were then computed, and are displayed in Table 7.4. As shown in Figure 7.3, the variable 'Influence' significantly accounted for the variance in punishment behaviour in response to

Table 7.3: correlation between social position and Proposer offer under effective or ineffective punishment

	Offer under effective punishment		Offer under ineffective punishment	
	r (n=21)	p	r (n=21)	p
Friends	-0.41	0.071¥	-0.22	0.215
Influence	-0.52	0.013*	-0.42	0.061¥
Socialise on Trip	-0.65	0.002**	-0.26	0.266
Socialise at Home	-0.53	0.016*	-0.16	0.475
Total Network	-0.63	0.003**	-0.37	0.103
(*p<0.05, **p<0.01, ¥p<0.09)				

potential Proposer offers. Individuals who were thought of as influential were significantly more sensitive to Proposer offers, i.e., they tended to assign more punishment points for low/unfair offers and fewer points to fairer offers.

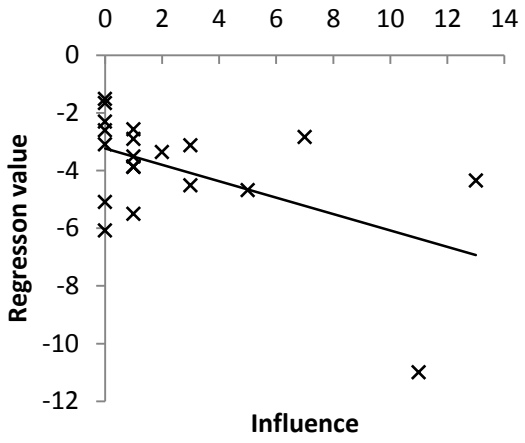


Figure 7.3: relationship between regression slope for punishment spending in response to Proposer offers and influence InDegree network position.

Table 7.4: the relationship between network position and third party punishment behaviour under effective and ineffective punishment conditions

	Gradient of punishment responses to potential Proposer Offers		Influence of punishment effectiveness on aggregate punishment behaviour		Difference between gradient of punishment responses to punishment effectiveness	
	r (n=21)	p	r (n=21)	p	r (n=21)	p
Friends	-0.22	0.35	-0.29	0.194	0.1	0.672
Influence	-0.52	0.04*	-0.38	0.097	-0.01	0.955
Socialise on Trip	-0.27	0.244	-0.06	0.801	0.11	0.65
Socialise at Home	-0.31	0.178	-0.21	0.352	0.23	0.343
Total Network	-0.38	0.088¥	-0.23	0.301	0.1	0.657

(*p<0.05, **p<0.01, ¥p<0.09)

We investigated whether the difference in overall punishment behaviour between the punishment effectiveness levels was affected by social network position. The mean of participant spending on punishment across potential Proposer offer levels was calculated for ineffective and effective punishment and then the former subtracted from the latter. These differences were then correlated with the individual network variables. As the central two columns in Table 7.4 show, social network position did not significantly affect the difference in average punishment behaviour between the two punishment effectiveness levels.

Finally, to investigate whether social network position interacted with effectiveness of punishment in determining the severity of punishment to potential Proposer offers, the difference between the punishment response gradients was calculated for each participant by subtracting the ineffective gradient from the effective gradient. The correlations between the gradient differences and the social network variables were then computed. As shown in the

right hand columns of Table 7.4, there were no significant correlations, suggesting that social network position did not interact with punishment effectiveness.

7.5 Discussion

7.5.1 Cooperation and social status

One of the fundamental assumptions made about economic games is that behaviour within them reflects how individuals behave outside of the laboratory (Henrich et al., 2010; Levitt & List, 2007; Rustagi et al., 2010) and individuals do have a strong sense of their position within a group (Anderson, Srivastava, et al., 2006) which would include the 'appropriate' way to behave (Anderson et al., 2008). The current study therefore used an individual's position in an informal social network, as defined by their InDegree score - the number of others who identified them in response to various social measures. InDegree is metric that represents prestige/dominance within the group (Scott, 2007), and it was used to investigate whether group position would have an effect on cooperative and punishment behaviour. A dominant position did affect behaviour within the Third Party Punishment Game, and by far the strongest relationships were found between the proposer behaviour and social status. Across the measures of social position, dominant/prestigious individuals made significantly smaller offers overall, and for a number of measures there was a significant interaction between social position and effectiveness of punishment.

One interpretation of these results is that low status individuals may behave pro-socially to raise their status within the group. By definition, individuals of low network status lack connections to other members, and sacrificing resources may be an attempt to instigate such reciprocal relationships. This is consistent with behaviour seen in non-human primates where low ranking members are willing to sacrifice resources in order to associate with higher ranking conspecifics (Stevens, Vervaecke, de Vries, & Van Elsacker, 2005; Watts, 2002).

Humans actively compete to improve their reputation within a group (Barclay & Willer, 2007; Roberts, 1998) and low ranking individuals may especially benefit from doing so. In fact, such 'mundane' cooperative/altruistic behaviour might be the preferred option for individuals without the status or the ability to perform more 'heroic' acts (Barclay, 2013).

Conversely, rather than low status individuals behaving more pro-socially, the results could be explained by high status individuals being more selfish. Being in a dominant position changes the way individuals interact with the environment and conspecifics (Maestripieri, 2012): dominant individuals see themselves as deserving a greater share of any resources (Sell, Tooby, et al., 2009), make less generous offers in economic games (Zak et al., 2009); and in everyday life they behave in a more selfish way than lower status individuals (Piff et al., 2012). Thus, the results of the current study can be seen as confirming this self-serving bias in the behaviour of dominant individuals, as prominent individuals kept more of their allocation for themselves.

Additionally, the results could explain what property of a dominant position affected how participants reacted to the specific game mechanisms. The established findings that there are differences in altruistic and/or cooperative behaviour depending on the presence of a third party (Henrich et al., 2010; Kim et al., 1998) and in response to the effectiveness of punishment (Falk et al., 2005) clearly show that the potential for punishment is factored into social decision making. While there was an overall effect of status on Proposer offers, the current study also found some interactions between status and the effectiveness of punishment, indicating that status altered how individuals responded to this 'threat'. If this was just a general effect of cooperative opportunities as described above, there should not have been a relationship between status and effectiveness.

Therefore, a more specific explanation for the results might be that dominant individuals did not fear the repercussions their unfair behaviour may incite. Dominance is defined as the ability to have priority access to resources (Cummins, 1996a) and dominant individuals can use both aggressive and coercive tactics to achieve this (Jensen, 2010; Little, Henrich, Jones, & Hawley, 2003) because they are aware others will not react to this behaviour. Indeed, we are very unwilling to punish those who are labelled as 'high status' (Eckel et al., 2010; Petersen et al., 2012). In essence, a dominant position means one does not have to feel threatened by others (see, Watkins et al., 2010; Watkins & Jones, 2012). Furthermore, Fessler and Holbrook (2013) recently demonstrated that simply being surrounded by allies lowers the perceived threat of a potential adversary, and conversely, the lack of allies is a strong predictor of being a victim of harassment (Sapouna et al., 2011; Smith et al., 2004). Thus, when asked to make decisions that might lead to punishment, more dominant individuals in the group did not perceive punishment, and especially effective punishment, to be as threatening as lower status individuals did. It should be noted however, that the latter result should be treated with caution as, while Table 3 shows that the relationship between network status and cooperation only occurred under effective punishment when analysed separately, only two variables (Socialise on Trip and Socialise at Home) showed a clear interaction effect.

Nevertheless, the results of the current study support the suggestion that dominant individuals feel able to behave in a selfish manner in general. It also offers some support for the suggested that this effect occurs because dominant individuals feel insulated by their social position from the anger and recrimination that low cooperation or other acts of unfairness invite, and that low status individuals may cooperate more fully to avoid punishment rather than to help generate social ties.

7.5.2 Punishment and network position

The threat of counter-punishment/retaliation might be the key cost to punishment behaviour (Nikiforakis, 2008; Rand et al., 2010). So, if more dominant individuals do feel less threatened by the possible recriminations from their actions, we might expect that dominant individuals would also be intrinsically less afraid of any possible retaliation from punishment, and would therefore punish more. As with the cooperative decision making, being in a dominant position was expected to bias the behaviour of dominant individuals when they made third party punishment decisions, and we expected that dominant individuals would punish unfairness more severely than low status individuals. The current study found some evidence for this relationship between punishment behaviour and social position. Participants who showed the greatest dominance for the 'Influential' variable did show a greater sensitivity to unfairness when making their punishment decisions. There was a marginal effect when the entire weighted network, i.e. 'overall' status, was compared to punishment behaviour, with more central individuals showing greater sensitivity to unfairness.

It has been suggested that dominant individuals should be more sensitive to unfairness between conspecifics in their social environment (Cummins, 1999) as they have an extra incentive to track the changes in the social hierarchy as indicated by acts of unfairness (Brosnan, 2011; Cummins, 1996a). Thus, more dominant individuals should be more sensitive to unfair behaviour in their vicinity. This was demonstrated to a certain extent by the results of the current study. While there were no actual in-game mechanisms to motivate or otherwise trigger this behaviour (for example, the potential to lose position, Maner & Mead, 2010); as stated previously, the study investigated whether just being in a dominant position would lead to dominant-type behaviour, and more dominant individuals did display the expected behaviour to an extent.

As also suggested for the cooperation data, individuals who were central in the network felt insulated from any repercussion for their actions, specifically the threat of retaliation which, when present, reduces instances of costly punishment (Janssen & Bushman, 2008; Nikiforakis, 2008; Rand et al., 2010). Because of the cost of retaliation, only high status individuals may feel able to punish because they can absorb the costs; here, because they had a large amount of social support (Fessler & Holbrook, 2013). While there was no explicit retaliation cost (see below), participants did make decisions in close proximity to one another. Punishment does respond to the possibility of being identified (Kurzban et al., 2007; Rockenbach & Milinski, 2011) and if the presence of 'eye-like' images can effect behaviour (Bateson et al., 2006) then the presence of other participants should have had a similar effect (for a general discussion on anonymity in lab experiments, see Levitt & List, 2007).

However, the effect of social status on the tendency to punish was limited, especially when compared to the strong association between cooperation and social status. One potential reason may be the lack of opportunity for actual retaliation. There was a clear effect of social status on cooperation because punishment was possible and effective, i.e. the punishment mechanism (effective and ineffective) within the game was clearly salient in participants cooperative decision making and this direct risk of punishment robustly established a relationship between altruistic/fair behaviour and social position. While, as stated previously, the presence of others could suggested reputational gain and therefore the risk of retaliation (but see, Fehr & Schneider, 2009), this vague risk might not have been salient to participants, and therefore the insensitivity to threat shown by dominant individuals was not a factor when participants made their punishment decisions.

7.5.3 Conclusion

The current study investigated the effects social status on cooperative and punishment behaviour. The study found that position in a social network significantly affected

cooperation, with dominant individuals making substantially lower offers than subordinates. This was especially apparent when potential punishment from a third-party was effective, and suggests that more dominant individuals behave more selfishly because they implicitly fear punishment less than those of low status. In regard to actual punishment behaviour, there was some evidence that higher status individuals were more sensitive to acts of unfairness taking place within the group; they punished unfairness more severely than low status. These results suggest that status in a social hierarchy can affect sensitivity to acts of unfairness between its members which results in the *altruistic* punishment of defectors and free-riders.

7.6 Study 10: reputation, cooperation and punishment in an informal social network

Study 9 found that position in an informal social network did have an effect on participants' behaviour. Those in a dominant social position were overall less likely than others to behave altruistically towards other members of the group and seemed less sensitive to the threat of punishment. As discussed above, these results support other studies which have demonstrated how dominant individuals generally behave less altruistically (Piff et al., 2012; Sell, Tooby, et al., 2009; Zak et al., 2009) and it therefore also validates the use of social network data, specifically InDegree data, as a measure of dominance. However, while theoretically more dominant individuals might be expected to be more sensitive to unfairness (Brosnan, 2011; Cummins, 1999) and thus punish more, there was only slight evidence that this occurred.

One potential reason for the strong effect of social position on cooperation, but not punishment, seen in Study 9 was the lack of any game mechanic that might have made punishment a riskier decision. While the rationale of (evolutionary) economic games is partly that individuals 'bring in' general social behaviours from the outside world to the lab, we have argued previously that the mechanics of games might allow individuals to behave 'out of character', for example that effective (and anonymous) punishment in laboratory experiments gives everyone the ability to punish cheaply in a way they could not in everyday

life (see Chapter 6, 6.1, and Guala, 2012). With this in mind, the current study added an additional cost to engaging in punishment.

It was not feasible to add another level to the game mechanism, such as a strategy-method retaliation-to-punishment round so, instead, to increase the prospect of a social penalty of behaviour, the current study made all decisions non-anonymous. Studies on how reputation affects costly punishment have given contradictory results. Several studies have demonstrated that punishers gain a positive reputation from the act (Barclay, 2006; Nelissen, 2008), and that the potential for reputational gain can motivate punishment (Bering, 2008; Kurzban et al., 2007). However, other studies have found that individuals are willing to hide punishment (Ostrom, 1990; Rockenbach & Milinski, 2011), possibly because of an implicit fear of retaliation. Thus, the addition of a reputation mechanism of sorts[8] could have a two-fold effect. On the one hand it might induce participants to punish because punishers of unfairness are well liked, and on the other, the risk of (social) retaliation from the target of punishment might make subordinate individuals less willing to do so, despite the rewards.

Study 9 only found limited evidence that social position affected punishment behaviour. Based on our interpretation of how dominance affects punishment, adding the reputation mechanism was expected to extenuate these marginal results. Accordingly, it was predicted that dominant individuals in the social network would behave less altruistically but would punish unfairness more severely than subordinate group members.

[8] 'Of sorts' because the mechanism is not embedded within the study itself per se; the reputational information was not used by participants to make future experimental decisions. Instead it relies on participants feel uncomfortable with their behaviour being discovered by other members of the group. Ideally all decisions should be incentivised (Balliet et al., 2011; Levitt & List, 2007, and see the discussion at the end of this chapter) however this would not have been possible given the study location.

7.7 Method

7.7.1 Participants and research context

Participants were a different group of 2nd year undergraduate students from the University of Exeter who were attending a week-long field course at a Field Studies Council site in South Wales, UK. Data collection took place in April 2013, two years after Study 9. In total, 32 students (20 females, 12 males; mean age=21) attended the course and completed the social network questionnaire. All students took part in the Third Party Punishment Game (TPPG).

7.7.2 Design

As with Study 9, a pen-and-paper strategy method version of the third party punishment game was used (Fehr, 2004) with participants being asked to make proposal and punishment decisions simultaneously. In the present version of the TPPG, Proposers were given 30 points and could transfer between 0 to 15 points to the Receiver. This limit was implemented as individuals rarely send more than half their points (Fehr, 2004) and to limit the number of strategic decisions participants had to make in the Third Party role. Third Parties still received 10points. This alteration was made to extend the maximum effective spending on punishment decisions, i.e. only by spending ten points would a third party remove all the points from a very selfish Proposer.

7.7.3 Procedure

On the 6th day of the field course participants were informed that the study was taking place later that evening. The study took place in the dining hall of the field centre with participants spaced widely apart. They were informed that participation was entirely voluntary and were told that they had the potential to earn up to £40 by taking part. Once all participants were seated, the instruction sheet was handed out. This sheet explained the various roles and how payment would be generated (see 7.7.4 Matching & Payment Procedure). Instructions were

read out verbatim, and participants were instructed to raise their hand if they were unclear about any details.

Participants were then informed that their decisions would not be anonymous. Participants were told that when they received payment, they would be told how many points they received from a named Proposer and how many deduction points they were awarded by a named Third Party. Furthermore, the information would be displayed on a classroom white-board. The instructions stressed that the "**DECISIONS ON THE SCORE CARDS ARE NOT ANONYMOUS**" [Emphasis and capitalisation as in instructions].

Unlike Study 9, punishment was always effective; i.e one point assigned by a Third Party removed three of a Proposer's points (1:3 ratio). The Receiver had no active role in the game but participants were made aware that part of their payoff for the experiment would depend on what others had sent to them. Thus, the number of points (and therefore, money) participants earned depended on a) their proposal decision, b) points they received from a Proposer, c) their behaviour AS the Third Party, and d) the punishment received FROM a Third Party.

At this point the Proposer decision card was handed out. As Study 9 found no presentation order effects, all participants made their Proposer decision first, and once all participants had made their allocation decision this card was collected by the experimenters and the Third Party decision card was handed out. The Proposer card asked participants to indicate "How many points did they wish to send to the Receiver" and The Third Party card asked participants to indicate "The amount of points I would assign to the Proposer in response to their possible behaviour" with boxes available for each possible offer the Proposer could make (0-15). The experimental session lasted approximately 20 minutes.

7.7.4 Matching & Payment Procedure

After the experiment, participants were randomly matched into triads (containing a Proposer, a Third Party, and a Receiver). Twenty-one triads were generated as each participant was assigned once in each role and acted in each role in a different triad. Based on the cooperative and punishment decisions made in each triad, a point total was generated for each participant. Half of the participants were randomly selected to receive payment and their points were converted at a ratio of 5:1 (5 points = £1 pound sterling). Two participants were randomly selected to receive payment on a 1:1 ratio of points to £. Payment was made privately to those selected the following day and those selected to receive their payment at a ratio of 5:1 earned a mean amount of £5.78. Participants selected to receive payment at a ratio of 1:1 earned a mean amount of £22.25.

7.7.5 Generating the social network

In the same session, and prior to taking part in the TPPG, participants were asked to complete the social network questionnaire. As with the TPPG, instructions on the sheet were also read out verbatim, and participants were instructed to raise their hand if they were unclear about any details. The questionnaire asked participants to indicate who they believed to be the most influential individuals on the field course, who they socialised with on the field course, who (of those attending the course) they socialised with at home (i.e., in Exeter), and who they thought of as close friends. There was no limit on the number of individuals participants could identify for each question. The individual network questions, "Close Friends", "Most Influence", "Socialise at Home", "Socialise on Trip", correlated on a scale with high reliability (α=0.70), therefore the additional measure "Total Network" was created by summing the ties indicated by participants in the network questions. This was a weighted network, i.e. it represented the overall strength of associations within the group, rather than just their presence or absence. For the full questionnaire, see Appendix E.

7.7.6 Statistical analysis

The software package UCINET (Borgatti et al., 2002) was used to generate the network variables. As social network data violate assumptions of independence (Lusher et al., 2010), it was not possible to use standard software such as SPSS for the analysis. Therefore UCINET was also used for the analysis of the effects of the network variables, as it takes into account the non-independence of the network data when conducting statistical tests.

7.8 Results

While all 32 participants took part in the TPPG, data from three participants were removed from the analyses as their responses indicated they either did not understand the instructions or were filling out the 15 strategy decisions at random, for example using a pattern (2,0,2,0 etc) or with extreme variation (1,0,0,8,0,10,2 etc).

7.8.1 Cooperation and social position

The mean allocation made by participants in the proposer role was 9.6 points (SD=4.7). Only one participant sent an offer of zero. A series of simple linear regressions were carried out on the data; no network variable predicted a significant amount of the variance in proposer behaviour.

7.8.2 Punishment and social position

Eleven of the 29 participants did not engage in any punishment, and the mean amount spent on punishment by those who did engage in any punishment was 2.3 points (SD=1.2). To investigate whether a participant's social network position explained any of the variance in punishment behaviour, the average points assigned as punishment across the punishment decisions was calculated. The gradients of the regression lines relating to punishment responses to each potential offer were also calculated, where a large negative number would indicate participants punished unfair offers far more severely than fairer offers.

A series of simple linear regressions were carried out with the average overall punishment as the outcome variable; no network variable predicted a significant amount of the variance in average spending on punishment. A further series of simple linear regressions were carried out with the slope data as the outcome variable; again no network variable predicted a significant amount of the variance in sensitivity to unfairness.

As over a third of participants chose not to punish, punishment was also coded for whether it occurred at all. A series of binary logistic regressions were carried out within SPSS 20 found that network position did not predict whether participants would spend any points at all on punishment. It was not possible to conduct this analysis within UCINet, however given the nature of network data (see 7.7.6), analysis within SPSS can be considered less conservative. Thus the lack of significant findings can be seen as evidence that the decision to punish or not was not affected by group position.

7.9 Discussion

The current study investigated whether adding the possibility of reputational gain to a Third Party Punishment Game would magnify the differences found in Study 9 between higher and lower status members of a social group. This proved not to be the case, as the results showed no effect of network status on either cooperative or punishment behaviour.

7.9.1 Cooperation and social status

Study 9 found a strong association between selfish behaviour and a dominant position within the group, with more dominant individuals contributing less overall and being less sensitive to the threat of effective punishment. It was suggested that this was the result of a) higher status individuals behaving in a way implicitly biased by their status (Piff et al., 2012; Sell, Tooby, et al., 2009) and, b) specifically because they were less sensitive to the risk of behaving unfairly than lower status individuals (Watkins & Jones, 2012). There is good

reason for this trend, as high status individuals are less likely to face punishment for their behaviour (Petersen et al., 2012; see also, the world - the banking sector; celebrities; and the millionaire's son who killed a family of four while DUI, Walker, 2013). Accordingly, in the current study, we predicted that adding the threat of gaining a negative reputation would magnify this result. This was not the case.

Humans care a great deal about our reputation (Felson, 1982) and are very good at tracking the behaviour of others (Nowak, 2008; Nowak & Sigmund, 1998). This information is vital when making decisions regarding trust and cooperation (Duffy & Feltovich, 2002; Sylwester & Roberts, 2013), in fact the ability to the ability to avoid free-riders is another factor that can encourage cooperation (Rand et al., 2011; Santos et al., 2006). Therefore, one interpretation of the results is that the risk to their reputations ensured all participants acted in a pro-social manner, i.e. that the threat of being seen as 'mean' by members of the group was enough to prevent the selfish behaviour seen in Study 9. It should be noted, that just such a motivated was evident in Study 2 (Chapter 3). While differences in the mechanics of the games preclude a direct comparison, for illustrative purposes the mean offer in Study 9 under effective punishment was 57% of the total possible offer (48% under ineffective punishment), and in the current study the mean offer was 65%. Nevertheless, even with this increased 'threat', the explanations given in Study 9 should still apply, i.e. dominant individuals should feel more immune from the repercussions of their actions.

An alternative explanation is that introducing a reputation mechanism led to competitive altruism (Barclay & Willer, 2007; Roberts, 1998) whereby all individuals in the group wanted to be seen as behaving pro-socially (see also, Rockenbach & Milinski, 2011). Indeed, while under every day circumstances only high status individuals are generally successful in such contests (Barclay, 2013; Bird & Smith, 2005b), here all participants had equal resources. Also, while generally dominant individuals find it easier to gain a reputation because they are

under surveillance (Rege, 2008), the fact all that answers would be made public may have also 'levelled the playing field' in a similar fashion that effective punishment mechanism allow everyone to punish as if they were dominant. Therefore, when reputation was at stake, individuals behaved pro-socially to gain a positive reputation and this masked any effect of sensitivity to the fear of being punished may have had.

7.9.2 Punishment and social status

The inclusion of a reputation mechanism did not have the predicted effect on punishment. It was predicted that by providing the possibility for reputation/retaliation from the target of punishment, that only dominant individuals would punish unfairness, or punish at all. This was not the case.

The literature on the effect of reputation on punishment is mixed. On the one hand, a number of studies have demonstrated that punishment can be evolutionarily stable if there is some indirect benefit from reputation (Panchanathan & Boyd, 2004; Santos et al., 2010), and that individuals do punish more when under surveillance (Bering, 2008; Kurzban et al., 2007), although any positive gain is contingent on others believing the punishment was justified (Kiyonari & Barclay, 2008). However, when punishers can be identified (even simply as 'the person who punished you') they are retaliated against (Cinyabuguma et al., 2006; Nikiforakis, 2008) and this restricts any act of punishment. Rockenbach and Milinski (2011) recently demonstrated that if given the choice individuals will openly display generosity but will pay to hide punishment, and Ostrom (1990) found that a great deal of 'policing' takes place covertly, probably because of the threat of retaliation. Therefore, because more dominant members of the group can be seen as being at a lower risk of retaliation, we expected that more dominant individuals would punish more than subordinate individuals, and in turn access the positive reputation from punishment. However, the current study found

that when it was possible to gain a reputation as a punisher, status within the group did not affect punishment behaviour.

One reason might be that, regardless of status, the willingness of *disinterested* third parties[9] to engage in punishment might be overstated. While punishment is common in the laboratory generally (Guala, 2012), when given the opportunity individuals will hide their punishment behaviour (Ostrom, 1990; Rockenbach & Milinski, 2011) and in everyday life individuals are actually very unwilling to confront others for breaking social norms, for example in response to racist comments (Kawakami et al., 2009), or in reporting criminal activity (Tarling & Morris, 2010). In fact, when given a choice, participants would rather assist victims of unfairness or reward co-operators than punish unfairness (Ottone, 2008; Rand, Dreber, et al., 2009). Therefore in the current study it may have been the case that all involved were simply unwilling to punish.

Even so, given that Study 9 did find some relationship between a dominant group position and sensitivity to unfairness also without a direct motivation to punish, it is therefore be more parsimonious to suggest the inclusion of a reputation led to an overall lack of punishment because all participants were unwilling to be seen as punishing a group member. This does however raise an interesting question as to whether this unease at being seen as a punisher would have been overcome by more dominant individuals if a) there was a direct benefit from punishing within the experiment (for example through a reward round, Nelissen, 2008), or b) the position as identified by InDegree had additional effects on an individual's endowment and payoff within the TPPG.

[9] The role of a *motivated* punisher will be discussed below in the General Discussion (6.11.1, see also 1.1 and the definition of 'costly' and third party' punishment).

7.9.3 Conclusion

The current study investigated whether a dominant position in an informal social network would affect cooperative and punishment behaviour when these behaviours would be visible to all other individuals in the group. In contrast to the relationships found in Study 9, where decisions were anonymous, the current study found no such connections between dominance and behaviour in the TPPG. This could indicate that concern about the negative reputational effects of costly punishment, for example retaliation from the target, occur regardless of an individual's social position. Alternatively, the results might suggest that while dominant individuals *should* face less of a threat from retaliation, when there is no direct additional incentive to punish, even this lesser risk is too much.

7.10 General discussion

Using a strategy-method Third Party Punishment Game, Study 9 found some evidence that dominant individuals, as measured by their InDegree centrality in a closed social network, were more sensitive to unfair offers than subordinates. That is to say dominant individuals punished low offers more compared to more generous offers. In doing so, it provided some support for the suggestion that when in a dominant position, individuals become more sensitive to unfairness in the social environment. Study 10 introduced a reputation-like mechanism into the TPPG in which the decisions participants made would be made known to one another. As there are positive reputational gains to be made from engaging in punishment, and because a dominant position was theorized to make participants feel insulated from any reprisals, we expected this addition to magnify the effect seen in Study 9. This was not the case. These results need to be considered in the light of the other chapters. Chapter 3 found that punishers are seen as both likeable and dominant, Chapter 5 found that dominant individuals are expected to punish unfair behaviour and are expected to face less risk from retaliation, and Chapter 5 found that when individuals receive a benefit associated

with dominance they are also willing to punish more than others. Given these results, the question is why these effects did not emerge when actual metrics of dominance were measured?

7.10.1 The selfish punisher: dominance and the motivation to punish

Humans are angered by acts of unfairness, and anger is a strong predictor of costly punishment behaviour (Falk et al., 2005). It has been suggested that punishment of unfair or un-egalitarian behaviour is a human universal (Fehr & Gächter, 2002). An alternative explanation is that punishment is a spiteful behaviour from which group beneficial effects are a secondary outcome (Jensen, 2010): individuals are not sensitive to inequality, but to disadvantageous inequality, i.e. when *they themselves* are worse off once all interactions are over (Leibbrandt & López-Pérez, 2008, 2011); and it has been suggested that such spiteful behaviour is beneficial in the long run if it supresses any more prosperous conspecific's advantage (see also, Gavrilets, 2012; Van De Ven et al., 2010). Regardless of the motive, it is an inescapable fact that punishment is costly to the punisher (the threat of retaliation, Dreber & Rand, 2012; production costs, Egas & Riedl, 2008; the need for it to be effective, Nikiforakis & Normann, 2008; and the overall cost of second-order free-riding, Yamagishi, 1988). If only by facilitating innate altruistic or spiteful motivations, a dominant position allows these costs to be mitigated and thus should affect punishment behaviour. This did not occur in the current chapter's studies.

Perhaps the main reason was that the studies in the current chapter offered no proximate motivation to punish. In Chapter 5, for example, while additional resources did lead to an increase in punishment, the effect was most pronounced when the punisher was in a position to derive continuing benefit from enforcing cooperation. This effect can also be seen when any benefit is dependent on winning an inter-group conflict (Abbink et al., 2010). Further support for this can be found in a recent model by Roberts (2013), who specifically suggested

that as long as dominant individuals receive addition benefits from group cooperation, then punishment should always be in a dominant individual's interest. Indeed, even if individuals are behaving spitefully alone, there should be some advantage gained from that behaviour, such as the ability to give oneself the best relative outcome (Leibbrandt & López-Pérez, 2011).

According to this line of reasoning, in the absence of a specific in-game mechanic that would provide a benefit to the behaviour there was no motivation for dominant individuals to punish. In fact this is the rationale as to why economic games should provide incentive rewards structures for participants (Balliet et al., 2011; Levitt & List, 2007). Other studies have shown that game mechanisms can alter behaviour based on individual difference, for example dominant (Maner & Mead, 2010) and Machiavellianism (Gunnthorsdottir et al., 2002). Study 9 found that dominant individuals made low offers and this predicted result occurred because this decision had direct in-game repercussions, the game mechanism provided a salient threat that dominant individuals did(not) react to. While Study 10 added a reputation-like mechanism was an attempt to introduce a 'threat' from punishment, the reputation gained or lost from punishment did not have any direct ramifications experimentally, i.e. there was no additional game mechanic that depended on this information.

If punishers are primarily sensitive to direct cost and rewards of their actions, this raises the question as to whether the reputation gained from an act of punishment is an actual motivating benefit from punishment. As mentioned previously, the results of anonymity manipulations in costly punishment studies are mixed, and Chapter 2 of this thesis found that in a hypothetical scenario, individuals were insensitive to the reputational benefits from punishment. In fact, Study 2 suggested that participants would actually refuse to punish a social defector lest their own actions be seen in a negative light. So, perhaps the reputational

benefits seen in Chapters 3 & 4 are only taken into account when they have tangible results, for example ensuring one is treated fairly in future encounters by the target of punishment (Barclay, 2006; Clutton-Brock & Parker, 1995), to attract a sexual partner (Farthing, 2005), or to directly attract new cooperative partner members (Individuals like a punishment environment, Rockenbach & Milinski, 2006).

7.10.2 Measuring dominance

An alternative explanation for the lack of predicted results is that our measure was not an accurate measure of dominance in a group. InDegree centrality is a measure of how often an individual is identified by others in the group and is considered a measure of prestige/dominance (Scott, 2007) insomuch as it indicates that an individual is known to many individuals and that their behaviour is under observation (see, Henrich & Gil-White, 2001; Rege, 2008). We assumed that because participants should be able to accurately assess their group position (Anderson et al., 2008; Anderson, Srivastava, et al., 2006; Cummins, 1996a), that the general social feeling of popularity and social position would affect behaviour. However a dominant position alone may not affect behaviour; for example Maner and Mead (2010) found that behaviour is the result of an interaction between personality (specifically, those on the dominance/motivation axis) and a dominant position.

Thus, while an individual might have been prestigious in the network, they may not themselves have had a ‘dominant’ personality. Equally, we did not explicitly ask participants to, for example, guess their position and our prediction hinges on them being aware they occupy a dominant position. However, it should be noted that the cooperation results from Study 9 do support the suggestion that the InDegree measure accurately reflected dominance, as the patterns of proposer behaviour were those we predicted to would be expected of dominant individuals. Furthermore, our own measure of dominance in Study 9 did correlate strongly with network position.

7.10.3 Future directions

Many of the questions posed above could be addressed with one or two additional studies. Because Study 9 & 10 cannot be directly compared, a further study might use a between-group design, where half of the participants face a threat to reputation, and the other half do not. The issue that prevented such an experiment was sample size. Study 9 & 10 took advantage of a closed network of participants who, as part of their formal education, would spend a week in close proximity to one another away from the University. As this was an opportunistic sample, it was not possible to recruit into the study and the field-course has attracted fewer students in recent years.

We had expected that reputation would provide a good motivator for behaviour; however this was not the case. Therefore, a future study may wish to include additional game mechanisms, such as a retaliation round, which would provide direct consequences for punishment decisions. This was considered, but the test environment made such an addition unfeasible. The lack of a computer network with which to run real-time calculations, the time it would have taken to run calculations by hand in real time, and the low sample size would have made a 'strategy method retaliation round' the only viable solution. To our knowledge, such a method has never been used before, and it would have added an extra layer of complexity to an already somewhat complicated set of instructions. Fundamentally, compromises had to be made between maximum data collection and over-exploiting the goodwill of students, who were sacrificing their report-writing/free time to take part. Nevertheless, there is evidence that social relationships do affect behaviour in economic games (Haan, Kooreman, & Riemersma, 2006), so should these limitations be overcome, investigating the effects of social position within a closed informal network is a clear area for future work.

7.10.4 General conclusion

Dominance, in our opinion, provides two important mechanisms that might explain the evolution of costly punishment on an individual level. Firstly, it explains *how* an individual *can* punish. Without listing the literature again, dominant individuals intrinsically have, or can obtain more, resources (indeed, this is the definition of dominance), their social position or physical formidability allows them to punish effectively while expending few resources, and these same characteristics diminish the threat of retaliation. In addition, because individuals do like punishers, or more likely, like *someone else* punishing defectors, the positive associations this invites leads to more social allies and resources for the dominant individual.

Secondly, dominance explains *why* an individual *should* engage in punishment. In the first place,, punishment of conspecifics to maintain position is part of a dominant individual's behavioural repertoire across taxa (Clutton-Brock & Parker, 1995) and this is also very much present in humans (Barash & Lipton, 2011; Maner & Mead, 2010). Dominant individuals might be more sensitive to norm violations (Brosnan, 2011; Lammers et al., 2010), as in any group-living animal, violations take the form of the resources allocation 'rules' of a dominance hierarchy itself (Cummins, 2005). Also, as demonstrated by Chapter 5 (Study 8) individual will punish more when they benefit more from group success and thus it is in their best interest to punish free-riding to the extent that subordinates can simply free-ride on a dominant individuals punishment behaviour (Roberts, 2013).

What the current chapter suggests however is that while dominance might explain both the *how* and *why* individuals engage in punishment, the latter may be more important than the former. Thus, costly punishment it is not so much a case of *"with great power comes great responsibility"* as *"to the victor, the spoils"*.

8 General Discussion

8.1 Research question

The thesis aimed to investigate an alternative explanation for the evolution of costly punishment; that punishment and its associated emotions have an evolutionary origin in dominance and status contests. The studies of this thesis support this claim: Chapter 3 demonstrated that when asked to engage in punishment, the decisions made by participants were affected by the possibility of acquiring a 'dominant' reputation, and that dominance as an individual-difference was a strong predictor of punishing; Chapter 4 showed that when observing costly punishment, participants perceived punishment to be a dominant behaviour; Chapter 5 found that when predicting the outcomes of punishment scenarios, participants consistently made decisions based around the dominance of those involved; and Chapter 6 found that, when placed in a position simulating dominance, participants were more likely to engage in punishment. Finally, Chapter 7 found some evidence that being central in a group affected sensitivity to unfairness.

The findings of the thesis suggests that a dominant position and the social 'rules' of behaviour that govern dyadic dominance relationships and wider hierarchies (see Cummins, 1996a) played a role in the evolution of costly punishment and the concepts of fairness that spark its occurrence. However, it will be important to analyse how strong a role this was. Dominance must a) be shown to offer a coherent explanation for costly punishment behaviour and b) it must also be evident that dominance is directly related to punishment decision-making, as opposed to simply allowing its occurrence.

8.1.1 Dominance and proximate punishment behaviour

The studies in the thesis show that the inherent advantages of a high dominance rank have a strong proximate impact on punishment. Chapter 2 posited that a dominant position could

reduce the cost of punishment to an individual, and as a result dominant individuals should punish more. This was found to be the case. In Studies 7 & 8 (Chapter 6), when given one advantage of a dominant position, more resources, 'dominant' participants punished more severely and more frequently than those not in a dominant position. Furthermore, when this dominant position was stable and as a result the benefits from group cooperation were predictable, participants in this position punished even more frequently and severely. Equally, when participants were asked to consider the status of characters in a punishment scenario in Studies 5 & 6 (Chapter 5), not only were dominants expected to be more successful than subordinates, but the latter were expected to face more retaliation than the former. Perhaps because of this, and contrary to some research (Bering, 2008) but supported by others (Rockenbach & Milinski, 2011), Studies 1 & 2 (Chapter 3) showed that participants were less willing to punish when there was an audience and groups were stable.

Such a result may seem to contradict the results of 7 & 8 (Chapter 6), but the difference actually further supports the role of dominance in the proximate decision to punish. In the studies reported in Chapter 6, the mechanics of the games gave participants an incentive to punish and removed the possibility of retaliation, both advantages of a dominant position. As a result participants in this position punished more because the benefits were higher and the costs were lower for them respectively. However the studies in Chapter 3 asked participants to imagine themselves in a 'real-life' scenario' and retaliation is a social reality (see also Chapter 5, and see Barash & Lipton, 2011). This was factored into the cost and resulted in an unwillingness to punish when participants would 'meet' the target of punishment again. While the studies in Chapter 3 & 6 used two different methods, the results are comparable to economic games where retaliation is and is not possible (Nikiforakis, 2008).

The fact that punishment decision-making is affected by costs directly associated with dominance (see Chapter 1 and 2) has ramifications for any benefits the behaviour may

generate, through for example, reputation and signalling. Study 2 (Chapter 3) and 4 (Chapter 4) showed that participants only responded to an individual if that person actually engaged in punishment; in Study 4 the attempt at punishment was seen as sufficient for the signal to be honest, and in Study 2 participants only responded positively to a person who actually punished, regardless of any 'cheap signal' of a desire to punish (see Boyd et al., 2010). Indeed, Chapter 4 showed that participants treated costly punishment as a unique form of aggression in the sense that punishers were both well liked and thought of as dominant (Study 3). Study 4 showed, in terms of dominance at least, the greater the (potential) cost of punishment, the stronger the signal (see, Nelissen, 2008); punishers were seen as more dominant when the target of punishment was unknown to the punisher and thus posed a potentially greater threat.

Nevertheless, while punishment may be a costly signal, Chapter 5 demonstrated that any indirect benefits from such a signal were mediated by dominance. While Study 5 found that only dominant individuals were seen as being able to punish successfully, Study 6 showed that participants perceived that potential punishers would either punish successfully or would not try. Not only were subordinates not expected to attempt punishment at all, but any reputational benefits from punishment were shown to be mediated by its success. Thus whether there are indirect benefits from costly punishment are generated from signalling, as proposed by Gintis et al. (2001) and Nelissen (2008) or from indirect reciprocity and reputation in general, as proposed by Barclay (2006) and Panchanathan and Boyd (2004), these benefits are only available to dominant individuals.

The proximate effects of dominance also have ramifications for Spite (Leibbrandt & López-Pérez, 2011) and Strong Reciprocity (Gintis, 2000) theories of punishment. A Spite theory suggests that the motivation for costly punishment is to reduce the relative fitness of others in the environment, and even if this is the case, the results of the studies in Chapters 5 show

participants believe only dominant individuals would actually engage in punishment. Equally, as shown by the studies in Chapter 6, individuals might have been spiteful insomuch as their punishment harmed others at a cost to themselves, but only participants with additional resources behaved in such a way. Alternatively, even if humans really have a group-selected other-regarding preference for the well-being of others (Camerer & Fehr, 2006; Fehr & Fischbacher, 2003; Gintis, 2000), the studies in Chapter 3 showed that when asked to imagine a 'real world' scenario individuals are actually very unwilling to punish unfairness (for a economic example, see Pedersen et al., 2013), Chapter 5 showed that only dominants are predicted to intervene at all, and Chapter 6 showed intervention only happens when the net cost is low and punishment is selfishly beneficial. In fact, one reason only marginal effects of dominance were found in Chapter 7 is because there was no direct selfish incentive to punish unfairness.

In sum, a great deal of previous experimental evidence shows that punishment is affected by the proximate costs and benefits of the behaviour (see Chapter 1), and a number of models have suggested that heterogeneity in the cost of punishment is vital to its evolutionary stability (de Weerd & Verbrugge, 2011; Frank, 1996). The interpretation of the past research literature offered in Chapter 2 and, more importantly, the results of the current studies demonstrate that dominance presents a coherent and logical explanation for proximate variation in punishment behaviour. At the very least, when future research attempts to investigate punishment, dominance much be considered as an individual difference that has important ramifications for behaviour.

8.1.2 Dominance and the origins of costly punishment

The previous paragraph suggests that dominance is an important determinant of punishment behaviour, but the interpretation it offers can be taken as somewhat narrow and static, as the very fact that participants were making judgements about status, retaliation, future

interactions etc. suggests that dominance and status retaliations are part of the social reasoning surrounding punishment. Indeed, dominance and status play an important role in general social interactions, for example mate choice and competition, (Buss, 1989; Gambacorta & Ketelaar, 2013) or when following suggestions and/or advice (Henrich & Gil-White, 2001; Nelissen & Meijers, 2011). Therefore it is important to understand how this state of affairs has come about.

This thesis does not just suggest that dominance affects punishment by lowering the costs, but that the behaviour has an evolutionary origin in dominance and in the cognition require to navigate social relationships and dominance hierarchies (Byrne & Whiten, 1997; Cummins, 1996a, 2005; Erdal & Whiten, 1994). If this is true, there should be evidence that punishment behaviour by participants and the judgements of punishers by participants were explicitly affected by dominance and status, i.e. the pattern of results should indicate participants that were reasoning about dominance as opposed to, for example, just the proximate outcomes. The studies do show such patterns.

Firstly, the results suggest that punishment is itself a signal *of* dominance. The studies in Chapter 4 demonstrated that punishers are seen as dominant and, interestingly Study 4 showed that, while ratings of dominance were not affected by the success of punishment, the rank of the punishing group member was, i.e. participants were reasoning about the dominance relationships of the group based on the outcome of punishment. That punishment might be a signal *of* dominance was also suggested by the finding in Study 4 as 'likability' was not affected by the level of risk to the punisher, whereas dominance was. This would not be the case if it was an independent signal of pro-sociality alone, as likeability should have also increased with risk if this was the case. While there are advantages to associating with dominant individuals (Fessler et al., 2013; Snyder et al., 2011), there are also negative consequences in terms of consistent asymmetries in resource allocation and reciprocity

(Schino & Aureli, 2009; Sell, Tooby, et al., 2009). Thus, while costly punishment should generate positive regard as such behaviour does deter free-riding, its dominance origins make it a double-edged sword, and participant responses indicated this was part of their reasoning.After all, while it is beneficial to all that free-riders are removed, the more accurate version of the popular maxim is "*the enemy of my enemy could still be my enemy"*.

Secondly, Chapter 5 found that dominants were predicted to be successful if they did attempt to punish, but also that participants *expected* dominants to punish: participants were not surprised that a dominant group member attempted to punish unfairness, whereas they were very surprised to read of a subordinate attempting to punish. Study 8 in Chapter 6 also showed that there are expectations of dominant individuals: it was found that the strategic use of costly punishment by dominant individuals was affected by reasoning about dominance relationships and resource monopolisation. While participants with additional resources did punish more in general, this effect was most apparent when groups were stable *and* when punishers overtly benefited at the expense of the group. Revolutionary coalitions are a constant threat to dominants in primate societies (Bissonnette et al., 2014) and it has been suggested that dominants need to balance their own monopolisation of resources with pros-sociality in order to prevent this (Gavrilets & Fortunato, 2014; Powers & Lehmann, 2014; Vehrencamp, 1983). In Study 8 therefore, the study mechanism interacted with the dominance-based reasoning of costly punishment to produce the result: though participants did not face the threat of a revolutionary coalition, they behaved as if this was possible, but *only* when it was made clear they were benefiting at the expense of others (i.e. the monopoly condition). This may be why there was only a marginal effect of dominance seen in Chapter 7: although dominant individuals could have punished in Studies 9 & 10, there was no direct motivation to.

The fact that dominants are willing to punish 'in the public good' has ramifications for the evolution of leadership. It is useful for groups to have leaders: not only are they effective at coordinating cooperative behaviour (Gillet et al., 2010), but coordinated punishment (Schoenmakers et al., 2014; Traulsen et al., 2012), or punishment by a signal punisher (O'Gorman et al., 2009), can drastically reduce the waste caused by costly punishment. Nevertheless, despite the fact that coordinated punishment can support its evolution (Boyd et al., 2010), Study 2 found that participants did not respond to a 'cheap' signal of a willingness to punish. Because dominants have a freedom of action (Van Vugt, 2006), the ability to force coordination (King et al., 2009), and a vested interest in maintaining cooperation (Chapter 6, see also, Gavrilets & Fortunato, 2014), dominants might have been the force behind the initial coordination of punishment. Indeed, as reported by Guala (2012), in non-state societies, punishment when it does occur is highly coordinated, and dominant individuals do have a greater voice in decision making (Henrich & Gil-White, 2001) and can act as mediators in conflicts (Diamond, 2012, Chapter 2).

Furthermore, and in support of the above suggestion that dominants are expected to punish, the studies in Chapters 4 & 5 showed that punishment is seen a dominance contest. Specifically, *not* punishing was seen as subordinate act; characters that were seen as not punishing because the scenario description stated they did not punish, or because participants believed they would not, were seen as subordinate. This further supports the suggestion made in Chapter 2 and throughout the thesis, that punishment is not just made possible by dominance; the behaviour itself is a dominant behaviour. Indeed, this can be seen in everyday life. As a recent international example, while the decision of the UK government *not* to intervene in Syria may have been rationally correct, the initial reaction in some circles (as summarised by Anne Perkins of the Guardian, 2014), was that the lack of intervention was seen as an act of weakness by the UK. The fact that a failure to punish unfairness in an

unrelated conflict makes one seem 'weak' adds further support to the suggestion that punishment is not just something dominants can do, but something they are required to do, or to mis-quote Silk (2003), to remain dominant an individual should "practise random acts of *costly punishment and senseless acts of third party intervention*".

Finally, Study 6, where the rank of both punisher and the aggressor were manipulated, showed that participants believed the most retaliation would take place when individuals were of similar rank, as would be expected in a dominance hierarchy (Wilson, 1980, Chapter 13). This was especially interesting as subordinates who punished a dominant were seen as being at less risk of retaliation than those involved in a subordinate/subordinate conflict, and supports the assertion made above that costly punishment behaviour may have evolved to *appear* to be in the public good to avoid revolutionary coalitions, and that dominants are sometimes expected to behave leniently with subordinates (see, Fiddick & Cummins, 2007). Fundamentally, the results of the empirical and theoretical parts of this thesis suggest that the instigation of an act of punishment by punishers and it interpretation by observers occur through the prism of dominance and status contests.

8.2 Practical implications: costly punishment, dominance and leadership

The hypothesis that that costly punishment has an evolutionary origin in dominance and status potentially has practical implications for every day society, specifically in leadership and defence of the common good. As discussed in greater detail in 6.11.3, as in the course of human evolution our greater social cognition allowed more coalitional conflicts, it became possible for subordinates to stop others becoming too dominant (Boehm, 1997; Gavrilets, 2012; Gavrilets et al., 2008). It has been argued throughout this thesis that costly punishment might be the *price* of a dominant/leadership position; we will allow an individual priority access to resources as long as they appear to act in a pro-social fashion. The appearance of such a *noblesse oblige* has been shown theoretically (Gavrilets & Fortunato, 2014; Roberts,

2013) and experimentally (Fiddick & Cummins, 2007), and indeed it was demonstrated in Study 8 where dominants behaved in a pro-social fashion when it was made clear their priority of access to resources was at the expense of others.

Unlike a Strong Reciprocity theory of punishment, which suggests that evolutionary pressures resulted in a change from the steep and tyrannical social structures of group-living primates to an other-regarding and egalitarian mind-set, a dominance-based approach to fairness suggests our apparent other-regarding behaviour is conditional on the suppressive power of conspecifics. Indeed, this suggestion is supported by the lurch towards despotism that occurred following the advent of agriculture (Betzig, 2014; Turchin et al., 2013): once our species was free from the constraints of an immediate return economy, allowing the monopolisation of long-term resources (Charlton, 1997), dominants no longer needed to act in quite such a pro-social fashion.

Such an effect was reflected in the results of Study 8. Dominants behaved in a pro-social fashion *only* when it was made clear their priority of access to resources was at the expense of others. As discussed in more depth in 6.11.3, and in 8.2.2 this framing likely activated psychological mechanisms evolved both to gain the maximum from group cooperation and to avoid triggering revolutionary coalitions. However, if such pro-social behaviour is contingent on possible threats and benefits from the group, we can ask how individuals in a dominant position will behave when their private success is no longer directly linked to the public good. This has important real world ramifications for issues from banking reform to habitat destruction and climate change, as without either adequate reward or coercive threat, we cannot expect dominant individual or coalitions/organisations to simply act in the name of the common good. While those of us at the bottom of the social hierarchy may implicitly expect dominants to *altruistically* punish (Chapter 5), we may find that, as with Chapter 10, those who could punish in the name of the public good simply don't.

Furthermore, by establishing the role that dominance plays in costly punishment, the specific advantages of a dominant position can be extrapolated to encourage more defence of the public good. For example, not recognising that punishment needs to be sufficiently costly to the target led an increase, rather than the expected decrease, in short-term child abandonment in an Israeli day-care centre (Gneezy & Rustichini, 2000a). More seriously, if individuals feel they are unable to punish effectively, other public good activities such as whistleblowing may simply not occur. Thisincludes behaviour by individuals, corporations and governments. It is not simply an attitude of 'why bother?' but potentially of the implicit sense of subordination that an inability to punish effectively represents, and thus an acceptance of 'unfairness'.

Equally, the case of Simon Singh mentioned in Chapter 2 highlights how public good can be derailed by the threat of retaliation. This is not an isolated case; science writer Ben Goldacre was sued for writing about AIDS denialism (See Chapter 10, Goldacre, 2010), and recently a friend of the author faced similar retaliation for his own science-based critique of the same movement (Myles Power, personal communication, and see Doctorow, 2014). Thus, one direct outcome would be to ensure anyone acting in the public good is protected as much as possible from retaliation by their target.

8.3 Future directions

The studies in this thesis have established that dominance plays an important role in proximate punishment behaviour. Moreover, they have shown that dominance makes an important contribution to explaining the evolutionary origins of costly punishment. However this is not to say that a dominance-theory of punishment is completely established, or that all possible aspects of dominance and status have been tested. Nevertheless, with dominance established as a valid testable hypothesis for the evolution and occurrence of costly punishment, there are a great number of avenues open for future research. The next section will detail a few of these possible avenues.

8.3.1 Dominance, punishment and retaliation

Perhaps the most pressing area to research would be the use of retaliation to an act of costly punishment. Retaliation has been mentioned often during this thesis, but it was only tested directly by Studies 5 & 6, and tangentially by Study 10. This was a conscious decision, as retaliation is a down-steam effect of punishment, and it was important to establish dominance as a credible explanation before more elaborate work was undertaken. With the role of dominance established, there are a number of retaliation-based studies that could be undertaken specifically from a perspective of dominance.

One possible future study would be to simulate the suggested effect of a dominant position, of allowing an individual to mitigate or avoid retaliation themselves, while being very able to dispense it. For example, one (identifiable) individual could be given a monopoly on retaliation in the same way one individual was given a monopoly on punishment in Studies 7 & 8; or following the punishment mechanism employed by Nikiforakis et al. (2009), only one individual would be able to retaliate effectively. One would expect that 'subordinate' participants would not punish the only, or only effective, retaliator, whereas the 'dominant' participant would be very willing to punish non-cooperation and, perhaps, even second-order free-riding. Interestingly, this might help further ascertain the motive of punishment, for example in relation to spite: an individual who cannot be retaliated against, but can retaliate, may use this ability to simply lower the resources of others.

Additionally, the disproportionate benefits mechanism from Study 8 could be added to such an experimental series. The results could be fascinating: would the ability to retaliate reduce the contributions by the dominant individual overall, or would it interact with benefit from group success as it did with Study 8? Equally, as suggested by a model by Roberts (2013), would a producer-scrounger pattern emerge, whereby others relied on the dominant to punish? Furthermore, would the *noblesse oblige* extend to *not* punishing second-order free-

riding, or would the additional resources make second-order punishing both economical and practical? In fact, if spiteful behaviour was found without disproportionate benefit, would such a benefit increase encourage 'pro-social' behaviour as it did in Study 8?

Such an experimental series could help explain how humans transitioned to a pool-punishment (Traulsen et al., 2012) and institutionalised system of laws: it is in the interest of dominants to 'encourage' everyone to defend the public good, and especially so if dominants have the monopoly on retaliatory aggression. We know for instance that individuals will punish more in the presence of a motivated third party (Kim et al., 1998), and the consistency of law and order in a society has been linked to punishment (Herrmann et al., 2008). Thus, on the one hand, the selfish punishment motivations of dominants might subsequently allow to others to voluntarily punish as in the model by Boyd et al. (2010), and on the other it might be in the best interests of dominant to coerce other to punish in the public good. The private interest of dominants in maintaining group cooperation could be the genesis of the Hobbsian Leviathan.

8.3.2 Punishment and usefulness: a test of prestige

Dominance also covers individuals who are prestigious, and focusing on simulating the specific *usefulness* of an individual would open up a new way of investigating the role of dominance in costly punishment. Indeed, one of the criticisms that can be made of experiments that use dynamic group sorting (for example, Rand et al., 2011; Wang et al., 2012) is that the costs are minimal and uniform, whereas dominant/prestigious individuals would likely find the process generally much cheaper. Masclet (2003), for example, used an ostracism mechanism as punishment, and the negative impact of ostracising a 'useful' individual could be implemented easily within such an experimental design by varying the within-subject costs for such ostracism; will only 'cheap-to-remove' (i.e. low social value) free riders be ostracised? Alternatively, in order to simulate the greater outside options and

lower search costs of dominants, some individuals might be given the ability to ostracise/sever ties more cheaply than others. While similar to effective punishment in principle, the ostracism mechanism is actually removing someone from a group, rather than removing resources; dominants can (potentially) afford to lose allies more than subordinates.

Though in principle the costs in economic games and models can represent non-physical as well as physical costs, they rely on the infliction of direct costs on a target. The effect of prestige/dominance could be tested less abstractly by including mechanisms in which certain participants were *actually* useful to others. For example, an experiment could include a mixture of public good contribution dilemmas and puzzle rounds similar to the tasks used by Maner and Mead (2010), where some participants are known to have more clues to help solve a group problem. One would expect that participants would be very reluctant to ostracise individuals who possessed such additional knowledge for free-riding, even if it were cheap/free to do (Bowles & Gintis, 2004). Additionally, Henrich and Gil-White (2001) suggested that prestigious individuals face certain obligations; thus an extension to the previous experiment might be to add multiple 'prestigious' individuals to groups, only one of whom is needed for the 'puzzle' task. We might expect prestigious individuals to suddenly begin to behave much more cooperatively and punish free-riding more, compared both to other non-prestigious participants, and to prestigious individuals who do not face competition.

8.3.3 Dominance, punishment and ally retention

Gürerk et al. (2006) demonstrated that individuals prefer an environment where punishment is possible, and from that it was suggested that individuals prefer an environment where someone *would* punish. One of the arguments made for the indirect benefits of punishment throughout the thesis has been that it aids in the recruitment and retention of social allies. This too is eminently testable. A simple experiment would be to run a standard public goods

game with punishment, followed by a self-assorting by participants in future rounds after being provided with information about cooperative and punishment behaviour.

An initial experiment would keep participants in the dark about the self-assorting to prevent any effect of competitive altruism (see Roberts, 1998), but such competition could be introduced in further iterations of the experiment, as could several rounds of group self-assorting. A further extension would be to give the post-self-assorting public goods game a single-punisher mechanism, therefore making the choice of the 'right' dominant individual, based on their reputation, more important: individuals would be expected to choose a dominant with a reputation for fairness and defending the public good. Different levels of disproportional benefits for dominant could be added as a follow up and we might expect, as shown in a model by Powers and Lehmann (2014), to find a threshold at which individuals will no longer tolerate the (excessive) private sequestering of resources in return for protection. Finally, the last suggestion could be further extended by including either direct inter-group competition or a 'risky' environment where group extinction is possible; this would potentially test the limits of exchanging equality for security.

8.3.4 Punishment and social network position

Separately, a future study could also carry out a more in-depth investigation into the effect of social network position on punishment behaviour. One approach would be to extend the work carried out in Study 9 & 10 and apply the method to a larger organisation. In fact, many unsuccessful attempts were made to recruit the local orchestra of around 100 members, as this would have provided a large network with overlapping informal (based on friendship etc) and semi-formal (based on skill, instrument etc) hierarchies. At the very least such a large network would provide a great deal more data to compare to punishment decisions in a punishment game similar to those run in Studies 9 & 10. Such an organisation would also provide a sufficient sample size allow experimental manipulation, for example the possibility

of reputational gain, or whether the proposer or receiver in the game was a close friend of the third party: one can envisage individuals playing repeated third party games where players are either anonymous, only known to the third party, or all are aware who is playing the game with them.

Additionally, Fessler and Holbrook (2013) found that the presence of friends reduced the threats posed by potential aggressors, and I have suggested that social allies are another cost-lowering advantage of a dominant position. Thus individuals of different social distance to one another could be invited to the lab for other reasons, and at the time asked to fill out questionnaires similar to those of Fessler and Holbrook (2013). Alternatively, and in traditional social psychological fashion, participants might face a confederate behaving in an anti-social manner (for example, Kawakami et al., 2009). One may expect that individuals surrounded by close social allies will punish more than those surrounded by more distant associates, although additional factors such as the participants' overall dominance would also likely affect any decision.

8.3.5 Perceptions and expectations of dominance and punishment

As noted by a number of researchers (for example Guala, 2012; Levitt & List, 2007), one of the challenges of an experimental economics approach is extrapolating 'real-life' behaviour from such an artificial environment. This is especially true for dominance, as while we can simulate a dominant position, developing an experiment that accounts for the social dynamics of everyday life would be difficult; indeed, one reason half the studies in this thesis used an experimental survey method was to avoid this problem; in comparison, explaining social position was simple and effective. As argued in Chapter 3, describing a social situation brings in many implicit assumptions and understandings that would be hard to evoke in an experimental environment. Therefore future studies could also make further use of this method.

Importantly, the survey method could be used to probe more deeply the general social cognition surrounding punishment behaviour in a social hierarchy. The 'bar room' scenarios as described in Chapters 4 & 5 could be expanded into more subtle situations and interventions. For example, would the most senior or popular member of a social group or, say, a group of students in a psychology department, be expected by participants to intervene in 'unfair' disputes between members or to stop (non-criminal) anti-social behaviour? Conversely, scenarios could be constructed where such individuals fail or refuse to intervene, with participant judgements of a) the status of dominant following this lack of action and b) how much they *deserve* their position being recorded. This could be seen in concert with the suggestions put forward in 8.6.2 and 8.6.3 in regards to the 'usefulness' of dominants and how, as suggested throughout this thesis, costly punishment might be both a signal and a justification of a dominant (and leadership) position.

8.3.6 Punishment and dominance in a virtual environment

Although useful, it would be unethical to run large naturalistic punishment studies on unsuspecting groups, for example in the vein of Sherif, Harvey, White, Hood, and Sherif (1961)'s seminal Robbers Cave experiment. However, with modern computing other alternatives are potentially available. This thesis started with a description of a social defection in an early online environment, and the descendants of such games could provide an interesting study area. One such game, EVE online (CCP, 2003) has gained quite a reputation as it is completely player-led; outside an initial starting area there are no rules or regulations apart from those generated by the players. Indeed, EVE regularly makes headlines in the technology world due to Machiavellian manoeuvrings of players and their alliances. Importantly, there are real costs to such behaviour. Such costs are not only measured in time; although some items take months to build (LaLone, 2013), a recent conflict resulted in the

loss of in-games goods valued at $200,000 and represented the coordinated efforts of over 2,000 unrelated individuals (Chalk, 2014; Purchese, 2014).

While not a ‘real’ environment, simulated environments are becoming more viable as a research medium and are showing interesting results; for example Patil et al. (2014) recently demonstrated that participants make very different decisions in the classic ‘trolley’ moral dilemma when viewing a real-time virtual simulation on a computer than when reading the scenario on paper. And what are economic games but a simulated environment? Thus, an environment like EVE or another Massive Multiplayer Online (MMO) game, where coalitions are vital to success, betrayal and losses are meaningful and, importantly, where there is no central authority, would be fascinating for investigating hypotheses about dominance and costly punishment. While it is doubtful actual experiments could be run in such an environment, a great deal of qualitative data could be gathered about how individuals self-organise, make social decisions and enforce group-norms and group-beneficial behaviours. With such data, or even with historic date (if it exists), it would be easy to test hypotheses about the how the presence of punishers, dominants and leaders, and general social network strength, affect group survival.

I will stop here, but the possibilities for further investigating the role of dominance in the evolution are wide and varied. As mentioned above there are many facets that could be tested; the effect of dominance on retaliation, the market value of a dominant individual within a group and their ability to attract social partners, as well as the effect of ‘every day’ informal social hierarchies on behaviour. This includes more general research into how naturalistic or manipulated dominance affects the appraisal of social situations. The latter especially could result in interdisciplinary research between the evolutionary psychological perspective and general social psychology, which has its own theories about dominance, status and social behaviour.

Finally, the results of this thesis make an important general point about the validity of both theoretical models and practical economic experiments. As argued in Chapter 2 and through the empirical chapters, the characteristics of a dominant position correspond to many of the factors that have been experimentally and theoretically shown to affect the occurrence and evolution of punishment: thus, when any future studies identify values and mechanisms that seem to allow costly punishment or other cooperative behaviour to occur or evolve, the strongest effort should be made to establish how these would be represented in everyday life.

8.4 Shadow of the Leviathan: dominance and the evolution of costly punishment

The thesis has presented both theoretical and experimental evidence that dominance has played a key role in the evolution of costly punishment. Firstly, dominants are the only individuals that can take part in confrontational behaviour, pro-social or otherwise, because they have the physical or social strength to mitigate many of its costs (Chapter 5, and Clutton-Brock & Parker, 1995). Secondly, because of the priority of access to resources dominants enjoy, it is in their best interest to impose distribution 'rules' on conspecifics and maintain a dominant position (Chapter 5 and 6, and Cummins, 1996a). Thirdly, as human social cognition increased and conflicts relied more on coalitions and alliances (Gavrilets et al., 2008; Pietraszewski et al., 2014), ideas of 'fairness' were extended to associates (Brosnan, 2011), to help support and recruit social allies (Chapter 3 & 4). Finally, because of the threat of revolutionary collations (which have a long evolutionary history, Bissonnette et al., 2014), it is in dominant's best interest to engage in 'altruistic' punishment as a justification for their monopolisation of resources (Chapter 6, and Gavrilets & Fortunato, 2014).

Thus, the hypothesis put forward here is that costly punishment is fundamentally a tool used by dominants to maintain their social position, whether by preventing the rise of a social challenger, or by behaving in a manner that would help recruit and maintain social allies as

they provide protection from harm. However this protection is in exchange for acceptance of a dominants priority of access to resources. While changes in the ecology of early human societies, still present in non-state societies today, may have suppressed overt dominant behaviour, the theoretical and experimental evidence presented here suggests that dominance is still very much a part of costly punishment. The costly punishment of anti-social behaviour is a strong force for encouraging cooperation and other pro-social behaviour and this has led many to suggest we do have a fundamental other-regarding preference, but this thesis argues that in fact it is fundamentally grounded in self-serving, dominance-based, instincts.

Hobbes (1651/1996) proposed that society can only exist if some entity can curtail the baser instincts of man, and we continue to seek "*some kind of talisman, a benevolent tyrant or a magical new technology, that can shelter us from power and crime and protect us from each other"* (Gray, 2013). Costly punishment, and the instinct to enforce fairness and cooperation between one-another, seems to collectively grant us this desire. However, much like Hobbes' *Leviathan*, costly punishment can only occur when there exists someone who is able and willing to impose their will on others without fear of the cost or reprisal. The shadow of this Leviathan extends across our species' evolutionary history and, as this thesis states, the shadow originates in the dominance and status contests of our ancestors. Only by appreciating this origin will be able to better understand costly punishment and both when and why individuals will act in defence of the public good.

.

9 References

Abbink, K., Brandts, J., Herrmann, B., & Orzen, H. (2010). Intergroup conflict and intra-group punishment in an experimental contest game. *The American Economic Review, 100*(1), 420-447.

Acheson, J. M. (1988). *The lobster gangs of Maine*: University Press of New England.

Albert, M., Guth, W., Kirchler, E., & Maciejovsky, B. (2007). Are we nice(r) to nice(r) people? *Experimental Economics, 10*(1), 53-69.

Alexander, C. S., & Becker, H. J. (1978). The use of vignettes in survey research. *Public opinion quarterly, 42*(1), 93-104.

Anderson, C., Ames, D. R., & Gosling, S. D. (2008). Punishing hubris: The perils of overestimating one's status in a group. *Personality and Social Psychology Bulletin, 34*(1), 90-101.

Anderson, C., Srivastava, S., Beer, J. S., Spataro, S. E., & Chatman, J. A. (2006). Knowing your place: Self-perceptions of status in face-to-face groups. *Journal of Personality and Social Psychology, 91*(6), 1094.

Anderson, S., Benjamin, E., Cavanagh, J., & Collins, C. (2006). Executive Excess 2006: Defense and Oil Executives Cash in on Conflict: Washington, DC: Institute for Policy Studies/United for a Fair Economy.

Anderson, S. P., Goeree, J. K., & Holt, C. A. (1998). A theoretical analysis of altruism and decision error in public goods games. *Journal of Public Economics, 70*(2), 297-323.

Andreoni, J. (1988). Why free ride. *Journal of Public Economics, 37*(3), 291-304.

Andreoni, J. (1995). Cooperation in public-goods experiments: kindness or confusion? *The American Economic Review*, 891-904.

Aureli, F., Cozzolino, R., Cordischi, C., & Scucchi, S. (1992). Kin-oriented redirection among Japanese macaques: an expression of a revenge system? *Animal Behaviour, 44*, 283-291.

Baldassarri, D., & Grossman, G. (2011). Centralized sanctioning and legitimate authority promote cooperation in humans. *Proceedings of the National Academy of Sciences, 108*(27), 11023.

Balliet, D., Mulder, L. B., & Van Lange, P. A. M. (2011). Reward, punishment, and cooperation: A meta-analysis. *Psychological Bulletin, 137*(4), 594.

Barash, D. P., & Lipton, J. E. (2011). *Payback: Why we retaliate, redirect aggression, and take revenge*: Oxford University Press.

Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution and Human Behavior, 27*(5), 325-344. doi: 10.1016/j.evolhumbehav.2006.01.003

Barclay, P. (2010). Altruism as a courtship display: Some effects of third-party generosity on audience perceptions. *British Journal of Psychology, 101*(1), 123-135. doi: 10.1348/000712609x435733

Barclay, P. (2013). Strategies for cooperation in biological markets, especially for humans. *Evolution and Human Behavior, 34*(3), 164–175.

Barclay, P., & Reeve, H. K. (2012). The varying relationship between helping and individual quality. *Behavioral Ecology, 23*(4), 693-698.

Barclay, P., & Willer, R. (2007). Partner choice creates competitive altruism in humans. *Proceedings of the Royal Society B: Biological Sciences, 274*, 749-753.

Barker, J. L., Barclay, P., & Reeve, H. K. (2013). Competition over Personal Resources Favors Contribution to Shared Resources in Human Groups. *PLoS ONE, 8*(3), e58826.

Barr, A. (2001). Social dilemmas and shame-based sanctions: experimental results from rural Zimbabwe: Centre for the Study of African Economies, University of Oxford.

Barrett, L., & Henzi, S. P. (2006). Monkeys, markets and minds: biological markets and primate sociality *Cooperation in primates and humans* (pp. 209-232): Springer.

Barthes, J., Godelle, B., & Raymond, M. (2013). Human social stratification and hypergyny: toward an understanding of male homosexual preference. *Evolution and Human Behavior, 34*(3), 155-163.

Bassett, J. F., & Moss, B. (2004). Men and women prefer risk takers as romantic and nonromantic partners. *Current Research in Social Psychology, 9*(10), 135-144.

Bateson, M., Nettle, D., & Roberts, G. (2006). Cues of being watched enhance cooperation in a real-world setting. *Biology Letters, 2*(3), 412-414 doi: 10.1098/rsbl.2006.0509

10.1016/j.evolhumbehav.2004

Baumard, N., & Liénard, P. (2011). Second- or third-party punishment? When self-interest hides behind apparent functional interventions. *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.1112212108

Bazzan, A., & Dahmen, S. R. (2010). Bribe And Punishment: Effects Of Signaling, Gossiping, And Bribery In Public Goods Games. *Advances in Complex Systems, 13*(6), 755-771.

Benard, S. (2013). Reputation systems, aggression, and deterrence in social interaction. *Social science research, 42*(1), 230-245.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior, 10*(1), 122-142.

Bergman, T. J., Beehner, J. C., Cheney, D. L., & Seyfarth, R. M. (2003). Hierarchical classification by rank and kinship in baboons. *Science, 302*(5648), 1234.

Bering, J. (2008). The Effects of Perceived Anonymity on Altruistic Punishment. *Evolutionary Psychology, 6*(3), 487-501.

Berman, J. (2013). You're Severely Underpaid, And Here's Proof. Retrieved 16/04/2014, from http://www.huffingtonpost.com/2013/12/27/2014-minimum-wage_n_4501830.html

Betzig, L. (2014). Eusociality: From the first foragers to the first states. *Human Nature*, 1-5.

Bird, R. B., & Smith, E. (2005a). Costly signaling and cooperative behavior. *Moral sentiments and material interests: On the foundations of cooperation in economic life*, 115–148.

Bird, R. B., & Smith, E. (2005b). Signaling theory, strategic interaction, and symbolic capital. *Current Anthropology, 46*(2), 221-248.

Bird, R. B., Smith, E., & Bird, D. W. (2001). The hunting handicap: costly signaling in human foraging strategies. *Behavioral Ecology and Sociobiology, 50*(1), 9-19.

Bissonnette, A., Franz, M., Schülke, O., & Ostner, J. (2014). Socioecology, but not cognition, predicts male coalitions across primates. *Behavioral Ecology*, aru054.

Boehm, C. (1997). Impact of the human egalitarian syndrome on Darwinian selection mechanics. *The American Naturalist, 150*(S1), S100-S121.

Boehm, C. (2000). Conflict and the evolution of social control. *Journal of Consciousness Studies, 7*(1-2), 1-2.

Bonetti, S. (1998). Experimental economics and deception. *Journal of Economic Psychology, 19*, 377-395.

Borgatti, S. P., Everett, M. G., & Freeman, L. C. (2002). UCINET for Windows: Software for social network analysis. *Harvard Analytic Technologies, 2006*.

Boseley, S. (2009). Science writer accused of libel may take fight to European court. Retrieved 03/07, 2014, from http://www.theguardian.com/society/2009/may/13/simon-singh-british-chiropractic-association

Bowles, S., & Gintis, H. (2004). The evolution of strong reciprocity: cooperation in heterogeneous populations. *Theoretical Population Biology, 65*(1), 17-28.

Boyd, R., Gintis, H., & Bowles, S. (2010). Coordinated Punishment of Defectors Sustains Cooperation and Can Proliferate When Rare. *Science, 328*(5978), 617-620. doi: 10.1126/science.1183665

Boyd, R., Gintis, H., Bowles, S., & Richerson, P. J. (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences, 100*(6), 3531-3535.

Boyd, R., & Richerson, P. J. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology, 13*(3), 171-195.

Brañas-Garza, P., Cobo-Reyes, R., Espinosa, M. P., Jiménez, N., Kovárík, J., & Ponti, G. (2010). Altruism and social integration. *Games and Economic Behavior, 69*(2), 249-257.

Branas, C. C., Richmond, T. S., Culhane, D. P., Ten Have, T. R., & Wiebe, D. J. (2009). Investigating the link between gun possession and gun assault. *American journal of public health, 99*(11), 2034.

Brandt, H., Hauert, C., & Sigmund, K. (2003). Punishment and reputation in spatial public goods games. *Proceedings of the Royal Society of London. Series B: Biological Sciences, 270*(1519), 1099-1104.

Brandt, H., Hauert, C., & Sigmund, K. (2006). Punishing and abstaining for public goods. *Proceedings of the National Academy of Sciences, 103*(2), 495-497.

Brosnan, S. F. (2006). Nonhuman species' reactions to inequity and their implications for fairness. *Social Justice Research, 19*(2), 153-185.

Brosnan, S. F. (2011). An evolutionary perspective on morality. *Journal of Economic Behavior & Organization*.

Brosnan, S. F., Talbot, C. F., Ahlgren, M., Lambeth, S., & Schapiro, S. J. (2010). Mechanisms underlying responses to inequitable outcomes in chimpanzees, Pan troglodytes. *Animal Behaviour, 79*(6), 1229-1237. doi: 10.1016/j.anbehav.2010.02.019

Bshary, R., & Grutter, A. S. (2002). Asymmetric cheating opportunities and partner control in a cleaner fish mutualism. *Animal Behaviour, 63*(3), 547-555.

Buckley, E., & Croson, R. (2006). Income and wealth heterogeneity in the voluntary provision of linear public goods. *Journal of Public Economics, 90*(4), 935-955.

Bull, G. (2003). *The Prince by Niccolo Machiavelli*. Bury: Penguin Classics.

Burnham, T. C. (2007). High-testosterone men reject low ultimatum game offers. *Proceedings of the Royal Society B: Biological Sciences, 274*(1623), 2327.

Burns, J., & Visser, M. (2006). Bridging the great divide in south africa: Inequality and punishment in the provision of public goods. *rapport nr.: Working Papers in Economics*(219).

Buss, D. M. (1989). Sex differences in human mate preferences: Evolutionary hypotheses tested in 37 cultures. *Behavioral and Brain Sciences, 12*(01), 1-14.

Byrne, R. W., & Whiten, A. (1997). Machiavellian intelligence. *Machiavellian intelligence II: Extensions and evaluations*, 1-23.

Camerer, C. (2003). *Behavioral game theory: Experiments in strategic interaction* (Vol. 544): Princeton University Press Princeton, NJ.

Camerer, C., & Fehr, E. (2006). When does" economic man" dominate social behavior? *Science, 311*(5757), 47-52.

Cant, M. A. (2011). The role of threats in animal cooperation. *Proceedings of the Royal Society B: Biological Sciences, 278*(1703), 170-178.

Cant, M. A., Hodge, S. J., Bell, M. B., Gilchrist, J. S., & Nichols, H. J. (2010). Reproductive control via eviction (but not the threat of eviction) in banded mongooses. *Proceedings of the Royal Society B: Biological Sciences, 277*(1691), 2219-2226.

Cant, M. A., & Johnstone, R. A. (2009). How Threats Influence the Evolutionary Resolution of Within-Group Conflict. *The American Naturalist, 173*(6), 759-771.

Carpenter, J., Bowles, S., Gintis, H., & Hwang, S.-H. (2009). Strong reciprocity and team production: Theory and evidence. *Journal of Economic Behavior & Organization, 71*(2), 221-232.

Casari, M. (2005). On the design of peer punishment experiments. *Experimental Economics, 8*(2), 107-115.

CCP. (2003). Eve online. *CCP Games*.

Chagnon, N. A. (1988). Life histories, blood revenge, and warfare in a tribal population. *Science, 239*(4843), 985-992.

Chalk, A. (2014). EVE Online Mega-Battle Breaks $300,000 In Real-World Losses. Retrieved 01/07, 2014, from http://www.escapistmagazine.com/news/view/131799-EVE-Online-Mega-Battle-Breaks-300-000-In-Real-World-Losses

Chan, K. S., Mestelman, S., Moir, R., & Muller, R. A. (1999). Heterogeneity and the voluntary provision of public goods. *Experimental Economics, 2*(1), 5-30.

Chapais, B. (Ed.). (1992). *The role of alliances in social inheritance of rank among female primates*. Oxford: Oxford University Press.

Charlton, B. G. (1997). The inequity of inequality. *Journal of Health Psychology, 2*(3), 413.

Cheney, D. L. (2011). Extent and limits of cooperation in animals. *Proceedings of the National Academy of Sciences, 108*(Supplement 2), 10902.

Cheney, D. L., & Seyfarth, R. M. (1989). Redirected aggression and reconciliation among vervet monkeys, Cercopithecus aethiops. *Behaviour*, 258-275.

Cheng, J. T., Tracy, J. L., Foulsham, T., Kingstone, A., & Henrich, J. (2013). Two ways to the top: Evidence that dominance and prestige are distinct yet viable avenues to social rank and influence. *Journal of Personality and Social Psychology, 104*(1), 103.

Cheng, J. T., Tracy, J. L., & Henrich, J. (2010). Pride, personality, and the evolutionary foundations of human social status. *Evolution and Human Behavior, 31*(5), 334-347.

Chirot, D., & McCauley, C. (2010). *Why not kill them all?: the logic and prevention of mass political murder*: Princeton University Press.

Choi, J., & Bowles, S. (2007). The coevolution of parochial altruism and war. *Science, 318*(5850), 636.

Christakis, N., & Fowler, J. H. (2010). *Connected: The Amazing Power of Social Networks and How They Shape Our Lives*. London: HarperPress.

Cinyabuguma, M., Page, T., & Putterman, L. (2006). Can second-order punishment deter perverse punishment? *Experimental Economics, 9*(3), 265-279.

Clutton-Brock, T., & Parker, G. (1995). Punishment in animal societies. *Nature, 373*, 209-216.

Cordingly, D. (2006). *Under the black flag: The romance and the reality of life among the pirates*: Random House LLC.

Crockett, M. J., Clark, L., Lieberman, M., Tabibnia, G., & Robbins, T. (2010). Impulsive choice and altruistic punishment are correlated and increase in tandem with serotonin depletion. *Emotion, 10*(6), 855-862.

Cummins, D. (1996a). Dominance hierarchies and the evolution of human reasoning. *Minds and Machines, 6*(4), 463-480.

Cummins, D. (1996b). Evidence for the innateness of deontic reasoning. *Mind & Language, 11*(2), 160-190.

Cummins, D. (1999). Cheater detection is modified by social rank: The impact of dominance on the evolution of cognitive functions. *Evolution and Human Behavior, 20*(4), 229-248.

Cummins, D. (2005). Dominance, status, and social hierarchies. *The handbook of evolutionary psychology*, 676-697.

Dasgupta, P. (2011). Dark matters: Exploitation as cooperation. *Journal of Theoretical Biology*.

Davis, D. D. (1993). *Experimental economics*: Princeton university press.

Dawes, C. T., Fowler, J. H., Johnson, T., McElreath, R., & Smirnov, O. (2007). Egalitarian motives in humans. *Nature, 446*(7137), 794-796. doi: 10.1038/nature05651

Dawkins, R. (1976). *The selfish gene*: Oxford university press.

de Bruyn, E., Cillessen, A., & Weisfeld, G. (2012). Dominance-popularity status, behavior, and the emergence of sexual activity in young adolescents. *Evolutionary psychology: an international journal of evolutionary approaches to psychology and behavior, 10*(2), 296.

de Quervain, D., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., & Fehr, E. (2004). The neural basis of altruistic punishment. *Science, 305*, 1254-1258.

De Waal, F. (1982/2007). *Chimpanzee politics: Power and sex among apes. 25th anniversary edition*: JHU Press.

de Weerd, H., & Verbrugge, R. (2011). Evolution of altruistic punishment in heterogeneous populations. *Journal of Theoretical Biology*.

Denga, K., Gintis, H., & Chua, T. (2011). Strengthening Strong Reciprocity. *Journal of Theoretical Biology, 268*, 141-145.

Diamond, J. (2012). *The World Until Yesterday: What Can We Learn from Traditional Societies?* St Ives: Penguin.

Dibbel, J. (1999). My tiny life. *London: Fourth Estate*.

Doctorow, C. (2014). AIDS deniers use bogus copyright claims to censor critical Youtube videos. Retrieved 18/07/2014, 2014, from http://boingboing.net/2014/02/15/aids-deniers-use-bogus-copyrig.html

Dreber, A., & Rand, D. G. (2012). Retaliation and antisocial punishment are overlooked in many theoretical models as well as behavioral experiments. *The Behavioral and brain sciences, 35*(1), 24.

Dreber, A., Rand, D. G., Fudenberg, D., & Nowak, M. A. (2008). Winners don't punish. *Nature, 452*(7185), 348-351. doi: 10.1038/nature06723

Drews, C. (1993). The concept and definition of dominance in animal behaviour. *Behaviour*, 283-313.

Duffy, J., & Feltovich, N. (2002). Do actions speak louder than words? An experimental comparison of observation and cheap talk. *Games and Economic Behavior, 39*(1), 1-27.

Dunbar, R. I. M. (1998). The social brain hypothesis. *brain, 9*, 10.

Dunbar, R. I. M. (2004). Gossip in Evolutionary Perspective. *Review of general psychology, 8*(2), 100.

Dunbar, R. I. M. (2010). *How Many Friends Does One Person Need?* London: Faber & Faber Limited.

Dunbar, R. I. M., Clark, A., & Hurst, N. L. (1995). Conflict and cooperation among the Vikings: Contingent behavioral decisions. *Ethology and Sociobiology, 16*(3), 233-246.

Eckel, C., Fatas, E., & Wilson, R. (2010). Cooperation and status in organizations. *Journal of Public Economic Theory, 12*(4), 737-762.

Eckel, C., & Grossman, P. (2001). Chivalry and solidarity in ultimatum games. *Economic Inquiry, 39*(2), 171–188. doi: 10.1111/j.1465-7295.2001.tb00059.x

Egas, M., & Riedl, A. (2008). The economics of altruistic punishment and the maintenance of cooperation. *Proceedings of the Royal Society B: Biological Sciences, 275*(1637), 871-878. doi: 10.1098/rspb.2007.1558

Ellis, L. (1994). *Social Stratification and Socioeconomic Inequality: Volume 2: Reproductive and Interpersonal Aspects of Dominance and Status* Westpoint CT`: Praeger.

Ellis, L. (1995). Dominance and reproductive success among nonhuman animals: a cross-species comparison. *Ethology and Sociobiology, 16*(4), 257-333.

Emery, N. J., & Clayton, N. S. (2004). The mentality of crows: convergent evolution of intelligence in corvids and apes. *Science, 306*(5703), 1903-1907.

Erdal, D., & Whiten, A. (1994). On human egalitarianism: an evolutionary product of Machiavellian status escalation? (Vol. 35, pp. 175-183): JSTOR.

Ernest-Jones, M., Nettle, D., & Bateson, M. (2010). Effects of eye images on everyday cooperative behavior: a field experiment. *Evolution and Human Behavior*.

Etzioni, A. (2010). Behavioral economics: A methodological note. *Journal of Economic Psychology, 31*(1), 51-54. doi: 10.1016/j.joep.2009.09.004

Falk, A., Fehr, E., & Fischbacher, U. (2005). Driving forces behind informal sanctions. *Econometrica, 73*(6), 2017–2030. doi: 10.1111/j.1468-0262.2005.00644.x

Farthing, G. W. (2005). Attitudes toward heroic and nonheroic physical risk takers as mates and as friends. *Evolution and Human Behavior, 26*(2), 171-185.

Farthing, G. W. (2007). Neither daredevils nor wimps: Attitudes toward physical risk takers as mates. *Evolutionary Psychology, 5*(4), 754-777.

Fast, N. J., & Chen, S. (2009). When the boss feels inadequate Power, incompetence, and aggression. *Psychological Science, 20*(11), 1406-1413.

Fehl, K., van der Post, D. J., & Semmann, D. Co evolution of behaviour and social network structure promotes human cooperation. *Ecology Letters*.

Fehl, K., van der Post, D. J., & Semmann, D. (2011). Co-evolution of behaviour and social network structure promotes human cooperation. *Ecology Letters, 14*(6), 546-551.

Fehr, E. (2004). Third-party punishment and social norms. *Evolution and Human Behavior, 25*(2), 63-87. doi: 10.1016/s1090-5138(04)00005-4

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature, 425*(6960), 785-791.

Fehr, E., Fischbacher, U., & Gächter, S. (2002). Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature, 13*(1), 1-25.

Fehr, E., & Gächter, S. (2000). Association, Cooperation and Punishment in Public Goods Experiments. *The American Economic Review, 90*(4), 980-994.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature, 415*(6868), 137-140.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The quarterly journal of economics, 114*(3), 817-868.

Fehr, E., & Schneider, F. (2009). Eyes are on us, but nobody cares: are eye cues relevant for strong reciprocity? *Proceedings of the Royal Society B: Biological Sciences, 277*(1686), 1315-1323. doi: 10.1098/rspb.2009.1900

Felson, R. B. (1982). Impression management and the escalation of aggression and violence. *Social Psychology Quarterly*, 245-254.

Fessler, D. M., & Haley, K. J. (2003). The Strategy of Affect: Emotions in Human Cooperation 12. In P. Hammerstein (Ed.), *Genetic and Cultural Evolution of Cooperation* (pp. 7–36). Cambridge, MA: MIT Press.

Fessler, D. M., & Holbrook, C. (2013). Friends Shrink Foes The Presence of Comrades Decreases the Envisioned Physical Formidability of an Opponent. *Psychological Science, 24*(5), 797-802.

Fessler, D. M., Tiokhin, L. B., Holbrook, C., Gervais, M. M., & Snyder, J. K. (2013). Foundations of the Crazy Bastard Hypothesis: Nonviolent physical risk-taking enhances conceptualized formidability. *Evolution and Human Behavior, 35*(1), 26-33.

Fiddick, L., & Cummins, D. (2007). Are perceptions of fairness relationship-specific? The case of noblesse oblige. *The Quarterly Journal of Experimental Psychology, 60*(1), 16-31.

Figueredo, A. J., Corral-Verdugo, V., Frías-Armenta, M., Bachar, K. J., White, J., McNeill, P. L., . . . del PilarCastell-Ruiz, I. (2001). Blood, solidarity, status, and honor: The sexual balance of power and spousal abuse in Sonora, Mexico. *Evolution and Human Behavior, 22*(5), 295-328.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics, 10*(2), 171-178.

Fischbacher, U., & Gächter, S. (2005). Heterogeneous social preferences and the dynamics of free riding in public goods. *Institute for Empirical Research in Economics - IEW, IEW - Working Papers, 261*.

Fischbacher, U., & Gächter, S. (2010). Social preferences, beliefs, and the dynamics of free riding in Public Good Experiments. *American Economic Review, 100*(1), 541-556.

Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are People Conditionally Cooperative Evidence from a public goods experiment. *Economics Letters, 71*(3), 397-404.

Flack, J. C., & De Waal, F. B. M. (2000). Any animal whatever. Darwinian building blocks of morality in monkeys and apes. *Journal of Consciousness Studies, 7, 1*(2), 1-29.

Flack, J. C., de Waal, F. B. M., & Krakauer, D. C. (2005). Social structure, robustness, and policing cost in a cognitively sophisticated species. *The American Naturalist, 165*(5), 126-139.

Flack, J. C., Girvan, M., De Waal, F. B. M., & Krakauer, D. C. (2006). Policing stabilizes construction of social niches in primates. *Nature, 439*(7075), 426-429.

Frank, S. A. (1996). Policing and group cohesion when resources vary. *Animal Behaviour, 52*, 1163-1169.

Frank, S. A. (2003). Repression of competition and the evolution of cooperation. *Evolution, 57*(4), 693-705.

Fraser, O. N., & Bugnyar, T. (2011). Reciprocity of agonistic support in ravens. *Animal Behaviour*.

Freeman, L. C. (1979). Centrality in social networks conceptual clarification. *Social networks, 1*(3), 215-239.

Fudenberg, D., & Pathak, P. A. (2010). Unobserved punishment supports cooperation. *Journal of Public Economics, 94*(1-2), 78-86. doi: 10.1016/j.jpubeco.2009.10.007

Gächter, S., Herrmann, B., & Thoni, C. (2004). Trust, voluntary cooperation, and socio-economic background: survey and experimental evidence. *Journal of Economic Behavior & Organization, 55*(4), 505-531. doi: 10.1016/j.jebo.2003.11.006

Gächter, S., Herrmann, B., & Thoni, C. (2005). Cross-cultural differences in norm enforcement. *Behavioral and Brain Sciences, 28*(6), 822-823.

Gächter, S., Renner, E., & Sefton, M. (2008). The long-run benefits of punishment. *Science, 322*(5907), 1510.

Galinsky, A. D., Gruenfeld, D. H., & Magee, J. C. (2003). From Power to Action. *Journal of Personality and Social Psychology, 85*(3), 453-466.

Gambacorta, D., & Ketelaar, T. (2013). Dominance and deference: Men inhibit creative displays during mate competition when their competitor is strong. *Evolution and Human Behavior, 34*(5), 330-333.

García, J., & Traulsen, A. (2012). Leaving the loners alone: Evolution of cooperation in the presence of antisocial punishment. *Journal of Theoretical Biology, 307*(21), 168–173.

García, J., & van den Bergh, J. C. J. M. (2010). Evolution of parochial altruism by multilevel selection. *Evolution and Human Behavior, 32*(4), 277–287.

Gardner, A., & West, S. A. (2004a). Cooperation and Punishment, Especially in Humans. *The American Naturalist, 164*(6), 753-764.

Gardner, A., & West, S. A. (2004b). Spite and the scale of competition. *Journal of Evolutionary Biology, 17*(6), 1195-1203. doi: 10.1111/j.1420-9101.2004.00775.x

Gavrilets, S. (2012). On the evolutionary origins of the egalitarian syndrome. *Proceedings of the National Academy of Sciences, 109*(35), 14069-14074.

Gavrilets, S., Duenez-Guzman, E. A., & Vose, M. D. (2008). Dynamics of alliance formation and the egalitarian revolution. *PLoS ONE, 3*(10), e3293.

Gavrilets, S., & Fortunato, L. (2014). A solution to the collective action problem in between-group conflict with within-group inequality. *Nature communications, 5*.

Georgesen, J., & Harris, M. J. (2006). Holding onto power: Effects of powerholders' positional instability and expectancies on interactions with subordinates. *European Journal of Social Psychology, 36*(4), 451-468.

Gest, S. D., Graham-Bermann, S. A., & Hartup, W. W. (2002). Peer experience: Common and unique features of number of friendships, social network centrality, and sociometric status. *Social Development, 10*(1), 23-40.

Gillet, J., Cartwright, E., & Vugt, M. v. (2010). Selfish or servant leadership? Evolutionary predictions on leadership personalities in coordination games. *Personality and Individual Differences*. doi: 10.1016/j.paid.2010.06.003

Gintis, H. (2000). Strong Reciprocity and Human Sociality. *Journal of Theoretical Biology, 206*(2), 169-179. doi: 10.1006/jtbi.2000.2111

Gintis, H., Smith, E., & Bowles, S. (2001). Costly signaling and cooperation. *Journal of Theoretical Biology, 213*(1), 103-119.

Gneezy, U., & Rustichini, A. (2000a). A Fine is a Price. *The Journal of Legal Studies, 29*(1), 1-17.

Gneezy, U., & Rustichini, A. (2000b). Pay enough or don't pay at all. *The quarterly journal of economics, 115*(3), 791-810.

Goldacre, B. (2010). *Bad science: Quacks, hacks, and big pharma flacks*: Random House LLC.

Goodman, S. (2014). Good Guys, Bad Guys and Guns. Retrieved 26/06, 2014, from http://www.huffingtonpost.com/sandy-goodman/good-guys-bad-guys-and-gu_b_5482186.html

Gordon, D., & Platek, S. (2009). Trustworthy? The Brain knows: Implicit neural responses to faces that vary in dark triad personality characteristics and trustworthiness. *Journal of Social, Evolutionary, and Cultural Psychology, 3*(3), 182-200.

Gray, J. (2013). A Point Of View: Bitcoin's freedom promise. Retrieved 24/06, 2014, from http://www.bbc.co.uk/news/magazine-22292708

Graziano, W. G., Jensen-Campbell, L., Todd, M., & Finch, J. (1997). Interpersonal attraction from an evolutionary psychology perspective: Women's reactions to dominant and prosocial men. *Evolutionary social psychology*, 141-167.

Gregory, S. W., & Webster, S. (1996). A nonverbal signal in voices of interview partners effectively predicts communication accommodation and social status perceptions. *Journal of Personality and Social Psychology, 70*(6), 1231.

Griskevicius, V., Tybur, J. M., Gangestad, S. W., Perea, E. F., Shapiro, J. R., & Kenrick, D. T. (2009). Aggress to impress: Hostility as an evolved context-dependent strategy. *Journal of Personality and Social Psychology, 96*(5), 980.

Guala, F. (2012). Reciprocity: Weak or strong? What punishment experiments do (and do not) demonstrate. *Behavioral and Brain Sciences, 35*(1), 1.

Gunnthorsdottir, A., McCabe, K., & Smith, V. (2002). Using the Machiavellianism instrument to predict trustworthiness in a bargaining game. *Journal of Economic Psychology, 23*(1), 49-66.

Gürerk, O., Irlenbusch, B., & Rockenbach, B. (2006). The Competitive Advantage of Sanctioning institutions. *Science, 312*(5770), 108-111 doi: 10.1126/science.1123633

Haan, M., Kooreman, P., & Riemersma, T. (2006). *Friendship in a public good experiment*: IZA.

Haley, K., & Fessler, D. (2005). Nobody's watching?Subtle cues affect generosity in an anonymous economic game. *Evolution and Human Behavior, 26*(3), 245-256. doi: 10.1016/j.evolhumbehav.2005.01.002

Hamilton, W. D. (1964). The genetic evolution of social behavior. *Journal of Theoretical Biology, 7*(1), 1–16.

Hand, J. L. (1986). Resolution of social conflicts: dominance, egalitarianism, spheres of dominance, and game theory. *Quart. Rev. Biol., 61*, 201-220.

Harcourt, A. H., & De Waal, F. B. (1992). Coalitions and alliances in humans and other animals.

Hardin, G. (1968). The tragedy of the commons. *Science, 162*(3859), 1243-1248.

Harrison, F., Sciberras, J., & James, R. (2011). Strength of Social Tie Predicts Cooperative Investment in a Human Social Network. *PLoS ONE, 6*(3), e18338.

Hawkes, K. (1991). Showing off: tests of an hypothesis about men's foraging goals. *Ethology and Sociobiology, 12*(1), 29-54.

Hawley, P. H. (1999). The ontogenesis of social dominance: A strategy-based evolutionary perspective. *Developmental Review, 19*, 97-132.

Hawley, P. H. (2014). Ontogeny and social dominance: A developmental view of human power patterns. *Evolutionary Psychology, 12*(2), 318-342.

Hawley, P. H., Little, T. D., & Card, N. A. (2008). The myth of the alpha male: A new look at dominance-related beliefs and behaviors among adolescent males and females. *International Journal of Behavioral Development, 32*(1), 76.

Helbing, D., Szolnoki, A., Perc, M., & Szabó, G. (2010). Punish, but not too hard: how costly punishment spreads in the spatial public goods game. *New Journal of Physics, 12*(8), 083005. doi: 10.1088/1367-2630/12/8/083005

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., . . . Ensminger, J. (2005). "Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences, 28*(06), 795-815.

Henrich, J., Ensminger, J., McElreath, R., Barr, A., Barrett, C., Bolyanatz, A., . . . Ziker, J. (2010). Markets, Religion, Community Size, and the Evolution of Fairness and Punishment. *Science, 327*(5972), 1480-1484. doi: 10.1126/science.1182238

Henrich, J., & Gil-White, F. J. (2001). The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evolution and Human Behavior, 22*(3), 165-196.

Herrmann, B., & Gächter, S. (2009). Reciprocity, culture and human cooperation: previous insights and a new cross-cultural experiment. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1518), 791-806. doi: 10.1098/rstb.2008.0275

Herrmann, B., Thoni, C., & Gächter, S. (2008). Antisocial Punishment Across Societies. *Science, 319*(5868), 1362-1367. doi: 10.1126/science.1153808

Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics. *Behavioral and Brain Sciences, 24*, 383–451.

Hilbe, C., & Sigmund, K. (2010). Incentives and opportunism: from the carrot to the stick. *Proceedings of the Royal Society B: Biological Sciences, 277*(1693), 2427-2433.

Hill, K., Barton, M., & Hurtado, A. M. (2009). The emergence of human uniqueness: Characters underlying behavioral modernity. *Evolutionary Anthropology: Issues, News, and Reviews, 18*(5), 187-200.

Hirschman, A. O. (1970). *Exit, voice, and loyalty: Responses to decline in firms, organizations, and states* (Vol. 25): Harvard university press.

Hobbes, T. (1651/1996). *Leviathan*. Oxford, University Press.

Hogg, M. A., van Knippenberg, D., & Rast, D. E. (2012). The social identity theory of leadership: Theoretical origins, research findings, and conceptual developments. *European Review of Social Psychology, 23*(1), 258-304. doi: 10.1080/10463283.2012.741134

Houser, D., & Kurzban, R. (2002). Revisiting kindness and confusion in public goods experiments. *The American Economic Review, 92*(4), 1062-1069.

Huston, T. L., Ruggiero, M., Conner, R., & Geis, G. (1981). Bystander intervention into crime: A study based on naturally-occurring episodes. *Social Psychology Quarterly*, 14-23.

Iredale, W., Van Vugt, M., & Dunbar, R. (2008). Showing Off in Humans Male Generosity as a Mating Signal. *Evolutionary Psychology, 6*(3), 386-392.

Iris, B., Herrmann, B., & Zeckhauser, R. (2009). Trust and the reference point for trustworthiness in gulf and western countries. *Working Paper Series, Harvard University*.

Jaffe, K. (2008). Evolution of shame as an adaptation to social punishment and its contribution to social cohesiveness. *Complexity, 14*(2), 46-52.

Janssen, M. A., & Bushman, C. (2008). Evolution of cooperation and altruistic punishment when retaliation is possible. *Journal of Theoretical Biology, 254*, 241-545.

Jennings, D. J., Carlin, C. M., Hayden, T. J., & Gammell, M. P. (2011). Third-party intervention behaviour during fallow deer fights: the role of dominance, age, fighting and body size. *Animal Behaviour, 81*(6), 1217-1222.

Jensen, K. (2010). Punishment and spite, the dark side of cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences, 365*(1553), 2635-2650. doi: 10.1098/rstb.2010.0146

Jensen, K., Call, J., & Tomasello, M. (2013). Chimpanzee responders still behave like rational maximizers. *Proceedings of the National Academy of Sciences, 110*(20), E1837-E1837.

Jenson, N. H., & Peterson, M. B. (2011). To Defer or To Stand Up? How Offender Formidability Affects Third Party Moral Outrage. *Evolutionary Psychology, 9*(1), 118-136.

Johnstone, R. A. (2001). Eavesdropping and animal conflict. *Proceedings of the National Academy of Sciences, 98*(16), 9177-9180.

Johnstone, R. A., & Bshary, R. (2004). Evolution of spite through indirect reciprocity. *Proceedings of the Royal Society B: Biological Sciences, 271*(1551), 1917-1922. doi: 10.1098/rspb.2003.2581

Jonason, P. K., & Webster, G. D. (2010). The dirty dozen: A concise measure of the dark triad. *Psychological Assessment, 22*(2), 420-432. doi: 10.1037/a0019265

Jones, B. C., DeBruine, L. M., Little, A. C., Watkins, C. D., & Feinberg, D. R. (2011). 'Eavesdropping'and perceived male dominance rank in humans. *Animal Behaviour*.

Jones, B. C., & Rachlin, H. (2006). Social discounting. *Psychological Science, 17*(4), 283.

Kahneman, D., Knetsch, J., & Thaler, R. (1990). Experimental Tests of the Endowment Effect and the.

Kamei, K., & Putterman, L. (2012). In Broad Daylight: Full Information and Higher-order Punishment Opportunities Promote Cooperation. *Working Paper, Brown University, Department of Economics*.

Kawakami, K., Dunn, E., Karmali, F., & Dovidio, J. F. (2009). Mispredicting affective and behavioral responses to racism. *Science, 323*(5911), 276-278.

Kazem, A. J., & Aureli, F. (2005). Redirection of aggression: multiparty signalling within a network. *Animal communication networks*, 191-218.

Keeley, L. (1996). *War Before Civilisation: The Myth of the Peaceful Savage*. Oxford: University Press.

Keltner, D., Gruenfeld, D. H., & Anderson, C. (2003). Power, approach, and inhibition. *Psychological review, 110*(2), 265.

Ketelaar, T., Koenig, B. L., Gambacorta, D., Dolgov, I., Hor, D., Zarzosa, J., . . . Wells, L. (2012). Smiles as signals of lower status in football players and fashion models: evidence that smiles are associated with lower dominance and lower prestige. *Evolutionary Psychology, 10*(3).

Kim, S. H., Smith, R. H., & Brigham, N. L. (1998). Effects of power imbalance and the presence of third parties on reactions to harm: Upward and downward revenge. *Personality and Social Psychology Bulletin, 24*(4), 353.

King, A. J., Douglas, C., Huchard, E., Isaac, N. J., & Cowlishaw, G. (2008). Dominance and affiliation mediate despotism in a social primate. *Current Biology, 18*(23), 1833-1838.

King, A. J., Johnson, D. D. P., & Van Vugt, M. (2009). The origins and evolution of leadership. *Current Biology, 19*(19), R911-R916.

Kiyonari, T., & Barclay, P. (2008). Cooperation in social dilemmas: Free riding may be thwarted by second-order reward rather than by punishment. *Journal of Personality and Social Psychology, 95*(4), 826-842. doi: 10.1037/a0011381

Krause, J., Croft, D., & James, R. (2007). Social network theory in the behavioural sciences: potential applications. *Behavioral Ecology and Sociobiology, 62*(1), 15-27.

Kurzban, R. (2012). *Why everyone (else) is a hypocrite: Evolution and the modular mind*: Princeton University Press.

Kurzban, R., Descioli, P., & O'brien, E. (2007). Audience effects on moralistic punishment. *Evolution and Human Behavior, 28*(2), 75-84. doi: 10.1016/j.evolhumbehav.2006.06.001

Kurzban, R., & Houser, D. (2005). Experiments investigating cooperative types in humans: A complement to evolutionary theory and simulations. *Proceedings of the National Academy of Sciences of the United States of America, 102*(5), 1803.

LaLone, N. (2013). EVE Online Players Blow Up Rare Ship. Retrieved 01/07/2014, 2014, from http://www.warcry.com/news/view/125771-EVE-Online-Players-Blow-Up-Rare-Ship

Lamba, S., & Mace, R. (2011). Demography and ecology drive variation in cooperation across human populations. *Proceedings of the National Academy of Sciences, 108*(35), 14426-14430.

Lamba, S., & Mace, R. (2013). The evolution of fairness: explaining variation in bargaining behaviour. *Proceedings of the Royal Society B: Biological Sciences, 280*(1750).

Lammers, J., Stapel, D., & Galinsky, A. (2010). Power Increases Hypocrisy: Moralizing in Reasoning, Immorality in Behavior. *Psychological Science, 21*(5), 737-744. doi: 10.1177/0956797610368810

Laville, S. (2012). Chris Huhne resigns over criminal charge in speeding case. Retrieved 09/06/2014, 2014, from http://www.theguardian.com/politics/2012/feb/03/chris-huhne-expected-resign-charges-speeding

Le Roux, A., Snyder-Mackler, N., Roberts, E. K., Beehner, J. C., & Bergman, T. J. (2013). Evidence for tactical concealment in a wild primate. *Nature communications, 4*, 1462.

Lehmann, L., & Feldman, M. W. (2008). War and the evolution of belligerence and bravery. *Proceedings of the Royal Society B: Biological Sciences, 275*(1653), 2877-2885.

Lehmann, L., & Keller, L. (2006). The evolution of cooperation and altruism–a general framework and a classification of models. *Journal of Evolutionary Biology, 19*(5), 1365-1376.

Lehmann, L., Rousset, F., Roze, D., & Keller, L. (2007). Strong reciprocity or strong ferocity? A population genetic view of the evolution of altruistic punishment. *American Naturalist*, 21-36.

Leibbrandt, A., & López-Pérez, R. (2008). The envious punisher: Understanding third and second party punishment with simple game. *Institute for Empirical Research in Economics, Working Paper No. 373*.

Leibbrandt, A., & López-Pérez, R. (2011). The dark side of altruistic third-party punishment. *Journal of Conflict Resolution, 55*(5), 761-784.

Levati, M. V., Sutter, M., & Van Der Heijden, E. (2007). Leading by example in a public goods experiment with heterogeneity and incomplete information. *Journal of Conflict Resolution, 51*(5), 793-818.

Levine, D. (1998). Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics, 1*, 593-622.

Levine, M., Taylor, P. J., & Best, R. (2011). Third Parties, Violence, and Conflict Resolution. *Psychological Science*.

Levitt, S. D., & List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *The journal of economic perspectives*, 153-174.

Lewis, R. J. (2002). Beyond dominance: the importance of leverage. *The Quarterly review of biology, 77*(2), 149-164.

Li, N. P., Valentine, K. A., & Patel, L. (2011). Mate preferences in the US and Singapore: A cross-cultural test of the mate preference priority model. *Personality and Individual Differences, 50*(2), 291-294.

Little, T. D., Henrich, C. C., Jones, S. M., & Hawley, P. H. (2003). Disentangling the "whys" from the "whats" of aggressive behaviour. *International Journal of Behavioral Development, 27*(2), 122.

Lusher, D., Robins, G., & Kremer, P. (2010). The application of social network analysis to team sports. *Measurement in physical education and exercise science, 14*(4), 211-224.

Lyle, H. F., & Smith, E. (2014). The reputational and social network benefits of prosociality in an Andean community. *Proceedings of the National Academy of Sciences, 111*(13), 4820-4825.

Maestripieri, D. (2012). *Games Primates Play: An Undercover Investigation of the Evolution and Economics of Human Relationships*: Basic Books (AZ).

Maner, J. K., & Mead, N. L. (2010). The essential tension between leadership and power: When leaders sacrifice group goals for the sake of self-interest. *Journal of Personality and Social Psychology, 99*(3), 482.

Manson, J. H., Wrangham, R. W., Boone, J. L., Chapais, B., Dunbar, R., Ember, C. R., . . . Nishida, T. (1991). Intergroup Aggression in Chimpanzees and Humans. *Current Anthropology*, 369-390.

Mappes, T., & Puurtinen, M. (2009). Between-group competition and human cooperation. *Proceedings of the Royal Society B: Biological Sciences, 276*(1655), 355-360. doi: 10.1098/rspb.2008.1060

Marlowe, F. W., Berbesque, J. C., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., . . . Tracer, D. (2008). More 'altruistic' punishment in larger societies. *Proceedings of the Royal Society B: Biological Sciences, 275*(1634), 587-592. doi: 10.1098/rspb.2007.1517

Marlowe, F. W., Berbesque, J. C., Barrett, C., Bolyanatz, A., Gurven, M., & Tracer, D. (2010). The 'spiteful'origins of human cooperation. *Proceedings of the Royal Society B: Biological Sciences, 278*, 2159-2164.

Masclet, D. (2003). Ostracism in work teams: a public good experiment. *International Journal of Manpower, 24*(7), 867-887.

Masclet, D., Noussair, C., Tucker, S., & Villeval, M. (2003). Monetary and nonmonetary punishment in the voluntary contributions mechanism. *The American Economic Review, 93*(1), 366-380.

Masclet, D., & Villeval, M. (2008). Punishment, inequality, and welfare: a public good experiment. *Social Choice and Welfare, 31*(3), 475-502. doi: 10.1007/s00355-007-0291-7

Massen, J. J., van den Berg, L. M., Spruijt, B. M., & Sterck, E. H. (2010). Generous leaders and selfish underdogs: Pro-sociality in despotic macaques. *PLoS ONE, 5*(3), e9734.

Mathew, S., & Boyd, R. (2011). Punishment sustains large-scale cooperation in pre-state warfare. *Proceedings of the National Academy of Sciences*. doi: PNAS 2011 : 1105604108v1-201105604.

Mathew, S., & Boyd, R. (2013). The cost of cowardice: punitive sentiments towards free riders in Turkana raids. *Evolution and Human Behavior, 35*(1), 58–64.

Maynard-Smith, J., Harper, D., & Brookfield, J. (1988). The evolution of aggression: can selection generate variability? *Philosophical Transactions of the Royal Society of London. B, Biological Sciences, 319*(1196), 557-570.

Maynard-Smith, J., & Parker, G. A. (1976). The logic of asymmetric contests. *Animal Behaviour, 24*(1), 159-175.

Maynard-Smith, J., & Price, G. (1973). The logic of animal conflict. *Nature, 246*, 15 - 18. doi: doi:10.1038/246015a0

McCullough, M. E., Kurzban, R., & Tabak, B. A. (2013). Cognitive systems for revenge and forgiveness. *Behavioral and Brain Sciences, 1*(1), 1-15.

McKay, R., Efferson, C., Whitehouse, H., & Fehr, E. (2010). Wrath of God: religious primes and punishment. *Proceedings of the Royal Society B: Biological Sciences, 278*(1713), 1858-1863 doi: 10.1098/rspb.2010.2125

McNamara, J. M., & Houston, A. I. (2002). Credible threats and promises. *Philosophical Transactions of the Royal Society B: Biological Sciences, 357*(1427), 1607-1616. doi: 10.1098/rstb.2002.1069

Mehrabian, A. (1994). *Manual for the revised trait dominance-submissiveness scale (TDS)*: University of California.

Melis, A. P., & Semmann, D. (2010). How is human cooperation different? *Philosophical Transactions of the Royal Society B: Biological Sciences, 365*(1553), 2663-2674. doi: 10.1098/rstb.2010.0157

Miller, K. L. (2010). The Darkest Figure of Crime: Perceptions of Reasons for Male Inmates to Not Report Sexual Assault. *Justice Quarterly, 27*(5), 692-712.

Mitani, J. C., Watts, D. P., & Amsler, S. J. (2010). Lethal intergroup aggression leads to territorial expansion in wild chimpanzees. *Current Biology, 20*(12), 507-508.

Molles, L. E., & Vehrencamp, S. L. (2001). Songbird cheaters pay a retaliation cost: evidence for auditory conventional signals. *Proceedings of the Royal Society B: Biological Sciences, 268*(1480), 2013-2019. doi: 10.1098/rspb.2001.1757

Moshagen, M., Hilbig, B. E., & Musch, J. (2011). Defection in the dark? A randomized-response investigation of cooperativeness in social dilemma games. *European Journal of Social Psychology, 41*(5), 638-644.

Mulder, L. B., Verboon, P., & De Cremer, D. (2009). Sanctions and moral judgments: The moderating effect of sanction severity and trust in authorities. *European Journal of Social Psychology, 39*(2), 255-269. doi: 10.1002/ejsp.506

Nakamaru, M., & Iwasa, Y. (2006). The coevolution of altruism and punishment: role of the selfish punisher. *Journal of Theoretical Biology, 240*(3), 475-488.

Nelissen, R. (2008). The price you pay: cost-dependent reputation effects of altruistic punishment. *Evolution and Human Behavior, 29*(4), 242-248.

Nelissen, R., & Meijers, M. H. (2011). Social benefits of luxury brands as costly signals of wealth and status. *Evolution and Human Behavior, 32*(5), 343-355.

Nikiforakis, N. (2008). Punishment and counter-punishment in public good games: Can we really govern ourselves? *Journal of Public Economics, 92*(1-2), 91-112. doi: 10.1016/j.jpubeco.2007.04.008

Nikiforakis, N., & Normann, H.-T. (2008). A comparative statics analysis of punishment in public good experiments. *Experimental Economics, 11*(4), 358-369.

Nikiforakis, N., Normann, H., & Wallace, B. (2009). Asymmetric enforcement of cooperation in a social dilemma. *Southern Economic Journal, 76*(3), 638-659.

Nikiforakis, N., Noussair, C. N., & Wilkening, T. (2012). Normative conflict and feuds: The limits of self-enforcement. *Journal of Public Economics, 96*(9), 797-807.

Noussair, C. N., & Tan, F. (2011). Voting on punishment systems within a heterogeneous group. *Journal of Public Economic Theory, 13*(5), 661-693.

Nowak, M. A. (2008). Generosity: A winners advice. *Nature, 456*, 579.

Nowak, M. A., Page, K. M., & Sigmund, K. (2000). Fairness versus reason in the Ultimatum Game. *Science, 289*, 1773-1775.

Nowak, M. A., & Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. *Nature, 393*(6685), 573-577.

Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature, 437*(7063), 1291-1298. doi: 10.1038/nature04131

O'Gorman, R., Henrich, J., & Van Vugt, M. (2009). Constraining free riding in public goods games: designated solitary punishers can sustain human cooperation. *Proceedings of the Royal Society B: Biological Sciences, 276*(1655), 323.

O'Gorman, R., Wilson, D., & Miller, R. (2005). Altruistic punishing and helping differ in sensitivity to relatedness, friendship, and future interactions. *Evolution and Human Behavior, 26*(5), 375-387. doi: 10.1016/j.evolhumbehav.2004.12.006

Ohtsubo, Y., Masuda, F., Watanabe, E., & Masuchi, A. (2010). Dishonesty invites costly third-party punishment. *Evolution and Human Behavior, 31*(4), 259-264.

Ohtsuki, H., Iwasa, Y., & Nowak, M. A. (2009). Indirect reciprocity provides only a narrow margin of efficiency for costly punishment. *Nature, 457*(7225), 79-82. doi: 10.1038/nature07601

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

Ostrom, E. (2006). The value-added of laboratory experiments for the study of institutions and common-pool resources. *Journal of Economic Behavior & Organization, 61*(2), 149-163.

Ostrom, E., Walker, J., & Gardner, R. (1992). Covenants With and Without a Sword: Self-Governance is Possible. *The American Political Science Review, 86*(2), 404-417

Osumi, T., & Ohira, H. (2010). The positive side of psychopathy: Emotional detachment in psychopathy and rational decision-making in the ultimatum game. *Personality and Individual Differences, 49*(5), 451-456. doi: 10.1016/j.paid.2010.04.016

Ottone, S. (2008). Are people Samaritans or Avengers. *Economics Bulletin, 3*, 1-8.

Panchanathan, K., & Boyd, R. (2004). Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature, 432*(7016), 499-502.

Patil, I., Cogoni, C., Zangrando, N., Chittaro, L., & Silani, G. (2014). Affective basis of judgment-behavior discrepancy in virtual experiences of moral dilemmas. *Social neuroscience, 9*(1), 94-107.

Paulhus, D. L., & Williams, K. M. (2002). The dark triad of personality: Narcissism, Machiavellianism, and psychopathy. *Journal of Research in Personality, 36*(6), 556-563.

Pawłwski, B., & Dunbar, R. I. (1999). Impact of market value on human mate choice decisions. *Proceedings of the Royal Society of London. Series B: Biological Sciences, 266*(1416), 281-285.

Pedersen, E. J., Kurzban, R., & McCullough, M. E. (2013). Do humans really punish altruistically? A closer look. *Proceedings of the Royal Society B: Biological Sciences, 280*(1758), 1-8.

Penton-Voak, I. S., & Perrett, D. I. (2000). Female preference for male faces changes cyclically:: Further evidence. *Evolution and Human Behavior, 21*(1), 39-48.

Perkins, A. (2014). The full impact of the UK's vote against intervention in Syria has yet to be felt. Retrieved 25/05/2014, 2014, from http://www.theguardian.com/commentisfree/2014/jan/01/uk-syria-vote-impact-parliament

Peter, L., Ottone, S., & Ponzano, F. (2010). Free-riding on altruistic punishment? An experimental comparison of third-party-punishment in a stand-alone and in an in-group environment. *POLIS department's Working Papers*.

Peters, K., & Kashima, Y. (Eds.). (In Press). *Gossiping as Moral Social Action: A Functionalist Account of Gossiper Perceptions*. New York: Psychology Press.

Petersen, M. B., Sell, A., Tooby, J., & Cosmides, L. (2012). To punish or repair? Evolutionary psychology and lay intuitions about modern criminal justice. *Evolution and Human Behavior*.

Peterson, M. B. (2011). *Moralization as a strategy of last resort: Lack of social support predicts moralization of private goods*. Paper presented at the Human Behaviour and Evolution Society, Montpellier, France.

Peterson, M. B. (2012). Moralization as protection against exploitation: do individuals without allies moralize more? *Evolution and Human Behavior, 34*(2), 78–85.

Pietraszewski, D., Cosmides, L., & Tooby, J. (2014). The Content of Our Cooperation, Not the Color of Our Skin: An Alliance Detection System Regulates Categorization by Coalition and Race, but Not Sex. *PLoS ONE, 9*(2), e88534.

Piff, P. K., Stancato, D. M., Côté, S., Mendoza-Denton, R., & Keltner, D. (2012). Higher social class predicts increased unethical behavior. *Proceedings of the National Academy of Sciences, 109*(11), 4086–4091.

Powers, S., & Lehmann, L. (2014). *The transition from leadership to despotism in Neolithic*. Paper presented at the EHBEA, Bristol, UK.

Pratto, F., Tatar, D. G., & Conway-Lanz, S. (1999). Who gets what and why: Determinants of social allocations. *Political Psychology, 20*(1), 127-150.

Price, M. E., Cosmides, L., & Tooby, J. (2002). Punitive sentiment as an anti-free rider psychological device. *Evolution and Human Behavior, 23*(3), 203-231.

Przepiorka, W., & Diekmann, A. (2013). Individual heterogeneity and costly punishment: a volunteer's dilemma. *Proceedings of the Royal Society B: Biological Sciences, 280*(1759), 2013-2247.

Purchese, R. (2014). It's a good day to be an Eve player. Retrieved 01/07, 2014, from http://www.eurogamer.net/articles/2014-01-28-its-a-good-day-to-be-an-eve-player

Raihani, N. J., Grutter, A. S., & Bshary, R. (2010). Punishers benefit from third-party punishment in fish. *Science, 327*(5962), 171-171.

Rand, D. G., Arbesman, S., & Christakis, N. A. (2011). Dynamic social networks promote cooperation in experiments with humans. *Proceedings of the National Academy of Sciences, 108*(8), 19193–19198.

Rand, D. G., Armao, J. J., Nakamaru, M., & Ohtsuki, H. (2010). Anti-social punishment can prevent the co-evolution of punishment and cooperation *Journal of Theoretical Biology, 265*, 624-632.

Rand, D. G., Dreber, A., Ellingsen, T., Fudenberg, D., & Nowak, M. A. (2009). Positive interactions promote public cooperation. *Science, 325*(5945), 1272-1275.

Rand, D. G., Ohtsuki, H., & Nowak, M. A. (2009). Direct reciprocity with costly punishment: generous tit-for-tat prevails. *Journal of Theoretical Biology, 256*(1), 45-57.

Rand, D. G., Tarnita, C. E., Ohtsuki, H., & Nowak, M. A. (2013). Evolution of fairness in the one-shot anonymous Ultimatum Game. *Proceedings of the National Academy of Sciences, 110*(7), 2581–2586.

Rege, M. (2008). Why do people care about social status? *Journal of Economic Behavior & Organization, 66*(2), 233-242.

Reuben, E., & Riedl, A. (2009). Public goods provision and sanctioning in privileged groups. *Journal of Conflict Resolution, 53*(1), 72-93.

Reuben, E., & Riedl, A. (2013). Enforcement of contribution norms in public good games with heterogeneous populations. *Games and Economic Behavior, 77*(1), 122-137.

Reuben, E., & Van Winden, F. (2008). Social ties and coordination on negative reciprocity: The role of affect. *Journal of Public Economics, 92*(1-2), 34-53. doi: 10.1016/j.jpubeco.2007.04.012

Riedl, K., Jensen, K., Call, J., & Tomasello, M. (2012). No third-party punishment in chimpanzees. *Proceedings of the National Academy of Sciences, 109*(37), 14824-14829.

Roberts, G. (1998). Competitive altruism: from reciprocity to the handicap principle. *Proceedings of the Royal Society of London. Series B: Biological Sciences, 265*(1394), 427-431.

Roberts, G. (2013). When Punishment Pays. *PLoS ONE, 8*(3), e57378.

Roberts, S. G. R., & Dunbar, R. I. M. (2010). The costs of family and friends: an 18-month longitudinal study of relationship maintenance and decay. *Evolution and Human Behavior, 32*(3), 186–197.

Rockenbach, B., & Milinski, M. (2006). The efficient interaction of indirect reciprocity and costly punishment. *Nature, 444*(7120), 718-723. doi: 10.1038/nature05229

Rockenbach, B., & Milinski, M. (2011). To qualify as a social partner, humans hide severe punishment, although their observed cooperativeness is decisive. *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.1108996108

Rotter, J. B. (1980). Interpersonal trust, trustworthiness, and gullibility. *American Psychologist, 35*(1), 1-7.

Rustagi, D., Engel, S., & Kosfeld, M. (2010). Conditional cooperation and costly monitoring explain success in forest commons management. *Science, 330*(6006), 961-965.

Sääksvuori, L., Mappes, T., & Puurtinen, M. (2011). Costly punishment prevails in intergroup conflict. *Proceedings of the Royal Society B: Biological Sciences, 278*, 3428-3436.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision-making in the ultimatum game. *Science, 300*(5626), 1755-1758.

Santos, F. C., Pacheco, J. M., & Lenaerts, T. (2006). Cooperation prevails when individuals adjust their social ties. *PLoS Computational Biology, 2*(10), e140.

Santos, M. D., Rankin, D. J., & Wedekind, C. (2010). The evolution of punishment through reputation. *Proceedings of the Royal Society B: Biological Sciences, 278*(1704), 371-377. doi: 10.1098/rspb.2010.1275

Santos, M. D., Rankin, D. J., & Wedekind, C. (2011). The evolution of punishment through reputation. *Proceedings of the Royal Society B: Biological Sciences, 278*(1704), 371-377.

Sapolsky, R. M. (2005). The influence of social hierarchy on primate health. *Science, 308*(5722), 648-652.

Sapouna, M., Wolke, D., Vannini, N., Watson, S., Woods, S., Schneider, W., . . . Aylett, R. (2011). Individual and social network predictors of the short term stability of bullying victimization in the United Kingdom and Germany. *British Journal of Educational Psychology*.

Schino, G. (2001). Grooming, competition and social rank among female primates: a meta-analysis. *Animal Behaviour, 62*(2), 265-271.

Schino, G., & Aureli, F. (2009). Reciprocal Altruism in Primates:: Partner Choice, Cognition, and Emotions. *Advances in the Study of Behavior, 39*, 45-69.

Schoenmakers, S., Hilbe, C., Blasius, B., & Traulsen, A. (2014). Sanctions as honest signals–The evolution of pool punishment by public sanctioning institutions. *Journal of Theoretical Biology, 356*, 36-46.

Scott, J. (2007). *Social network analysis: A handbook*. London: Sage.

Sell, A., Cosmides, L., Tooby, J., Sznycer, D., von Rueden, C., & Gurven, M. (2009). Human adaptations for the visual assessment of strength and fighting ability from the body and face. *Proceedings of the Royal Society B: Biological Sciences, 276*(1656), 575-584.

Sell, A., Tooby, J., & Cosmides, L. (2009). Formidability and the logic of human anger. *Proceedings of the National Academy of Sciences, 106*(35), 15073-15078.

Sell, J., Lovaglia, M. J., Mannix, E. A., Samuelson, C. D., & Wilson, R. K. (2004). Investigating conflict, power, and status within and among groups. *Small Group Research, 35*(1), 44-72.

Sherif, M., Harvey, O. J., White, B. J., Hood, W. R., & Sherif, C. W. (1961). *Intergroup conflict and cooperation: The Robbers Cave experiment* (Vol. 10): University Book Exchange Norman, OK.

Shinada, M., Yamagishi, T., & Ohmura, Y. (2004). False friends are worse than bitter enemies:"Altruistic" punishment of in-group members. *Evolution and Human Behavior, 25*(6), 379-393.

Sidanius, J., & Pratto, F. (2004). *Social Dominance Theory: A New Synthesis*. New York: Psychology Press.

Sigmund, K. (2007). Punish or perish? Retaliation and collaboration among humans. *Trends in Ecology & Evolution, 22*(11), 593-600. doi: 10.1016/j.tree.2007.06.012

Sigmund, K., Hauert, C., & Nowak, M. A. (2001). Reward and punishment. *Proceedings of the National Academy of Sciences of the United States of America, 98*(19), 10757–10762.

Silk, J. B. (2003). Practice random acts of aggression and senseless acts of intimidation: The logic of status contests in social groups. *Evolutionary Anthropology: Issues, News, and Reviews, 11*(6), 221-225. doi: 10.1002/evan.10038

Silk, J. B., & House, B. R. (2011). Evolutionary foundations of human prosocial sentiments. *Proceedings of the National Academy of Sciences, 108*(Supplement 2), 10910–10917.

Singer, T., Seymour, B., O'Doherty, J. P., Stephan, K. E., Dolan, R. J., & Frith, C. D. (2006). Empathic neural responses are modulated by the perceived fairness of others. *Nature, 439*(7075), 466-469. doi: 10.1038/nature04271

Singh, S. (2008). Beware the spinal trap. Retrieved 03/07, 2014, from http://www.theguardian.com/commentisfree/2008/apr/19/controversiesinscience-health

Skyrms, B. (1996). *Evolution of the Social Contract*. Cambridge: University Press.

Skyrms, B., & Pemantle, R. (2000). A dynamic model of social network formation. *Proceedings of the National Academy of Sciences of the United States of America, 97*(16), 9340.

Smith, P. K., Talamelli, L., Cowie, H., Naylor, P., & Chauhan, P. (2004). Profiles of non victims, escaped victims, continuing victims and new victims of school bullying. *British Journal of Educational Psychology, 74*(4), 565-581.

Snyder, J. K., Fessler, D. M., Tiokhin, L., Frederick, D. A., Lee, S. W., & Navarrete, C. D. (2011). Trade-offs in a dangerous world: Women's fear of crime predicts preferences for aggressive and formidable mates. *Evolution and Human Behavior, 32*(2), 127-137.

Sommerfeld, R. D., Krambeck, H. J., Semmann, D., & Milinski, M. (2007). Gossip as an alternative for direct observation in games of indirect reciprocity. *Proceedings of the National Academy of Sciences, 104*(44), 17435–17440.

Stanish, C., & Levine, A. (2011). War and early state formation in the northern Titicaca Basin, Peru. *Proceedings of the National Academy of Sciences, 108*(34), 13901-13906.

Stevens, J. M. G., Vervaecke, H., de Vries, H., & Van Elsacker, L. (2005). The influence of the steepness of dominance hierarchies on reciprocity and interchange in captive groups of bonobos (Pan paniscus). *Behaviour, 142*(7), 941-960.

Stirrat, M., & Perrett, D. I. (2010). Valid facial cues to cooperation and trust: Male facial width and trustworthiness. *Psychological Science, 21*(3), 349-354.

Sylwester, K., & Roberts, G. (2013). Reputation-based partner choice is an effective alternative to indirect reciprocity in solving social dilemmas. *Evolution and Human Behavior, 34*(3), 201–206.

Számadó, S. (2011a). The cost of honesty and the fallacy of the handicap principle. *Animal Behaviour, 81*(1), 3-10. doi: DOI: 10.1016/j.anbehav.2010.08.022

Számadó, S. (2011b). Long-term commitment promotes honest status signalling. *Animal Behaviour*.

Tan, F. (2008). Punishment in a linear public good game with productivity heterogeneity. *De Economist, 156*(3), 269-293.

Tarling, R., & Morris, K. (2010). Reporting crime to the police. *British Journal of Criminology, 50*(3), 474.

Tennie, C. (2012). Punishing for your own good: the case of reputation-based cooperation. *Behavioral and Brain Sciences, 35*(01), 40-41.

Tibbetts, E. A., & Izzo, A. (2010). Social Punishment of Dishonest Signalers Caused by Mismatch between Signal and Behavior. *Current Biology, 20*(18), 1637-1640. doi: 10.1016/j.cub.2010.07.042

Tiedens, L. Z., & Fragale, A. R. (2003). Power Moves: Complementarity in Dominant and Submissive Nonverbal Behavior. *Journal of Personality and Social Psychology, 84*(3), 558-568.

Topalli, V., Wright, R., & Fornango, R. (2002). Drug dealers, robbery and retaliation. Vulnerability, deterrence and the contagion of violence. *British Journal of Criminology, 42*(2), 337-351.

Traulsen, A., Röhl, T., & Milinski, M. (2012). An economic experiment reveals that humans prefer pool punishment to maintain the commons. *Proceedings of the Royal Society B: Biological Sciences, 279*(1743), 3716-3721.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *Quarterly review of biology*, 35-57.

Turchin, P., Currie, T. E., & Whiteshouse, H. (2013). *Understanding the Dynamics of Inequality over the Long Term: A Cultural Evolution Approach*. Paper presented at the Ringberg Meeting 2013, Bavaria.

Turchin, P., & Gavrilets, S. (2009). Evolution of complex hierarchical societies. *Social Evolution and History, 8*(2), 167-198.

Vaillancourt, T., & Hymel, S. (2006). Aggression and social status: the moderating roles of sex and peer valued characteristics. *Aggressive Behavior, 32*(4), 396-408.

Van De Ven, N., Zeelenberg, M., & Pieters, R. (2010). Warding Off the Evil Eye : When the Fear of Being Envied Increases Prosocial Behavior. *Psychological Science, 21*(11), 1671-1677.

Van Vugt, M. (2006). Evolutionary origins of leadership and followership. *Personality and Social Psychology Review, 10*(4), 354-371.

Van Vugt, M., Jepson, S. F., Hart, C. M., & De Cremer, D. (2004). Autocratic leadership in social dilemmas: A threat to group stability. *Journal of Experimental Social Psychology, 40*(1), 1-13.

Van't Wout, M., Kahn, R. S., Sanfey, A. G., & Aleman, A. (2006). Affective state and decision-making in the Ultimatum Game. *Experimental Brain Research, 169*(4), 564-568.

Vehrencamp, S. L. (1983). A model for the evolution of despotic versus egalitarian societies. *Animal Behaviour, 31*(3), 667-682.

Von Rueden, C., Gurven, M., & Kaplan, H. (2008). The multiple dimensions of male social status in an Amazonian society. *Evolution and Human Behavior, 29*(6), 402-415.

Walker, T. (2013). Ethan Couch: Texas quadruple murderer – or a victim of 'affluenza'? Retrieved 29/05/2014, from http://www.independent.co.uk/news/world/americas/ethan-couch-texas-quadruple-murderer--or-a-victim-of-affluenza-9004308.html

Wallace, B., Cesarini, D., Lichtenstein, P., & Johannesson, M. (2007). Heritability of ultimatum game responder behavior. *Proceedings of the National Academy of Sciences, 104*(40), 15631-15634. doi: 10.1073/pnas.0706642104

Wang, J., Suri, S., & Watts, D. J. (2012). Cooperation and assortativity with dynamic partner updating. *Proceedings of the National Academy of Sciences, 109*(36), 14363-14368.

Wasserman, S. (1994). *Social network analysis: Methods and applications*: Cambridge university press.

Watkins, C. D., Fraccaro, P. J., Smith, F. G., Vukovic, J., Feinberg, D. R., DeBruine, L. M., & Jones, B. C. (2010). Taller men are less sensitive to cues of dominance in other men. *Behavioral Ecology, 21*(5), 943-947.

Watkins, C. D., & Jones, B. C. (2012). Priming men with different contest outcomes modulates their dominance perceptions. *Behavioral Ecology, 23*, 539–543.

Watts, D. P. (2002). Reciprocity and interchange in the social relationships of wild male chimpanzees. *Behaviour*, 343-370.

West, S. A., Griffin, A. S., & Gardner, A. (2007). Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. *Journal of Evolutionary Biology, 20*(2), 415-432.

Whiten, A., & Byrne, R. W. (1988). Tactical deception in primates. *Behavioral and Brain Sciences, 11*(02), 233-244.

Widdig, A., Streich, W. J., & Tembrock, G. (2000). Coalition formation among male Barbary macaques(Macaca sylvanus). *American Journal of Primatology, 50*(1), 37-51.

Wilson, D. S., & O'Gorman, R. (2003). Emotions and actions associated with norm-breaking events. *Human Nature, 14*(3), 277-304.

Wilson, E. O. (1980). Sociobiology: the abridged version: Cambridge (MA), Harvard University Press.

Wilson, M. L., & Wrangham, R. W. (2003). Intergroup relations in Chimpanzees. *Annu. Rev. Anthropol, 32*, 363-392.

Winking, J., & Mizer, N. (2013). Natural-field dictator game shows no altruistic giving. *Evolution and Human Behavior, 34*(4), 288–293.

Wolff, I. (2012). Retaliation and the role for punishment in the evolution of cooperation. *Journal of Theoretical Biology, 315*, 128–138.

Wong, M. Y. L., Buston, P. M., Munday, P. L., & Jones, G. P. (2007). The threat of punishment enforces peaceful cooperation and stabilizes queues in a coral-reef fish. *Proceedings of the Royal Society B: Biological Sciences, 274*(1613), 1093-1099. doi: 10.1098/rspb.2006.0284

Yamagishi, T. (1988). The provision of a sanctioning system in the United States and Japan. *Social Psychology Quarterly*, 265-271.

Zahavi, A., & Zahavi, A. (1997). *The Handicap Principle: A Missing Piece of Darwin's Puzzle*. Oxford: University Press.

Zak, P. J., Kurzban, R., Ahmadi, S., Swerdloff, R. S., Park, J., Efremidze, L., . . . Matzner, W. (2009). Testosterone administration decreases generosity in the ultimatum game. *PLoS ONE, 4*(12), e8330.

Zizzo, D. J., & Oswald, A. J. (2001). Are people willing to pay to reduce others income? *Social Interactions and Economic Behavior, 63*, 39-65.

10 Appendix: vignettes and instructions

10.1 Appendix A: chapter 3

10.1.1 Study 1 scenario

Part 1: You have been placed in a student work group of 5 people including yourself to prepare a presentation for a course module. You have also been informed this [*will be your work group for all subsequent group assignments for the rest the year* / *this will be your work group for this project only*]

Tasks needed to complete the presentation have been discussed and divided by the group between its members as equally as possible and the deadline is two weeks from its announcement. The presentation is worth 50% of the marks for that particular module, meaning failing the presentation completely would most likely mean failing the module. However, not long before the presentation needs to be given one member informs the group they have done nothing in the past weeks. Their lack of work had nothing to do with any personal or family problems and they have completed all individual assignments on time. Despite this, the presentation is a success and the group receives a good passing grade.

After session is over, [*you let it be known openly to the course that you will* / *you tell the group that you will / you decide to privately]* approach the tutor inform them that one member of the group did nothing for the presentation and demand the person's mark be reduced.

Part 2: In fact, before you have a chance to do anything another group member, Avery, tells the group they are not happy with the situation, and will inform the tutor about the person who did not contribute to the presentation.

10.1.2 Study 2 scenario

At the beginning of term you entered a university-wide competition to win a two-week alpine skiing holiday and have been informed you were among 20 students to win the prize. Winners will stay in one of four chalets close to one another (5 students to each) and have the opportunity to both participate in excursion and events organised by the university and to explore/travel/ around the area, as well as ski, as you see fit.

You realise you do not know anyone else who won but are nevertheless excited. After attending a meeting concerning the trip you are relieved to hear no one knows one another and realise this will be a good opportunity to meet people outside your course and activity groups

Around a week into the trip there is tension in your cottage as another house-mate, Charlie, is not doing their fair share of the cleaning and household tasks: In fact they repeatedly leave the kitchen a mess and rarely clean up after themselves. One evening you bump into another house-mate, Alex, in the lounge who *[Reveals they too are angry at the way Charlie is behaving and believes you should both confront Charlie about their behaviour next time you are all in the same room / Reveals they don't really mind that Charlie is leaving a mess and that it would create needless tension to bring it up during this trip / Tells you about the fantastic bar/restaurant they came across earlier in the day]*.

A few days later you, Charlie and Alex are in the lounge together and you decide it's time to confront Charlie about his behaviour. You tell him calmly that while it is a holiday you think it's unfair they not only don't help to keep the place clean, but won't even clean up the mess

they themselves create. You look to Alex, who [*tells Charlie they agree with everything you say, and that they think Charlie should do more around the cottage / says nothing and looks away*].

10.2 Appendix B: chapter 4

10.2.1 Study 3 scenario

10.2.1.1 Example of the 'newspaper article' stimuli

BBC Mobile | News | Sport | Weather | iPlayer | TV | Radio | More | Search BBC News

NEWS UK

Home | World | UK | England | N. Ireland | Scotland | Wales | Business | Politics | Health | Education | Sci/Environment | Technology | Entertainment & Arts

8 March 2011 Last updated at 12:16

Local man prevents mugging

A have-a-go hero was today praised by the local community for his bravery in foiling the attempted mugging of an elderly gentleman late on Tuesday evening.

Harold White, aged 77, was walking home alone after visiting a friend and must have seemed a tempting target that night. Having confronted his intended victim, the mugger forced Mr White into an alley way off the main street and demanded he hand over any valuables he had.

John Taylor stopped the mugging in it's tracks

Thankfully help was at hand. A local man, John Taylor (pictured), interrupted the assault, causing the assailant to flee the scene. Speaking to the press, Mr Taylor , who is a 5ft 10, had this to say about the ordeal

"I was walking home after a cinema trip and saw two men arguing ahead of me. Suddenly the younger one wrestled the other into the alley nearby. I knew I had to do something"

"I demanded the guy leave Harold alone, we exchanged *words* and he refused. We struggled and I knocked him to the ground. He got up and ran off, and I called the police"

Police have praised John's bravery in preventing a crime being committed but reiterated they advise the public against such actions and to instead call the police . They continue that while in this case no weapon was involved, Mr Taylor potentially placed himself in danger by becoming directly involved.

The assailant, described as muscular and around 6ft in height, is now being hunted by police with local hospitals and clinics being asked to report anyone arriving with broken nose or similar injury. Anyone with information that may be relevant to the case should contact their local police station.

Top Stories

Gaddafi renews attack on rebels

Fair access levy for universities
UK arrest over Stockholm bombing
Royal couple in Northern Ireland
State pension 'to be simplified'

Features & Analysis

Foot and mouth
Sounds and images of the UK countryside 10 years ago

In pictures
Samba schools dance their way through Rio de Janeiro Carnival

Not amused
The people who just don't care about the royal wedding

Referendum views
Matthew Elliott from the No to AV campaign

Most Popular

Shared | Read | Video/Audio

Net set to make cookies crumble	1
Tiger kills lion in Turkish zoo	2
The royal wedding naysayers	3
State pension 'to be simplified'	4
Police pay review wants perks cut	5
Sheen fired from popular US show	6
Gaddafi renews attack on rebels	7
Royal couple in Northern Ireland	8
Recycled cardboard packaging risk	9
UK arrest over Stockholm bombing	10

10.2.1.2 Third Party Punishment condition

A have-a-go hero was today praised by the local community for his bravery in foiling the attempted mugging of an elderly gentleman late on Tuesday evening. Harold White, aged 77, was walking home alone after visiting a friend and must have seemed a tempting target that night. Having confronted his intended victim, the mugger forced Mr White into an alley way off the main street and demanded he hand over any valuables he had. Thankfully help was at hand. A local man, John Taylor (pictured), interrupted the assault, causing the assailant to flee the scene. Speaking to the press, Mr Taylor, who is 5'10'' had this to say about the ordeal

> *"I was walking home after a cinema trip and saw two men arguing ahead of me. Suddenly the younger one wrestled the other into the alley nearby.* [I box mainly for the exercise, but I can defend myself so / I've never been in a fight in my life but] *I knew I had to do something"*
>
> *"I demanded the guy leave Harold alone, we exchanged words and he refused. We struggled and I knocked him to the ground. He got up and ran off, and I called the police"*

Police have praised John's bravery in preventing a crime being committed but reiterated they advise the public against such actions and to instead call the police. They continue that while in this case no weapon was involved, Mr Taylor potentially placed himself in danger by becoming directly involved.

The assailant, described as muscular and around 6^{ft} in height, is now being hunted by police with local hospitals and clinics being asked to report anyone arriving with broken nose or

similar injury. Anyone with information that may be relevant to the case should contact their local police station.

10.2.1.3 Second party punishment scenario

A local man was praised by the local community today for his bravery after fending off an attempted mugging on Tuesday evening. John Taylor (pictured) was walking home alone after a trip to the cinema with friends and must have seemed a tempting target that night. The unidentified assailant confronted Mr Taylor sometime after 11pm. Mr Taylor had this to say about the ordeal

> *"I was walking home after seeing a film and suddenly this guy stepped out in front of me and demanded I give him my wallet and phone.* [I box mainly for the exercise, but I can defend myself so / I've never been in a fight in my life but] *I wasn't going to just hand over my things"*
>
> *"I demanded the guy leave me alone, we exchanged words and he refused. We struggled and I knocked him to the ground. He got up and ran off and I called the police"*

Police have praised Mr Taylor's bravery in defending himself but reiterated they advise the public against such actions. They continue that while in this case no weapon was involved, Mr Taylor potentially placed himself in danger by escalating the confrontation with his attacker.

The assailant, described as muscular and around 6^{ft} in height, is now being hunted by police with local hospitals and clinics being asked to report anyone arriving with broken nose or

similar injury. Anyone with information that may be relevant to the case should contact their local police station

10.2.1.4 'Bar fight' scenario

Today the clean-up began after a fight broke out early on Saturday evening. Police were called after two men began brawling outside a local bar, breaking tables and knocking over passers-by as they went. Eye witnesses report hearing raised voices from the two people involved before the fight broke out. As yet there is no information as to what caused the altercation and footage from the bar's CCTV cameras does not appear to suggest either man was drunk. By the time police arrived the fight had subsided and one of those involved; Mr John Taylor was arrested. He had this to say

> *"I don't really know what happened. It was early in the afternoon and I got into an argument with this random person. Not even sure what it was about now. The whole thing was stupid"*
>
> *"Not sure who started anything physical.* [I box mainly for the exercise, but I can defend myself so / I've never been in a fight in my life]. *We struggled and I knocked him to the ground. He got up and his friends lead him away"*

Police later released John and are seeking witnesses to help identify the other man involved, described as muscular and around 6^{ft} in height. Local hospitals and clinics being asked to report anyone arriving with broken nose or similar injury and anyone with information that may be relevant should contact their local police station.

10.2.1.5 'Flash mob' scenario

The flash-mob phenomena arrived here yesterday as the city centre was ground to a halt by a recreation of the famous Radiohead music video to the hit signal "Just". For those who don't know, a flash mob is a random collection of individuals who, after signing up to a website, receive instructions guiding them to a specific place at a specific time. Once the signal is given each person carries out a pre-arranged action: from a song and dance number to stripping naked, or in this case lying on the floor. The mob dispersed before any comments from the participants could be obtained. But one onlooker, John Taylor, (Pictured), *[a keen amateur boxer / who was on his way home from a hair salon]* had this to say

> *"It was all a little surreal really. One minute the centre is filled with people the next half are lying on the floor"*
>
> *"I could see some people were confused nervous initially, but it's just a bit of a fun. And I think everyone eventually saw the funny side of it. Certainly made the day more eventful"*

City officials say that while there is nothing illegal in what the mob did, in future they would prefer some notice of future events in case other members of the public are worried by the spectacle and overload the police with calls.

10.2.2 Study 4 scenario

You are sitting alone in a local bar when a group of people arrive together and sit at a table in front of you nearby. The bar is quite full and they soon realise there are more of them than chairs with no opportunity to get more.

You see one of the standing people, a man in a GREY shirt, go over to one of the seated individuals and, un-humorously, forcefully demand the seated man gives up his seat. After a few seconds the seated man grudgingly gets up and stands away.

Another member of the group, a man in a BLUE shirt, notices this taking place and turns to the male in the GREY shirt. He then however says nothing to the man in the GREY shirt before resuming his previous conversation [No intervention condition].

Another member of the group, a man in a BLUE shirt, notices this taking place and turns to the man in the GREY shirt, angrily berating him for making the seated guy move. After a pause, the man in the GREY shirt stares at him and laughs before turning away and starting a conversation with the person next to him [Unsuccessful intervention condition].

Another member of the group, a man in a BLUE shirt, notices this taking place and turns to the man in the GREY shirt, angrily berating him for making the seated guy move. After a pause, the man in the GREY shirt gets up and returns to the other standing people and the other man returns to his seat [Successful intervention condition].

10.3 Appendix C: chapter 5

10.3.1 Study 5 scenario

You have been part of a university sports team for around a year. Following a practice session you and a number of other team members have gone to a local bar. You and the team regularly go to the same bar after practice and the staff always seem happy to have you all there.

Once you arrive, the team claims the one remaining free table. However, as the bar is quite full, there are not enough chairs for everyone. You and a number of others therefore have to stand.

Nearby, two strangers are sitting at another table and after a few minutes one of them heads to the bar to order drinks. Seeing this, one of standing members of your team goes over to the table and proceeds to take the now vacant chair. They laugh off the protests of the other stranger, daring them to take it back, and return to your team's table with the chair.

You see that another member of your team, who is [*popular/unpopular*] in the team and the [*most/least]* skilled member, has also noticed this interaction. Standing up, this person angrily berates the chair-taker, and tells them they "won't tolerate this behaviour" and will "[*beat the crap out of them / make sure they never represent the team competitively again*]" if the chair isn't returned and should they ever do anything like this in the future.

10.3.2 Study 6 scenario

You have been part of a local sports team for around a year. Following a practice session you and a number of other team members have gone to a local bar. You and the team regularly go to the same bar after practice and the staff always seem happy to have you all there.

Once you arrive, the team claims the one remaining free table. However, as the bar is quite full, there are not enough chairs for everyone. You and a number of others therefore have to stand. Nearby, two strangers are sitting at another table and after a few minutes one of them heads to the bar to order drinks. Seeing this, one of standing members of your team, who is [*popular/unpopular*] in the team and considered to be one of the [*most/least]* skilled member, goes over to the table and proceeds to take the now vacant chair. They laugh off the protests of the other stranger, daring them to take it back, and return to your team's table with the chair.

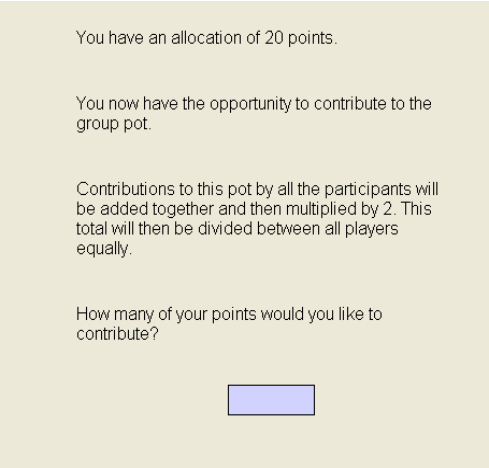
As the ‘chair-taker’ returns, you see that another member of your team, someone who is quite is [also] [*popular/unpopular*] in the team and [also] one of the [*most/least]* skilled members, has also noticed this interaction. They seem to be very agitated by it.

10.4 Appendix D: chapter 6

10.4.1 Study 7 participant instructions and comprehension questions

10.4.1.1 Contribution round instructions

In each round in stage 1, each participant will receive **20 points** and will have the opportunity to contribute between 0-20 of these points to the **group pot** (see screen below). Any points you do not contribute will be kept by you



Once all participants have made their contribution decisions, they will be added together and the total will be multiplied by 2. This total will then be **divided equally** between all participants. (see screen below)

"You" contributed 8 points, Player 1 contributed 15 points, Player 2 contributed 5 points and Player 3 contributed 14 points. 8+15+5+14=42, 42*2=84, 84/4=21, so each participant received 21points from the group pot. As "you" kept 12 points for yourself, your total for that round is 33 (12+21=33).

The round is now over. Any points you earned will be added to your overall total and are 'safe'. Before the next round begins, the groups will be randomised again. You will play between 5 and 10 rounds. Beyond Round 5 there is a 75% chance the game will continue to the next round.

Using the information provided, please answer the questions below to ensure your have understood the instructions

1) How many points are you allocated at the beginning of each round _____
2) By how much is the Group Pot multiplied by before being divided _____
3)

	Contributions
Player 1:	7
Player 2:	14
Player 3:	3
Player 4:	16

What is the total pot before it is multiplied _____

The total pot after it is multiplied _____

How many points will each player receive from the pot? _____

What is Player 4's total score for this round? _____

What is Player 1's total score for this round? _____

10.4.1.2 Punishment round instructions

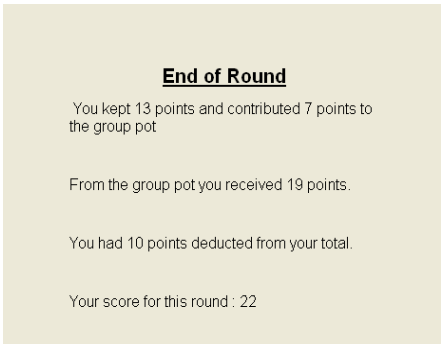
In stage two, the rounds will proceed as in stage 1. In addition however, a random member of the group will be able to assign 'deduction' points to other group's members. After the contribution part of the round, **one randomly selected** group member will see the screen below

[As the player with the ability to assign deduction points, you will also gain an additional [50% / 25% / 10%] of the group pot (see screen below). For instance, here the total pot is 93 points, so in addition to your share of 23 points, you will get an additional [46/23/9] points. ***These points are NOT taken from any other player****]*

If selected you will be asked to assign between 0-20 deduction points to one player if you so wish (assigning 0 if you do not wish to assign any).

Each point assigned to the selected player will remove **one** of their points

Each point you assign as a deduction point will remove **one** of your points



Once this decision has been made the round ends and the player who had deduction points assigned to them will be told how many points they had deducted (as above). The round is now over. As with Stage 1, the groups will be randomised before the next round to begins.

You will play between a 5 and 10 rounds. Beyond Round 5, there is a 75% chance the game will continue to the next round. Using the information provided, please answer the questions below to ensure your have understood the instructions.

1) What is the maximum number of deduction points you can assign to another _____
2) For deduction point you assign, how many points are removed from the player you assign them to? _____
3) For deduction point you assign, how many points are removed from your score for that round? _____
4)

	Contribution	Total Score
Player 1:	7	28
Player 2:	12	33
Player 3:	15	36
You:	9	30

If you assign Player 1 FIVE deduction points what will their new Total Score be? _____

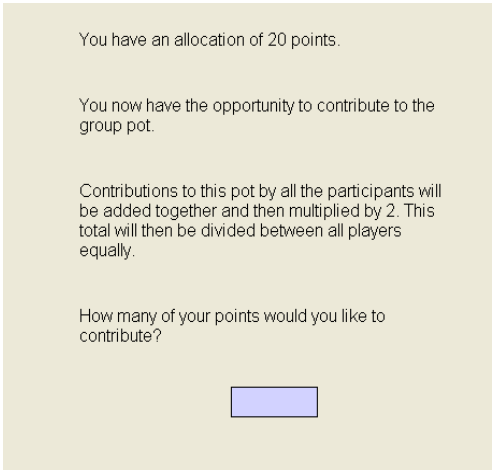
If you assign Player 2 SEVEN deduction points, what will your Total Score be? _____

The total group pot was 86. As the deducting player, how many additional points will you received from the group pot? _____

10.4.2 Study 8: participant instructions and comprehension questions

10.4.2.1 Contribution round

At the start of each round each participant will receive **20 points** and will have the opportunity to contribute between 0-20 of these points to the **group pot** (see screen to the right). Any points you do not contribute will be kept by you



Once all participants have made their group pot contribution decisions, they will be added together and the total will be multiplied by 2. *[25% of this total will then be removed (more on this in a moment)]*. This [the remaining] total will then be **divided equally** between all participants (see screen bottom/right)

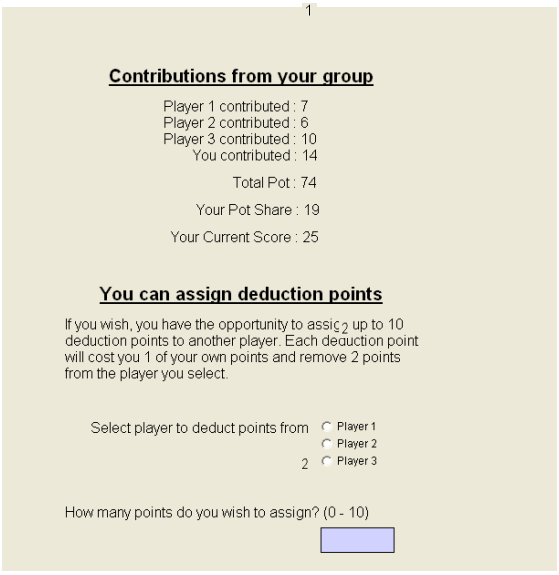
"You" contributed 8 points, Player 1 contributed 15 points, Player 2 contributed 5 points and Player 3 contributed 14 points. 8+15+5+14=42, 42*2=84, 84/4=21, so each participant received 21points from the group pot. As "you" kept 12 points for yourself, your total for that round is 33 (12+21=33).

PLEASE NOTE that the 'player 1' etc labels are

arbitrarily assigned at this screen by the computer. That is to say, in the next round, anyone could be 'player 1' etc

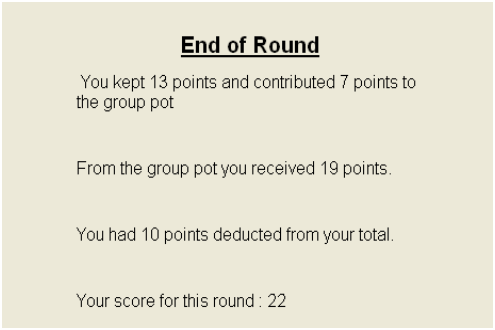
10.4.2.2 Punishment round instructions

For three of the four players in each group, the round ends at the previous screen. However, one player will be able to assign 'deduction' points to other group's members. This randomly selected group member will see the following screen (right)



[If selected, you will be asked if you would like to assign up to 20 'deduction points' to one player (assigning 0 if you do not wish to assign any)]

[If you are the player with the ability to assign deduction points, you will also gain an additional 25% of the group pot (see screen to the right). For instance, here the total pot is 92 points, so in addition to your share of 23 points, you will get an additional 23 points. **These points are NOT taken from any other player]**



[If you are the player with the ability to assign deduction points, you will also be given the 25% of the group pot that was previously removed (see screen to the right). So in addition to your share of the group pot (15), the points you kept (10), you will get an additional 20 points, giving you 45 points]

Each deduction point assigned to the selected player will **deduct** one of their points

Each point you assign as a deduction point will 'cost' you one of your points

Once this decision has been made the round ends. The player who had deduction points assigned to them will be told how many points they had deducted but **NOT** who assigned them (right). The round is now over and any points you have are 'safe' and added to your total.

Note, the person selected to have the ability to assign deduction points [WILL KEEP THIS ROLE for the duration of the game / IS RANDOMLY ASSIGNED each round]

You will play between 6 and 12 rounds. Beyond Round 6, there is a 75% chance the game will continue to the next round. ***[In every subsequent round you will be in a group with the same individuals / Before the start of each new round, the groups will again be randomly assigned****, so it is unlikely you will play against the same group of individuals twice]*

Using the information provided, please answer the questions below to ensure your have understood the instructions. Remember you will also play a few practice rounds to get used to how the game is played

1) How many points are you allocated at the beginning of each round _____
2) By how much is the Group Pot multiplied by before being divided _____ equally between all players
3) What is the maximum number of deduction points you can assign _____ to another player
4) If you assigned 4 deduction points to another player, how many _____ points would be deducted from that player?
5) If you assigned 9 deduction points to another player, how many _____ points would be removed from your own score?

6) In this game, do you (circle your response)
    a. Play with the same players each round, or
    b. Play with random players each round

7) The player randomly assigned to be able to allocate deduction points (circle your response)...
    a. Keeps this role for the whole game, or
    b. The role is randomly assigned at the start of the next round

Please also indicate:

Age: _____ Sex: _____ Nationality: _____

10.5 Appendix E

10.5.1 Social network questionnaire – Study 9 & 10 used identical questionnaires

You will now be asked some questions about your relationship with the other students on this field course.

The questions are designed to reflect your opinion, so there are no right or wrong answers. To maintain participant anonymity and confidentiality, once data collection is complete, the list of names and numbers will be removed. Thus, as stated above, once data collection is complete, there will be no way of identifying you or any other participant. Please answer as honestly as you can.

For example,

Question: "Which people have you never met before today"

Answer: 1, 24, 17

Once again, please be assured all information will be kept anonymous and confidential

1) What is your number on the sheet _____
2) Who would you say are your closest friends on this field course? _____
3) Who do you think are the most influential members of this field course _____
4) Who did you socialise most with during this trip? _____
5) If you were going for an average evening out, who would you be going with from this field trip? _____

10.5.2 Study 9 participant instructions

During this game you will fill out information for both the Proposer and Third-party roles and will be randomly matched into groups of three once your information has been collected.

For example, your Proposer data will be matched to the Third-Party decisions of another and to a Receiver, your Third-party data will be matched to a Receiver and a Proposer, and you will be the receiver for the Proposer and Third-Party data of others.

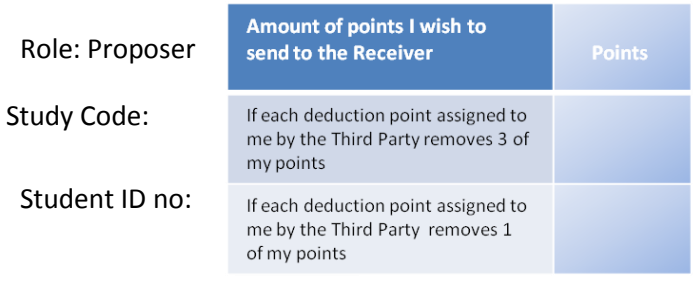
You will be given a score card for both the Proposer and Third-Party role. As well as your decisions, please ensure you write the last 4 digits of your student ID number (or surname if you can't remember). This information is for experimenter use only and **you cannot be identified by other participants.**

Next to where it says "round number" please indicate the order in which you completed the Proposer and Third-Party score cards: Write 1 on the card you did first, and 2 on the card you did second.

Once you have filled out the first card it will be collected in and the second will be handed out. Once this is done the experiment is over and you may collect your thank you chocolate

The Proposer role

As the Proposer, you have an allocation of **20 points**. You must decide how many points (if any) to send to an anonymous Receiver. You may choose to send any whole number amount between 0 (no points sent to the receiver) and 10 points.



Any points you do not send will be yours and added to your total earnings for the experiment

As mentioned above, a Third-Party will randomly assigned to your decision (See next page). They will have the opportunity to assign 'deduction point's to you. Each deduction point they assign you will **either remove THREE points from your total or ONE point.** We will ask you to make you 'send' decision for both potential deduction schemes (see score card for clarity).

Whether your group uses the THREE points or ONE point deduction scheme will be decided at random.

The Receiver role

The decisions you make as a Proposer will determine how many points the Receiver assigned to you will receive. Equally, what you receive as a 'Receiver' will be affected by the Proposal decisions of others.

You will not have anything to do for this role.

The Third-Party

When making a decision as the Third-Party, you will have an **allocation of 10 points**.

You have the choice assign the Proposer 'deduction points'. Each deduction point you assign will 'cost' you one of your points.

Any of the 10 points you don't 'spend' on deduction points will be kept by you

Each deduction point you assign to the Proposer will **either remove THREE of their points OR ONE of their points**. We will ask you to make decisions for both deduction schemes. As

mentioned above, whether your group uses the THREE points or ONE point deduction scheme will be decided at random.

These decisions will affect the Proposer's points only.

As you do not know how many (if any) points the Proposer will send, you will be asked to enter how many deduction points you would allocate **in response** to a given number of points sent by the Proposer. For each possible amount if points sent by the proposer you can assign between 0-10 deduction points (See example below).

Think of it like this, you are indicating that:

If the Proposer sent 3 points to the Receiver, you would wish to assign them ____ deduction points in response. However if the Proposer sent 10 points to the Receiver, you would wish to assign them _____ deduction points etc

Study Code:_____

Role: Third-Party

Student ID no:

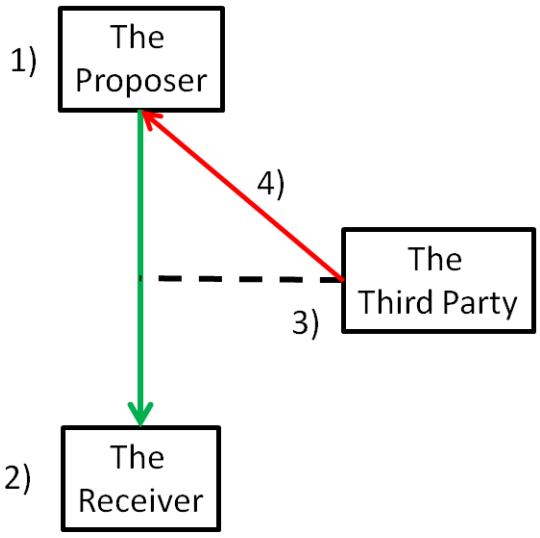
Amount of points sent by the Proposer		1	2	3	4	5	6	7	8	9	10
Amount of deduction points I would wish to assign in response	If each deduction point removed **THREE** of the Proposers points										
	If each deduction point removed **ONE** of the Proposers points										

10.5.3 Study 10 participant instructions

You are about to take part in a "third party interaction game"

Normally, this game involves groups of three with each person taking a different role. One person takes the role of The Proposer, one person takes the role of The Receiver, and one person takes the role of The Third Party

The game goes as follows



1) The Proposer: The Proposer is given an allocation of 20 points. They can choose, if they wish, to send between 0-10 of these points to the Receiver. **The Proposer keeps any points they do not send**.

2) The Receiver: The Receiver plays no active role in the game.

3) The Third Party: The Third Party is given an allocation of 10 points and observes how many points, if any, the Proposer sends to the Receiver. If they so choose, the Third Party can assign up to 10 deduction points to the Proposer (4). Each deduction point 'costs' the Third Party 1 of their points, and removes 3 of the Proposer's total. For example 2 deduction points would remove 6 points from the Proposer's total, and leave the Third Party with 8 points. **Any points the Third Party does not assign are kept by them**

So, the number of points (and therefore £) each person gets is based on the following

a) The Proposer: The amount of points they kept, minus any deductions by the Third Party
b) The Receiver: The amount of points they received from the Proposer
c) The Third Party: The amount of points they did not spend on deduction points

Because we're a small group, the game you will be playing will be a little different; because here you will be making decision as the Proposer AND the Third Party.

1) Proposer: You have 30 points and can send between 0-15 points to the Receiver

How many points do you wish to send to the Receiver?	Points

Today

Normally this sort of game is done at a computer, but because that's not possible, you will all make Proposer and Third Party decisions at the same time.

Once the cards (Right and below) have been taken in, these proposals will be randomly matched to someone else in the group

You will play a version of the Third Party interaction game using what is known as the "Strategy" Method. Essentially, you will make decisions in all roles.

In a moment you will be given a score card like the ones shown above and to the right, which asks you to make both Proposer and Third Party decisions.

Once you have made the decisions, these will be randomly matched by computer after the experiment.

2) Third Party: You have been given an allocation of 10 points. You may assign between 0-10 of these points to the Proposer as 'deduction' points. Remember, you can assign between 0-10 points for EACH option below

Amount of points SENT by the Proposer TO the Receiver	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
How many deduction points do you wish to assign in response?																

So, the total amount of points you can earn during the experiment depends on

a) The points you kept when making your Proposer decision, minus any deduction points you were assigned by a Third Party
b) The points sent to you BY a Proposer you were matched with
c) The points you did not spend as deduction points AS the Third Party, once your decisions were matched to a Proposer

There's more!

Because it takes time to match everyone together and work out everyone's points, those chosen to receive their points as £ will be told after dinner tomorrow.

When you receive your money, you WILL BE TOLD who your Proposer was and how much they sent you, and who your Third Party was and how many deduction points they assigned to you. The decisions will also go up on the classroom whiteboard.

THIS MEANS YOUR DECISIONS ON THE SCORE CARDS ARE NOT ANONYMOUS