Exploring the impact of neurocriminological ideas on public sentencing preferences and moral behaviour

Thesis for DPhil Criminology

94,110 words

Michaelmas Term 2018

Robert Blakey

Centre for Criminology, Faculty of Law

St Hilda's College

Abstract

While the focus of criminology has traditionally been the sociological correlates of criminal behaviour, there is growing interest in the neurobiological correlates of offending. Theorists predict that this emerging area of research, known as neurocriminology, will change how people view crime, punishment and their own behaviour. The veracity of this prediction will rest on the question of whether neurocriminology can challenge the belief that offenders make a morally relevant *choice* to offend (or not). By suggesting that offenders are less able to make such choices, neurocriminology could lead offenders to appear more dangerous yet less blameworthy. Indeed, learning about neurocriminology may actually promote deviant behaviour, if people internalise the idea that they are not really responsible for their actions, because their brain 'made them do it'. This thesis tested such hypotheses through two online survey experiments and three field experiments, with a total of 8153 participants. In every study, participants learned about some neurocriminological idea that could explain why a person, without any mental disorder, committed an impulsive offence – a common combination of factors in many individual acts of crime. The communication of neurocriminology significantly reduced attributions of blame, albeit to a small extent, and had no (or even an ameliorating) effect on perceived dangerousness. In the largest study, after exposure to neurocriminology, participants punished the offender significantly less, a response to ascribing him a significantly more moral character and significantly less self-control; though these effects remained small. Neurocriminological evidence also led participants to recall significantly more acts of prior deviance, but not to act less morally in the study. The fear that neurocriminology could promote deviance might explain the final finding: participants in the richest and poorest boroughs of London – but not those in between – were significantly more inclined to disrupt a scientific campaign aimed at communicating neurocriminology to offenders (rather than non-offenders). In sum, the implications of communicating neurocriminology were in the direction of promoting lenience for offenders, yet the effects were small. Hence this thesis concludes that lay people will not relinquish the notion that offenders *deserve* to be punished in the face of neurocriminology. Given its function in communicating censure, the human tendency to attribute blame may be too ingrained to be overthrown by science, at least at the present point in time.

Table of Contents (Shortened Version)

Table of Cases 12

Chapter 1: Introduction 13

Chapter 2: Overview of studies 19

Chapter 3: The experimental method 33

Chapter 4: Attitudes towards crime and punishment 41

Chapter 5: The impact of neurocriminological ideas on attributions of blame and the unconscious: A conceptual framework 71

Chapter 6: The impact of neurocriminological ideas on attributions of blame and the unconscious: The empirical work 94

The theatre study 103

The online study 118

Chapter 7: The impact of neurocriminological ideas on attributions of blame and the 'real self': A conceptual framework 145

Chapter 8: The impact of neurocriminological ideas on attributions of blame and the 'real self': The empirical work 178

The museum study 181

Chapter 9: The impact of neurocriminological ideas on attributions of dangerousness 219

The theatre study 231

The online study 235

The museum study 238

Chapter 10: The impact of neurocriminological ideas on moral behaviour 246

The Facebook study 247

The museum study 268

Chapter 11: Limits on the impact of neurocriminological ideas 274

The museum study 287

The postcard study 289

Chapter 12: Conclusion 308

Acknowledgements 322

References 323

Glossary 347

Appendices 350

Tables of Contents (Full Version)

Table of Cases 12

Chapter 1: Introduction 13

Chapter 2: Overview of studies 19

i. Linking the science to the offender 19

ii. The psychological approach 23

The theatre study (the young offender) 23

iii. The mock court approach 24

The online study (the rational and the irrational offender) 25

The postcard study (the act of communicating neurocriminology) 26

The Facebook study (the psychopathic brain) 26

iv. The philosophical approach 27

The museum study (the caused offender) 27

v. Thesis structure 28

vi. Conclusion 31

Chapter 3: The experimental method 33

i. The experimental method 33

i. Vignette design 36

ii. Conclusion 39

Chapter 4: Attitudes towards crime and punishment 41

i. How do neurocriminologists explain crime? 42

Philosophical assumptions of neurocriminology 42

The assumption of determinism 43

The assumption of the conscious brain 45

The assumption of the social brain 46

Empirical findings of neurocriminology 47

The role of the brain in psychopathy and offending 49

The role of the brain in self-control and offending 50

The role of genes and the environment in offending 51

ii. How do neurocriminologists justify punishment? 57

iii. How do lay people explain crime? 64

iv. How do lay people justify punishment? 66

v. Conclusion 69

Chapter 5: The impact of neurocriminological ideas on attributions of blame and the unconscious: A conceptual framework 71

i. Will neuroscience increase belief in determinism? 73
ii. When do people blame the caused offender? 75
iii. How might neuroscience reduce attributions of blame? 80
Mechanism one: The unconscious offender 82
Mechanism two: The consciously helpless offender 83
Mechanism three: The unconsciously caused offender 85
Mechanism four: The abnormally caused offender 86
iv. Conclusion 92
Chapter 6: The impact of neurocriminological ideas on attributions of blame and the unconscious: The empirical work 94
i. The psychological approach 94
ii. The mock court approach 96
Effects on verdicts 97
Effects on sentencing 98
iii. Beyond criminal justice 99
iv. Genetic and philosophical testimony 100
v. The role of neuroscience in youth justice 102
vi. The theatre study 103
Hypotheses 103
Method 104
Design 104
Participants 105
The play 105
Procedure 106
Survey items 107
Results 108
The recommended age of criminal responsibility 108
Attributions of moral responsibility 108
Summary 112
Discussion 112
vii. The online study 118
Is the impulsive actor unable or unwilling to exercise self-control? 119
The neuroscientific response 120
The legal response 120
The intuitive response in theory 120

The intuitive response in practice 122
Is a lapse from high self-control willed or caused? 124
The neuroscientific response 124
The legal response 124
The intuitive response 125
Hypotheses 126
Method 127
Design 127
Participants 128
Procedure 128
Results 131
Effects of the disposition 131
Causal attributions 132
Mediation 133
Effects of the Source 134
Fault attributed for the Impulsive Disposition 134
The control condition 135
Summary 135
Discussion 135
How did people explain the offending of the impulsive and controlled actors? 136
Is a lapse from high self-control willed or caused? 139
Is the impulsive actor unable or unwilling to exercise self-control? 139
How do lay people represent the brain? 142
viii. Conclusion 143
Chapter 7: The impact of neurocriminological ideas on attributions of blame and the 'real self': A conceptual framework 145
i. Do people blame the offender whose brain has been manipulated? 147
Do people conceive of the brain as the 'real self'? 151
Do people consider situational influences obstructions to the expression of the 'real self'? 152
Do people consider a manipulated brain an obstruction to the expression of the 'real self'? 154
Do people consider an abnormal brain an obstruction to the expression of the 'real self'? 158

Do people consider the situational brain an obstruction to the expression of the 'real self'? 164

ii. Some potentially confounding correlates of perceptions of the 'real self' 168

iii. The Deep Self Model 171

Do people consider genetic and environmental influences to constitute the Deep Self? 172

Do people consider situational brain activity to constitute the Acting Self? 173

iv. Conclusion 176

Chapter 8: The impact of neurocriminological ideas on attributions of blame and the 'real self': The empirical work 178

i. Two untested mechanisms of neuroscientific influence 179

Will there be a gap between the lay and legal response to neurocriminology? 180

ii. The museum study 181

Hypotheses 184

Method 184

Design 184

Participants 185

Procedure 186

The follow-up survey 191

Results 192

The hypothetical world 192

Mediation analysis 193

The real world 194

Mediation analysis 195

Summary 197

Discussion 197

How did neuroscience affect moral perceptions of the person? 200

How did neuroscience affect attributions of self-control? 201

How limited were the effects of neuroscience? 203

What were the effects of sequentially presenting the hypothetical and real worlds? 207

The current study 209

Hypotheses 210

Results 210

The follow-up survey 214
Summary 215
Discussion 215
Were there facilitatory effects of the hypothetical world? 216
iii. Conclusion 218
Chapter 9: The impact of neurocriminological ideas on attributions of dangerousness 219
i. How might neuroscience aggravate sentences? 220
ii. The clinical approach 223
iii. The mock court approach 224
iv. The criminal context 228
v. The theatre study 231
Hypothesis 231
Results 232
Discussion 233
vi. The online study 235
Hypotheses 235
Results 236
The control condition 236
Discussion 237
vii. The museum study 238
Hypothesis 239
Results 240
Discussion 240
viii. Conclusion 243
Chapter 10: The impact of neurocriminological ideas on moral behaviour 246
i. The Facebook study 247
Moral alarm 247
How might neuroscience promote immoral behaviour? 248
The behavioural consequences of belief in free will 249
The behavioural consequences of belief in a just world 250
The behavioural consequences of belief in fatalism 250
The believability of neuroscience 252
Facebook analysis 253
Hypotheses 254

Method 254
Design 254
Participants 255
Procedure 255
Results 263
Summary 264
Discussion 265
ii. The museum study 268
Results 269
Discussion 269
Practical implications 270
iii. Conclusion 272
Chapter 11: Limits on the impact of neurocriminological ideas 274
i. Does positive retributivism limit the influence of neurocriminology? 276
ii. Do consequentialist concerns limit the influence of neurocriminology? 279
Does the wrongfulness of crime motivate blame? 279
Can censure be maintained without blame? 281
What is the consequentialist function of blame? 282
How does blame fulfil this consequentialist function? 284
iii. The museum study 287
Results 287
Discussion 288
iv. The postcard study 289
The lost letter technique 290
Returning the postcard: Communicating neurocriminology to offenders 290
Returning the postcard: A moral act 293
Hypotheses 295
Method 295
Design 295
Participants 296
Postcard content 296
London boroughs 297
Procedure 298

Results....298
Summary....302
Discussion....302
v. Conclusion....306
Chapter 12: Conclusion....308
i. Do neurocriminological ideas reduce attributions of blame?....311
A blameworthy alternative to the free self in a neuroscientific world..312
ii. Do neurocriminological ideas increase attributions of dangerousness?316
iii. Do neurocriminological ideas increase immoral behaviour?....317
iv. Do people seek to censor neurocriminological ideas from offenders?.317
v. Future research....318
vi. Final note....319
Acknowledgements....322
References....323
Glossary....347
Appendices....350
i. Appendix A: Experimental materials and variables for the theatre study 350
The adolescent brain....350
The prefrontal cortex....350
The limbic system....351
Example envelope....352
Example survey....353
The full script for Brainstorm....354
ii. Appendix B: Experimental materials and variables for the online study 377
The survey questions....379
iii. Appendix C: Experimental materials and variables for the museum study 385
The offence....385
The science....386
Every condition....386
The biosocial condition....386
The neurobiological condition....386
Every condition....387
The biosocial condition....387

The neurobiological condition 387
Follow-up to the museum study 393
iv. Appendix D: Experimental materials and variables for the Facebook study 393
Psychopathy 393
Free will 394
Determinism 394
Dualism 395
The survey questions 396
The crying baby 396
The trolley and the switch 396
The trolley and the bridge 397
v. Appendix E: Experimental materials and variables for the postcard study 402
Example postcards 402

Table of Cases

Roper v. Simmons (2005) 543 U.S. 551.....86

Graham v. Florida (2010) 560 U.S. 48.....86

Miller v. Alabama (2012) 567 U.S. __.....86

Chapter 1: Introduction

It has been claimed that brain science is capable of 'radically transforming our approach to criminal justice' (Greene and Cohen, 2004, p.1784). This transformative potential of neuroscience primarily regards not the implications of the science for the law as it stands, but the implications of the science for public support of the law and legal practice. While there has been a surge in legal commentaries on the implications of neuroscience for the content of the law (Jones et al., 2013), far less attention has been directed towards the criminological question of how lay people will respond to neuroscientific evidence. In particular, such evidence might change sentencing preferences, support for the criminal justice system (CJS) and the propensity for moral deviance.

Whereas neuroscience is the broader study of the brain, neurocriminology is the study of how the volume and activity of different brain regions correlate with criminal behaviours in particular (neuroscientists refer to the volume and activity of brain regions as the structure and function of those regions). While some neuroscientists specifically identify as neurocriminologists (e.g., Raine, 2013), this thesis uses the term 'neurocriminologist' to refer to all academics who publish (empirical or library) research on the relationship between the brain and criminal behaviour. Similarly, the term 'neuroscience' is used in its broadest sense to refer to any information about the brain. A glossary can be found at the end of the thesis, defining all technical terms.

The advance of neurocriminology is likely to open up profound normative questions concerning what institutions of law and punishment can and should seek to achieve. For example, some researchers have used neuroscience to support the philosophy of determinism, the idea that all behaviour is the inevitable product of prior uncontrollable events (Harris, 2012; Libet, 1985). If neuroscience does support

determinism, one might question whether the CJS should continue to pursue positive retributivism; that is punish offenders on the basis that they are blameworthy for 'choosing' to offend. To be clear from the outset, I am a determinist who believes that determinism challenges the positive retributivism. However, this thesis does not aim to document evidence for this view; I do not aim to discuss the normative questions of whether neuroscience can support determinism, whether determinism is true or whether the truth of determinism *should* bring positive retributivism to an end. These questions do not need to be answered, in order to investigate the conclusions that lay people draw from neuroscience – the focus of the current research.

The conclusions that people draw from neuroscience about the appropriateness of positive retributivism pose important implications for the CJS. First, neuroscience could destabilise public support for positive retributivism. Greene and Cohen (2004, p.1776) expect the science to expose a 'gap between what the law officially cares about and what people really care about', arguing that only lay people (and not current interpretations of criminal law) will find neuroscience a threat for positive retributivism. Since 'a broadly retributive theory of punishment…remains in the ascendency in the Anglo-American world' (Matravers, 2016, p.145), one possible implication of neuroscience is that the CJS loses legitimacy in the eyes of a neuroscientifically informed public. Second, neuroscience may shift public support towards alternative justifications for punishment; for example, towards policies that prolong the detention of offenders until they are no longer 'dangerous', thereby lengthening sentences (Snead, 2007). Hence it is unclear whether the conclusions people draw from neurocriminology will ultimately reduce or increase public punitiveness (Aspinwall, Brown and Tabery, 2012).

This thesis is concerned with what lay people will seek to achieve through punishment after learning about neuroscience; in particular, neuroscience that can explain traits, behaviours or tendencies implicated in criminal acts. The addressed research questions concern whether, how and why communicating neuroscience to the public could change their attitudes towards crime and punishment. Attitudes towards crime and punishment are defined broadly as beliefs that entail the evaluation of crime or punishment. This includes beliefs that entail evaluations of a particular offender, such as attributions of choice, blame and dangerousness, and beliefs that entail evaluations of punishment, such as the prison sentence recommended for a particular offender. The definition also includes beliefs that reveal more implicit evaluations of moral decisions, such as the willingness of people to cheat, resist impulse and reason in favour of the greater good. The willingness of people to engage in such behaviours implicitly reveals the perceived appropriateness of tendencies that have been linked to criminal acts (Hare, 1993; Hirschi, 1969).

The empirical research conducted for this thesis captured the attitudes of convenience samples; those are people who were conveniently available to participate in the research at the time it was conducted. The samples were recruited through a theatre, a museum and online recruitment channels. Although no efforts were made to exclude lawyers or neuroscientists from the sample, it was generally the case that participants had no expertise in law or neuroscience. The focus on the perspectives of lay people is justified by the importance of favourable public opinion for maintaining the legitimacy of the CJS (Robinson and Darley, 1997). The focus is also justified by the fact that legally untrained people act as jurors and almost all judges are untrained in neuroscience. Hence, although it remains a task for future research, one might expect the findings of this thesis to generalise to judges, police officers, victims and offenders.

This thesis begins in chapter two with an overview of all five studies, the key research questions and methodological approach adopted, and the theoretical relationships between the aims of each study. Chapter three evaluates the methodological approach common to all five studies; that being the experimental approach, which aims to establish the causal effect of learning about a neurocriminological idea on sentencing preferences and moral behaviour. The thesis aims to capture the changes in attitudes that follow exposure to neurobiological explanations of criminal behaviour and related tendencies. In order to make predictions about such changes, one must first consider the attitudes that neurocriminologists and lay people currently hold towards crime and punishment. Hence chapter four discusses how crime is explained and how punishment is justified, first, by neurocriminologists and second, by lay people. By comparing the attitudes of lay people with the attitudes of neurocriminologists, one can make predictions about how the two perspectives might collide.

For specific neurocriminologists, there is a clear relationship between the neurobiological causes of offending and the most appropriate justification for punishment: since criminal behaviour is caused by the brain and offenders are not morally responsible for the condition of their brains, offenders do not *deserve* punishment; in turn, punishment is only justified if it reduces the likelihood of further offending (Greene and Cohen, 2004). One might also expect this relationship to hold for the lay person: the adopted explanation of offending might inform the recommended form of punishment (Carroll et al., 1987). Hence one would expect the recommended punishment to change with attributions for offending. This rationale provides the basis for predicting changes in lay attitudes following exposure to neuroscience.

The empirical content of this thesis regards the effects of posing scientific challenges to lay explanations of criminal behaviour. The succeeding chapters address these challenges in the following manner. If neuroscientific evidence successfully changes lay attributions for offending, this raises two potential implications for punishment. First, some neurocriminologists expect neuroscience to *rightfully* challenge the claim that offenders *deserve* punishment (Greene and Cohen, 2004). Second, some theorists fear that neuroscience will *wrongfully* fortify the claim that offenders are dangerous beings in need of detention to prevent reoffending (Snead, 2007). Hence chapters five to nine consider whether lay people will perceive offenders as *deserving* punishment less, yet requiring detention more, following exposure to neurocriminology. The thesis discusses potential shifts in attitudes towards blame first (in chapters five to eight) and dangerousness second (in chapter nine).

Chapter ten investigates the effects of communicating neurocriminology on moral attitudes and moral behaviours that could render the recipient of the science more likely to offend. Hence, whereas chapter nine concerns the attributions of dangerousness made by those exposed to neurocriminological ideas, chapter ten assesses the possibility that such exposure might lead people to believe that the science provides a free license to transgress moral rules; such an effect might be expected if lay people infer that evidence of neurobiological causation undermines the blameworthiness of both others (chapters five to eight) and themselves (chapter ten) for immoral behaviour.

Chapter eleven ties together the notions of blame and dangerousness by suggesting that moral violators might be blamed for their actions in order to discourage them from committing future violations (and thereby becoming ‘dangerous’). Consequently, attributions of blame might be sustained in spite of neurocriminology

simply because blame is a useful concept – it communicates censure for the violation of moral rules. For the sake of clarity, chapters five to nine continue the structure of prior research by maintaining the divide between blame and dangerousness (e.g., Carlsmith and Darley, 2008). Nevertheless, chapter eleven ultimately entwines the two concepts: attributions of blame may remain strong in spite of neuroscientific arguments, if blame fulfils the important function of discouraging people from violating moral rules.

Five original studies were conducted for this thesis, including two online survey experiments and three field experiments, with a total of 8153 participants. In every study, some participants were presented with neuroscience, while other participants were not, such that the impact of presenting the neuroscience could be measured. In three of these studies, participants engaged in mock sentencing scenarios, making decisions about the blameworthiness, dangerousness and sentencing of a hypothetical offender. Overall, it was found that participants considered the offender no more dangerous, but significantly less blameworthy, after learning about neuroscience, though the effects were small. In this respect, the thesis concludes that neurocriminology will *not fundamentally* change public attitudes towards offenders.

The remaining two studies measured the moral behaviour of participants and their willingness to communicate a neuroscientific explanation of offending, respectively. After learning about neuroscience, participants did *not* behave less morally, but did recall significantly more instances of acting immorally in the recent past. Finally, in the richest and poorest boroughs of London, participants were significantly more reluctant to communicate a neuroscientific explanation of offending to offenders (in comparison to non-offenders). In this respect, the thesis concludes that neurocriminology will *not fundamentally* impact moral behaviour, though people may be reluctant to communicate the science as a result of fearing its impact.

Chapter 2: Overview of studies

The question of how much people change their attitudes after learning about neuroscience might invite a simple response: surely the answer depends on how much the person learns about the science. When the person is led to make the same philosophical assumptions that most neurocriminologists would make, a greater change in attitudes should result. This chapter discusses three approaches to empirically assessing the impact of communicating neurocriminological ideas. Each of these approaches varies by the extent to which they explicitly link the presented neurocriminological idea to the judged offending behaviour; in other words, the extent to which the participant is led to interpret the idea in the way that most neurocriminologists would interpret the idea. The five studies conducted for this thesis are then discussed in relation to these three methodological approaches. Finally, the structure of the thesis is justified in respect to the four key research questions of the thesis and the order in which those questions are addressed.

i. Linking the science to the offender

One would expect the degree to which the vignette links the neuroscientific explanation to the judged offending – and in this sense, does the linking *for* the participant – to affect the observed degree of attitude change. In contexts where the researcher has reason to expect people to resist a particular neurocriminological idea, the vignette may need to better link the neuroscientific explanation to the judged offending to achieve the same attitudinal outcome. By explicitly asserting the common philosophical assumptions of neurocriminology, the science may gain influence that would otherwise be undermined by other features of the vignette.

Where a study leads participants to link the neuroscience to the judged offender, one might criticise this study for inducing demand characteristics; in other words, one might argue that any effect of neuroscience is unsurprising if participants have read that the brain contributed to the offending behaviour. However, it is important to note that in the current research, none of the vignettes linked the neuroscientific explanation of offending to the retributive argument that the offender deserves less blame and punishment for his offending behaviour. Therefore, neuroscience was never explicitly presented as a mitigating factor; it was the decision of the participants whether or not to make the inference that the sentence warranted mitigation on the basis of the presented science.

It is also not an obvious outcome that lay people would find neuroscience mitigating. There is a wealth of research to suggest that lay people find offenders very much blameworthy for their behaviour even after being asked to assume that their behaviour can be entirely explained by uncontrollable events (e.g., Nahmias et al., 2005). In other words, people appear resistant to even strong mitigating arguments. Moreover, some theorists have argued that lay people will treat neuroscience as aggravating evidence, believing offenders to be more dangerous after learning about neuroscience (Snead, 2007). Other theorists have commented that neuroscience might exert both mitigating and aggravating effects (Aspinwall et al., 2012). Hence, it remains a theoretically important question to consider whether or not people consider neuroscience to constitute convincing mitigating evidence.

The degree to which the neuroscientific explanation is linked to the judged offending represents a key dimension by which the studies of this thesis differ. This dimension can be segmented into three approaches, starting with the approach that least leads participants to make the inferences that most neurocriminologists would make:

that is the psychological approach, which does not relate the science to the judged offender, followed by the mock court approach, which does relate the correlational science to the judged offender, and finally, the philosophical approach, which presents the science as a complete causal explanation of the judged offending. The five studies conducted for this thesis will be discussed in relation to each one of these approaches.

The studies conducted for the thesis were designed to answer four key research questions:

1. Does the communication of neurocriminological ideas reduce the extent to which lay people attribute blame to offenders?
2. Does the communication of neurocriminological ideas increase the extent to which lay people attribute dangerousness to offenders?
3. Does the communication of neurocriminological ideas increase immoral behaviour among lay people?
4. Do lay people seek to censor neurocriminological ideas from offenders?

A comparison of the key features of all five studies can be found in Table 1, including the answers that each study suggests to the four research questions. The table is followed by a summary of the main research questions and findings of each study, including a description of how the studies relate to each other.

Table 1.

A Summary of the Fives Studies Conducted for this Thesis.

	The psychological approach	The mock court approach			The philosophical approach
	The theatre study	The online study	The postcard study	The Facebook study	The museum study
Participants judged:	A young offender	A rational or an irrational offender	The act of communicating neurocriminology	The psychopathic brain	A neurobiological or biosocial offender
Suggested answers to the four research questions					
Do neurocriminological ideas reduce attributions of blame?	Slightly	Slightly	-	-	Slightly
Do neurocriminological ideas increase attributions of dangerousness?	No	No	-	-	No
Do neurocriminological ideas increase immoral behaviour?	-	-	-	No	-
Do people seek to censor neurocriminological ideas from offenders?	-	-	No	-	No
Key features of the communicated neuroscience					
It mentioned the brain	✓	✓	✓	✓	✓
It mentioned genes and the environment					✓
It used causal language			✓		✓

It linked the science to the judged offender		✓	✓	✓	✓
The key dependent variables					
Attributions of blame	✓	✓			✓
Attributions of dangerousness	✓	✓			✓
Moral attitudes and behaviours				✓	✓
The willingness to communicate the science to offenders			✓		
Sample size	728	821	832	750	5,084

ii. The psychological approach

The psychological approach consists of the studies that do the least linking of the neuroscientific explanation to the judged offender for the participants. In this type of study, participants read a (correlational or causal) explanation of (the relevant or all) behaviour, before making judgements about a specific offender. One of the thesis studies adopted this approach. The target of the measured attitudes is stated in brackets; for example, the theatre study measured attitudes towards young offenders.

The theatre study (the young offender)

There has been growing public interest in adolescent brain development, with the emergence of popular publications for adolescents and their parents, such as The Teenage Brain (Jensen and Nutt, 2016), Blame My Brain (Morgan, 2013) and Brainstorm (Siegel, 2014). This area of research, known as adolescence neuroscience, identifies the physical correlates of the stronger impulses experienced, and the weaker

capacity for impulse control available, during adolescence (Steinberg and Scott, 2003). Since young people are far more likely to offend than adults (Hirschi and Gottfredson, 1983) and poor impulse control predicts offending (Gottfredson and Hirschi, 1990; Pratt and Cullen, 2000), the poorer impulse control of young people may contribute to their disproportionate participation in offending (Hirschi and Gottfredson, 1983; Steinberg, 2012).

Lay people may believe that poor impulse control compromises the capacity of adolescents to act upon reason and thereby exercise rational choice (Steinberg, 2013). Consequently, young offenders may be attributed less blame for their impulsive actions and greater scope to develop impulse control. This study tested whether 728 people would reach such conclusions after watching a play about adolescent brain development. However, such attitude change would only result if participants linked the neuroscientific explanation to the survey questions: the play did not relate neuroscience to crime, blame, reoffending or the specific offender. Instead, the participants were required to infer that neuroscience regarding a general category of people – adolescents – applied to a particular category member – the specific young offender described in the vignette. Hence, in this study, participants responded to a neurobiological explanation of the behavioural tendencies characteristic of a particular age category. The findings were published in the journal Frontiers in Psychology.

iii. The mock court approach

The mock court approach consists of studies that more explicitly link the neuroscientific explanation to the judged offending. In this type of study, participants read a correlational explanation of a specific mental trait or disorder that directly affects the described offender, before making judgements about that offender. The trait or disorder

is one that concerns the capacity of the offender to exercise rational choice. Three of the thesis studies adopted this approach.

The online study (the rational and the irrational offender)

In the theatre study, participants did not change their attitudes towards adult offenders after learning about adolescent brain development. Perhaps this is unsurprising: it may be unreasonable to expect people to make inferences from adolescent brain development to the brain of the impulsive adult offender. Hence, in this study, the neuroscientific explanation was more explicitly linked to the judged offending: while the relationship between neuroscience, blame and reoffending remained unstated, the actual brain of the judged offender was described. In this respect, the vignette featured the type of neuroscientific testimony that can be presented at the sentencing stage of real court cases (e.g., Denno, 2015). In this vein, the 829 participants also read a detailed description of the offence, the prosecution testimony and the defence testimony.

In prior mock court research, the focus has always been upon offenders with neurobiological characteristics that *compromise* their capacity to exercise rational choice (e.g., Gurley and Marcus, 2008). Yet, in a CJS informed by neuroscience, there will also exist offenders with neurobiological characteristics that *support* their capacity to exercise rational choice. Hence the online study also measured attitudes towards the individual who offends on impulse yet is normally controlled and so normally capable of exercising rational choice. Therefore, in this study, participants responded to a neurobiological explanation of a particular (impulsive or controlled) character. The findings were published in the journal Frontiers in Psychology.

The postcard study (the act of communicating neurocriminology)

The majority of the thesis studies measured the effects of communicating neurocriminology on lay attitudes towards offenders. This gives rise to the question of whether the resulting attitude change is beneficial and therefore, whether neurocriminology *should* be communicated to promote such a change in attitudes. The postcard study was designed to empirically measure what the public think about this normative question; that is whether the public support the dissemination of neurocriminology. The return rate of postcards, each of which promoted a neurocriminological idea, was used to measure public approval for communicating the science. Once again, in this study, the neuroscientific explanation was more explicitly linked to offending: while the relationship between neuroscience, blame and reoffending remained unstated, the brain was described as the complete cause of 'most offending'. Hence, in this study, participants responded to the *act* of communicating neuroscience. The study design was published in Frontiers in Psychology.

The Facebook study (the psychopathic brain)

A key finding of the postcard study was that certain people *do* wish to censor the communication of neurocriminology. Therefore, the Facebook study was designed to probe the motives for that censorship. One possible explanation for the censorship is that lay people fear the change in attitudes that could arise from the dissemination of neurocriminology (e.g., offenders failing to take responsibility for their behaviour and thereby legitimising their desires to reoffend). Hence this study tested whether this potential fear is warranted; or specifically, whether communicating the neuroscience of psychopathy would promote the attitudes exhibited by psychopathic offenders. In this study, participants were presented with personalised feedback about their own brains

(or minds) and effects on their moral decision making were measured. Once again, the neuroscientific explanation was more explicitly linked to the relevant moral behaviour: while the relationship between neuroscience, blame and reoffending remained unstated, the participant read a description supposedly of their own brain. Hence, in this study, participants responded to a neurobiological explanation of psychopathy that was related to their own brain. The study design and findings were published, in two separate articles, in Frontiers in Psychology.

iv. The philosophical approach

The philosophical approach consists of the studies that most explicitly link the neuroscientific explanation to the judged offending. In this type of study, participants are asked to assume that a *causal*, *complete* and *universally applicable* explanation of all behaviour is true, before making judgements about a specific offender (Nichols, 2011). One study adopted this approach.

The museum study (the caused offender)

In this study, neuroscience was applied beyond the adolescent age category of concern to the theatre study. In fact, the brain was depicted as the *complete cause* of every decision to either resist or give in to impulse. Note this form of neuroscience therefore contrasts with the neuroscience depicted in the online and theatre studies; these two studies simply explained the behavioural *tendency* to act impulsively, void of explicitly causal language. In the museum study, neuroscience was also applied to the very offender and the very offence that participants were subsequently asked to judge; here termed the caused offender, since participants read a complete causal explanation of his actions. For the first round of judgements, participants were also asked to assume that all the presented science was true. Hence the neuroscientific explanation was explicitly

linked to the judged offending, yet the relationship between the brain and blame remained unstated.

This study also extended the breadth of the neurocriminological perspective by asking participants to consider either the brain or genes and the environment as causes of offending. In the online study, neuroscience was used to explain the (impulsive or controlled) character of the offender. In the museum study, neuroscience was used to explain a single act, with the participant left to make either a situational or dispositional interpretation of that act; in other words, left to decide whether the impulsive act reflected an unusually tempting situation or an unusually impulsive character. Although both the online and museum studies described impulsive offences, in the museum study, that impulse resulted from a rare opportunity to steal money rather than an emotive desire for vengeance. In this respect, the participant could decide whether to attribute stealing to the rare situation the offender found himself within or to his disposition. The museum study therefore presented the opportunity for neuroscience to strengthen a situational (rather than a dispositional) explanation of offending. Overall, in this study, participants responded to a causal and universal application of neuroscience to all instances of impulse control.

v. Thesis structure

The studies conducted for this thesis aimed to answer four key research questions:

1. Do neurocriminological ideas reduce attributions of blame? This question was addressed in chapters five to eight by the following studies:
    - The theatre study
    - The online study
    - The museum study

2. Do neurocriminological ideas increase attributions of dangerousness? This question was addressed in chapter nine by the following studies:
    - The theatre study
    - The online study
    - The museum study
3. Do neurocriminological ideas increase immoral behaviour? This question was addressed in chapter ten by the following studies:
    - The Facebook study
    - The museum study
4. Do people seek to censor neurocriminological ideas from offenders? This question was addressed in chapter eleven by the following studies:
    - The museum study
    - The postcard study

Each empirical study aimed to answer not just one of the above four questions but two, three or even all four questions. Therefore, each study is discussed across multiple chapters; for example, where a study finds something regarding blame and dangerousness, that study is discussed in both chapters five to eight about blame and chapter nine about dangerousness. Every study bears relevance to multiple types of attitude and so multiple chapters; the complete procedure of the study is described the first time that study is discussed.

The structure of this thesis is to discuss the above four research questions in sequence. The sequential structure of the thesis is important since the reasoning found in earlier chapters informs the reasoning found in subsequent chapters. There is a relationship between the dependent variables of blame, dangerousness and moral

behaviour that justifies the order in which the four research questions are tackled. It is important to discuss attributions of blame (chapters five to eight) before perceptions of dangerousness (chapter nine); since the hypothesis that neurocriminologically informed people might consider offenders to be more dangerous (in chapter nine) derives from the hypothesis that the same people might consider offenders to be less blameworthy (in chapters five to eight). After learning about neurocriminology, people might be more likely to consider offenders dangerous (chapter nine) on the basis that they are less able to control their behaviour since that behaviour has a neurobiological cause. Therefore, it is important to first examine whether and why a person might consider an offender to have less control over his/her behaviour when that behaviour has a neurobiological cause (chapters five to eight); since this is the premise for the argument that people may consider such offenders to be more dangerous (chapter nine).

The notion that people consider offenders whose brains 'made them do it' to have less control over their behaviour also acts as the premise for the hypothesis of chapter ten: that people are less likely to behave morally after learning about neurocriminology and subsequently, perceiving themselves to have less control over their own moral behaviour. Hence, again, it is appropriate to discuss the impact of neurocriminological ideas on attributions of control (chapters five to nine) before considering their impact on moral behaviour (chapter ten).

Finally, the notion that people may behave less morally after learning their brains 'made them do it' (chapter ten) acts as the premise for the hypothesis of chapter eleven: that people 'use' attributions of blame to discourage immoral behaviour. If the concept of blame really does reduce immoral behaviour, perhaps people also unconsciously 'know' this fact and thereby seek to protect the concept of blame from neurocriminological attack. Hence, again, it is appropriate to discuss the impact of

neurocriminological ideas on moral behaviour (chapter ten) before considering how people might 'use' attributions of blame to maintain moral behaviour (chapter eleven). In all of these respects, the sequential structure of the thesis ensures its coherence and flow, whereby the reasoning found in later chapters builds upon the reasoning found in earlier chapters.

vi. Conclusion

This chapter contextualised the five studies conducted for this thesis under three approaches to empirically assessing the impact of communicating neurocriminological ideas; these approaches and the respective studies were:

1. The psychological approach
    - The theatre study
2. The mock court approach
    - The online study
    - The postcard study
    - The Facebook study
3. The philosophical approach
    - The museum study

The three approaches differ according to how much they lead people to make the same philosophical assumptions that most neurocriminologists would make. Whereas the philosophical approach explicitly links a neurocriminological idea to the judged offending, the psychological approach allows participants to make that link for themselves. Subsequently, the sequential structure of the thesis was justified, in terms of the need to discuss the dependent variable of blame first, followed by dangerousness, moral behaviour and finally, the wish to censor the science from offenders. In turn,

these four dependent variables form the basis of the four key research questions of the thesis.

Chapter 3: The experimental method

The empirical component of this thesis adopted the experimental method to establish causal relationships between exposing lay people to neurocriminological ideas and their subsequent attitudes towards crime and punishment. This chapter first justifies the use of the experimental method in this particular field of research regarding public opinion of sentencing, while also describing the various features of the experimental method. The chapter proceeds by considering the choice of vignettes used in the current research, including the decision for those vignettes to describe acts of impulsive offending. Finally, the three approaches to conducting this research (as raised in the previous chapter) are discussed in respect to their relative methodological strengths and weaknesses.

i. The experimental method

The experimental method is especially suited to studies of sentencing preferences, given evidence that people report using and actually use *different* justifications for punishment (Carlsmith et al., 2002; Carlsmith, 2008). This finding of a discrepancy between people's sentencing 'theory' and their practice is consistent with a broader theory of lay reasoning: people tend to reach decisions on the basis of emotion and intuition *before* generating rational justifications for those decisions (Haidt, 2001). Hence the justifications that people explicitly support differ from the justifications that people implicitly apply in the practice of making concrete decisions. In mock sentencing scenarios, lay people tailor the sentences they recommend to retributive concerns, yet explicitly, they also express equal support for consequentialist concerns, such as general deterrence (Carlsmith et al., 2002). Consequently, there is no correlation between the implicitly measured sensitivity of people to different sentencing rationales and their stated support for those rationales (Carlsmith, 2008). Such findings suggest

that people either deliberately disguise their motives for punishment or, more likely, do not know their motives. This gap between sentencing behaviour and sentencing opinion illustrates the importance of deciding which – behaviour or opinion – is the phenomenon of interest.

While, in the current research, lay people could have simply been asked to report their inferences from neurocriminology to punishment, such a method would capture the effects that lay people *expected* neurocriminology to have on their attitudes; in other words, such a method would capture sentencing opinion. Instead, this thesis is concerned with the actual attitudinal effects of neurocriminology and therefore sentencing behaviour. It was therefore appropriate for the empirical research to adopt the experimental method and in particular, the independent-groups design: in every study, each participant was randomly allocated to one of several groups; also known as conditions. The participants in a particular group were presented with a particular neurocriminological idea that differed from the idea presented to participants in a different group. All participants then answered the same series of questions about a hypothetical offender or offending in general. Since these questions were answered using rating scales, the average responses between groups could be compared.

The random allocation of participants to conditions ensured that any difference in the average responses between conditions must have resulted from a difference in the neurocriminological idea presented in that condition. Response differences could not be attributed to differences in the traits of respondents, since, following random allocation, traits should not differ systematically between conditions. Consequently, it was possible to determine the effect of presenting each neurocriminological idea, as reflected in response differences. In addition to groups of participants who learnt about a neurocriminological idea, known as experimental groups, there were control groups;

those are groups where participants were not exposed to any neurocriminological idea. As a result, it was possible to determine the attitudinal effect of exposure to neurocriminology compared to no exposure.

Since participants were unaware of the alternative conditions (or alternative neurocriminological ideas) in each condition, their responses did not result from a conscious comparison of conditions (or ideas). In this respect, the independent-groups design contrasts sharply with the repeated measures design; in just one of the studies, participants were sequentially presented with two different neurocriminological ideas – a repeated measures design. In this case, participants may have adjusted their responses to the second idea in direct comparison or opposition to the idea presented first. The repeated measures design therefore encounters the same problem as directly asking participants for their opinions: the findings are limited by the degree of insight people have into their own sentencing practices. Since such insight is greatly limited (Carlsmith et al., 2002; Carlsmith, 2008), the independent-groups design is more capable of capturing the effects that neurocriminology *implicitly* poses for sentencing *practice*. People may not endorse those effects if the findings were subsequently explained to the very participants who generated them. In this respect, the empirical research seeks to capture 'the power of experimentation to disentangle the manifest from the latent functions of societal practices' (Fincher and Tetlock, 2015, p.640). Hence the experimental method was chosen.

Neuroscience can be communicated in different forms. In the majority of studies, the science was explained in the text displayed onscreen. In such an online vignette experiment, participants read a vignette about neuroscience and a hypothetical offender, before making judgements about the offender. The online experiment allows for the greatest experimental control, such that the precise effect of a very particular

neurocriminological idea can be isolated from the similar effects of similar ideas. The online experiment also typically allows more variables to be measured, including the hypothesised mechanisms of the effects, also known as mediators. This thesis consists not only of online vignette studies but also three field experiments, conducted at the National Theatre, at the Science Museum and on the streets of London. Hence the effects of communicating neurocriminology were also studied at real-life sites of science communication.

i. Vignette design

Previous research has presented brain mechanisms as an explanation of behaviour in general (e.g., Nahmias, Morris, Nadelhoffer and Turner, 2005) or severe mental illness in particular (e.g., Schweitzer and Saks, 2011). In the latter case, participants typically read a neurobiological explanation of schizophrenia or its main symptom, psychosis. Psychosis describes the hallucinations and delusions that prevent schizophrenic offenders from understanding the real consequences of their criminal acts. If the offence is committed during a psychotic episode, the diagnosis of schizophrenia is clearly relevant to establishing the intentionality of the offender. In this respect, psychosis unambiguously compromises the capacity for the affected individual to exercise rational choice.

The major contribution of this thesis is to present neuroscience as an explanation of a trait, impulsivity, that does not render a person *incapable* of exercising rational choice. In the majority of the studies conducted for this thesis, brain mechanisms are described as an explanation of impulse control (Morse, 2004b) and in one study, as an explanation of the emotional capacity to *appreciate* the difference between right and wrong (Morse, 2008b). Both impulse control and emotional sensitivity concern the

capacity to *act upon* rational knowledge of the difference between right and wrong, as opposed to *just having* that knowledge. Hence the question of whether impulse control and emotional sensitivity are also components of the capacity for rational choice is more contentious. This thesis aims to identify the effect of communicating neuroscience in relation to offenders with these milder deficits in rational choice; specifically, the impulsive or emotionally stunted (rather than the hallucinatory or delusional) offender.

In three studies, the described offending was impulsive, yet the cause of the impulse (i.e., the source of provocation) still varied between the studies. These studies took place in a museum, a theatre and online respectively; hence the thesis refers to these studies as the museum study, the theatre study and the online study. In these studies, the offences ranged from a university graduate who committed opportunistic theft to purchase a luxury holiday (the museum study) to a 'serious assault against a stranger who insulted [the offender]' (the theatre study) and a case where the offender used a hammer to cause permanent brain injury; in other words, an act of Grievous Bodily Harm that was clearly disproportionate to the source of provocation (the online study). While the fourth study (the postcard study) also described a neurobiological explanation of impulsive offending, a different outcome measure was taken: the outcome of how much people supported the act of communicating neuroscience to offenders. The fifth study (the Facebook study) departed from the use of neuroscience to explain impulsive offending; instead, the empathy deficit characterising psychopathy was explained, with moral attitudes measured.

Every study conducted for this thesis has both methodological strengths and weaknesses. While those specific strengths and weaknesses can be found in the chapter that discusses the relevant study, some of these methodological points are common to all the studies. For example, every study is limited by the fact that it only documents the

effects of communicating neurocriminological ideas that are presented briefly and, with exception for the theatre study, through text-based communications. Therefore, this research cannot test the possible effects of more powerful interventions, such as a series of classes about neurocriminology, attended over several months or years. Nevertheless, this limitation does not negate the value of the current research, since, in real-life, neurocriminological ideas are presented in short text-based communications, such as in the media. Similarly, where neurocriminological ideas are presented as expert testimony in court, exposure to these ideas will be relatively brief.

Other methodological considerations represent the advantages and disadvantages of the broader approach that each study adopts. For example, the mock court approach derives ecological validity from the fact that it represents the way neuroscience is presented in the courtroom; hence the findings from this approach may be more applicable to real life. In contrast, the approach of experimental philosophy lacks ecological validity, since it asks participants to report their attitudes towards an offender *after assuming* that the presented science is a true, complete and causal explanation of the judged offending. Given that people would not necessarily make those assumptions in real life, this approach lacks ecological validity, yet gains internal validity: for example, if participants did not change their attitudes in response to neuroscience, the approach of experimental philosophy would be more suited to establishing *why* not; specifically, whether participants did not believe the science, did not apply the science to this particular offender or did not consider the science relevant to sentencing. The approach of experimental philosophy isolates the lattermost process by asking participants to assume the science is true and applicable. In turn, it becomes possible to narrow the focus of enquiry and test a specific reason why participants did not change their attitudes; hence this approach offers high internal validity.

In contrast, studies that adopt the psychological approach lack internal validity. These studies do not lead participants to make assumptions about the presented science and its relevance to the offender or even offending in general. Consequently, it would become difficult to establish which features of the presented science were responsible for that manipulation failing to change the measured attitudes; that is whether participants did not believe the science, apply the science or infer its relevance to punishment. Yet this methodological weakness also constitutes a strength: since the tested means of science communications has multiple features, it also probably better represents real-life forms of science communication.

For this thesis, data was collected under all three approaches: the psychological approach, the mock court approach and the approach of experimental philosophy. Therefore, an important contribution of this thesis is the range of complementary approaches adopted by each empirical study; the strengths of one approach compensate for the weaknesses of another approach. Whilst the specific procedures of each study will be described in the chapter where the study first appears, in this chapter, the primary research questions of each study are related to its methodological approach. The theoretical and methodological relationships between the studies are also discussed, with the aim of illustrating the empirical coherence of the thesis. Every study received individual ethics approval from the University of Oxford Central Ethics Committee.

ii. Conclusion

The five studies conducted for this thesis adopted the experimental method to assess the causal relationships between people learning about neurocriminological ideas and their subsequent attitudes towards crime and punishment. In each study there was an independent variable – some neurocriminological idea – and at least one dependent

variable – attributions of blame, attributions of dangerousness, moral behaviour and/or the willingness to communicate the neurocriminological idea to offenders. The chapter began by justifying the use of the experimental method in this particular research field, paying particular attention to the discrepancy between public opinion in theory and public opinion in practise. The design of the vignettes, including the focus on impulsive offending, was then discussed. Since the current research adopted vignette designs from all three empirical approaches (the psychological approach, the mock court approach and the philosophical approach), the various studies were able to draw upon the relative methodological strengths and weaknesses of each approach.

This thesis now proceeds to address the four key research questions:

1. Do neurocriminological ideas reduce attributions of blame?
2. Do neurocriminological ideas increase attributions of dangerousness?
3. Do neurocriminological ideas increase immoral behaviour?
4. Do people seek to censor neurocriminological ideas from offenders?

The next chapter sets the scene for all four of these research questions by highlighting the discrepancies between how lay people and neurocriminologists tend to explain criminal behaviour – discrepancies that, if communicated to lay people, might influence their attributions of blame, their attributions of dangerousness and their moral behaviour.

Chapter 4: Attitudes towards crime and punishment

It is a core claim of this thesis that lay people and specific neurocriminologists have remarkably different views about the causes of crime and the appropriate justifications for punishment. It is possible that the discrepancies in their views simply reflect a public that is uninformed by neuroscience. Hence this thesis aims to understand how lay people view crime and punishment *after* learning about neuroscience. Although there is clearly no single lay perspective and no single scientific perspective on crime and punishment, there are certainly trends in how lay people and neurocriminologists differ in their views. This chapter describes those discrepancies between the dominant attitudes of lay people and neurocriminologists.

First in this chapter, the field of neurocriminology is introduced, with a particular focus on the philosophical assumptions that most neurocriminologists make about the relationship between the brain and behaviour. The chapter then presents specific empirical findings regarding that relationship, giving examples of the association between the activity of specific brain regions and the traits typically involved in particular forms of criminal behaviour. Subsequently, the chapter discusses justifications for punishment that are compatible with neuroscientific perspectives on criminal behaviour; a particular emphasis is given to the controversial argument of Greene and Cohen (2004). Finally, the question is addressed of how lay people tend to explain crime and justify punishment. Overall, this chapter aims to identify discrepancies between the views that lay people currently hold and the views that Greene and Cohen expect lay people to hold after learning about neurocriminology. The nature of these discrepancies provides the basis for predicting whether and how the attitudes of lay people might change in response to neurocriminology.

i. How do neurocriminologists explain crime?

In the study of crime, neurocriminology is a new field that aims to identify the brain structures and brain functions that correlate with criminal behaviour (Raine, 2013). Neurocriminologists can make two types of claim about crime causation: philosophical assumptions and empirical findings. It would be impossible to document empirical evidence for the grand causal assumptions of some neurocriminologists, since the methodological tools of the field are limited. In particular, it would be unethical and scientifically unfeasible to induce criminal behaviour through experimental manipulations of brain activity. Therefore, the field is reliant on correlational studies, in which one cannot measure all the forms of behaviour and brain activity that might covary with the particular antisocial and neurobiological traits under study. Nevertheless, it remains possible that lay people might also make causal assumptions after engaging with specific empirical findings.

In order to empirically study the attitudinal effects of communicating neurocriminology, one must first identify the type of neuroscience that will be presented to participants. Although theorists expect neurocriminology to have major attitudinal effects, the type of neuroscience expected to have this effect is rarely specified (e.g., Greene and Cohen, 2004). Hence this thesis begins by depicting the spectrum of neurocriminology, ranging from its common philosophical assumptions to examples of empirical findings. Such a spectrum provides a context for positioning the forms of neuroscience described to participants in the empirical studies of this thesis.

Philosophical assumptions of neurocriminology

Ultimately, it is only feasible that the empirical findings of neurocriminology could affect punitive attitudes if people made *causal* inferences from such findings; in other

words, if people made the philosophical assumption that the correlations between brain and behaviour identified in neuroscience reflect, in fact, causal effects of the brain on behaviour. In this context, the correlational nature of the data reflects a limit on the methodological capacity of neuroscience rather than the absence of causality. The current capability of neurocriminological research is strictly correlational; the science can only identify associations between brain and behaviour that could always be the spurious effects of unmeasured variables. Nevertheless, this thesis aims to anticipate a future in which advances in neuroscientific technology render it increasingly difficult to reject a causal interpretation of the brain in criminal behaviour.

The empirical studies of this thesis differed according to whether the assumption of causality was stated or not. In some studies, participants read that the brain *caused* the offending behaviour, whereas, in other studies, participants read only that there was an association between the brain and offending. Both forms of neuroscience are reviewed in this chapter; that which makes the key philosophical assumptions (such as the assumption of causality) and that which avoids those assumptions. First, I examine these philosophical assumptions of neurocriminology: the assumption that brain activity is the product of uncontrollable prior events, that brain activity is responsible for conscious thoughts, desires and plans, and that brain activity is the product of uncontrollable social (not just genetic) factors. These assumptions are hereon labelled the assumption of determinism, the assumption of the conscious brain and the assumption of the social brain.

The assumption of determinism

A common philosophical foundation of neuroscience is determinism: the idea that every behaviour has a preceding cause, which, in turn, has a preceding cause and so on, until one reaches causal events that occurred before the actor was born (Morse, 2008a). In

this respect, neuroscience tends to reject the existence of a free will that is incompatible with determinism (hereon referred to as incompatibilist free will); such a free will would have to entail a part of the conscious mind that was entirely unaffected by prior genetic and environmental events:

> Only a small number of philosophers and scientists believe that human beings possess libertarian freedom of action and will…because it is so implausible (Morse, 2008a, p.13-14).

There is no data to evidence this 'small number'; hence empirical research is still required to survey the deterministic beliefs of scientists.

Nevertheless, in the field of neurocriminology, key authors reveal their causal interpretation of the empirical data; for example, writing that 'every action is caused by brain events' (Tonry, 2011, p.300) or that 'all behaviour is caused by our brains' (Raine and Glenn, 2014, p.161). It is not entirely clear what such authors mean by 'causal'. One can distinguish between probabilistic and deterministic causation. Probabilistic causation refers to there being a greater probability of criminal behaviour arising if certain conditions are present. In contrast, deterministic causation refers to criminal behaviour being inevitable if certain conditions are present. No neurocriminologist believes that any single brain characteristic would inevitably cause a person to behave criminally. However, Greene and Cohen (2004) do believe that if certain complex combinations of brain characteristics are present, criminal behaviour is inevitable. According to Morse (2008a, p.13-14), this underlying belief in deterministic causation is widespread among scientists:

> Most philosophers and scientists believe that the universe is deterministic or universally caused, or nearly so.

Regardless of their underlying beliefs, scientists can only make probabilistically causal claims from individual empirical studies due to methodological limits in

empirical tools. Brain scanning technologies cannot detect brain activity beyond certain spatial and temporal resolutions. Researchers are limited in the experiments that are ethical and practical to perform. Therefore, current empirical methods could not possibly evidence the truth of determinism, even if it were to be true. This thesis is not concerned with whether the world is deterministic, but with how lay people interpret the causal assumptions that (at least some) neuroscientists make (e.g., Greene and Cohen, 2004). Where causality is mentioned in the current research, participants are left to decide exactly what it means.

Most neurocriminologists do not consider the idea of the brain causing behaviour controversial. The claim that the brain causes criminal behaviour is not a claim that the unconscious mind or genes cause criminal behaviour. Conscious thought is also the product of brain activity, which, in turn, is caused by genes and the environment, including social factors (Greene and Cohen, 2004). Yet beyond neuroscience, lay people may overattribute the brain to genes and unconscious mental activity (Nahmias, 2006; Nahmias, Coates and Kvaran, 2007). More generally, the communication of neurocriminology may render salient the uncontrollable sources of criminal intent – caused by variation in the activity and volume of different brain regions and therefore, ultimately, genetic and environmental influences (Raine, 2013).

The assumption of the conscious brain

If lay people are reluctant to attribute conscious thoughts to a neurobiological chain reaction (Nahmias et al., 2007), they may be unreceptive to neurobiological explanations of premeditated offending. In turn, they may be unreceptive to neurobiological arguments that seek to challenge the moral responsibility, or document the dangerousness, of premeditated offenders. Hence in this thesis, the empirical studies do not seek to communicate the neuroscience of premeditated offending. Instead, the

empirical focus is the neuroscience of impulse control; where lack of impulse control, also known as self-control, has been proposed as a general theory of crime (Gottfredson and Hirschi, 1990).

People may more readily expect impulses to result from a physical mechanism, such as the brain, and consequently, anticipate the need for a physical interface between impulses and impulse control. Hence the focus on impulsive offending reflects the predicted receptivity of the lay person to neurobiological explanations in this particular context. Nevertheless, it is important to recognise that neurocriminologists do not intend their claims to be limited to impulsive offending. In fact, the brain is conceived to cause every form of (criminal) behaviour:

> The universe seems so sufficiently regular and lawful that rationality demands that we must adopt the hypothesis that universal causation is approximately correct (Morse, 2008a, p.14).

From the scientific perspective of 'universal causation', both premeditated and impulsive offending are equally caused by preceding uncontrollable events; both intent and impulse are forms of brain activity.

The assumption of the social brain

For most neurocriminologists, the social causes of brain activity are considered no more controllable than the genetic causes:

> Humans do not necessarily have more control over psychological or social causal variables than over biological causal variables (Morse, 2008a, p.14).

For those scientists who accept determinism, every cause is uncontrollable and no cause is more controllable than any other cause. In turn, no case of offending is more controllable than any other case and no offender is more morally responsible for criminal behaviour than any other offender:

> As a general matter, it is *always* true that our brains made us do it (Monterosso and Schwartz, 2012).
>
> For the determinist, biological causes…pose no new or more powerful…challenge to responsibility than non-biological or social causes (Morse, 2008a, p.14).

While most neurocriminologists therefore equate causes with equal uncontrollability, the lay person does not. Unlike scientists, lay people tend to reject the idea that every behaviour is caused by prior uncontrollable events (Nichols and Knobe, 2007; Sarkissian, Chatterjee and Brigard, 2010). By rejecting this philosophy of determinism, the lay person can consider certain influences (e.g., the brain) less controllable than other influences (e.g., friends). Hence, for the lay person, neurocriminology *could* pose a 'new or more powerful…challenge to responsibility' (Morse, 2008a, p.14).

Neurocriminologists typically stake claims about the neurobiological causes of criminal behaviour, whereas sociological criminologists typically stake claims about the social causes of criminal behaviour. Nevertheless, most neurocriminologists would not deny the latter claims, instead recognising that 'there are many causes that impinge on behaviour'. These researchers simply assert that 'all of them…must exert their influence through the brain' (Greene and Cohen, 2004, p.1781). Therefore, any cause of criminal behaviour must, at some stage, become a neurobiological cause before expressing itself in behaviour. The same, however, is not true of genetic factors. It is important to distinguish genes from the brain, or genetic factors from neurobiological factors. Since the brain is a vehicle for all other influences, brain activity changes across situations and the lifespan, whereas genes do not.

Empirical findings of neurocriminology

When communicating neuroscience in court, expert witnesses present not its common philosophical assumptions but specific empirical findings; for example, brain scans

conducted on the offender, highlighting abnormalities in particular brain regions and the cognitive functions of those brain regions in unimpaired people. Such scans suggest that the offender lacks the neurobiological hardware that gives people the capacity to exercise this cognitive function. In this context, neuroscience derives its legal relevance from what the scans suggest about the mind of the offender. The brain is the physical organ implicated in operations of the mind; the conscious and unconscious cognitive computations that generate behaviour. Only certain of these computations are legally relevant to assessing responsibility; those that provide the capacity for the offender to exercise rational choice. The capacity for rational choice provides the foundation for legal attributions of responsibility (Morse, 2008a). While its precise definition can differ by jurisdiction, it refers to the general capacity of the offender to form intentions from a basic knowledge of reality and illegality, and to act upon those intentions through self-control.

In this respect, neuroscience is presented in court for the same reason that clinical diagnoses are presented. Certain mental illnesses clearly undermine the capacity of the offender to exercise rational choice. Mental illness diagnoses have long been considered at sentencing and at trial, for example, to support pleas that the defendant is Not Guilty by Reason of Insanity (NGRI) – typically defined as the inability to know that the offending act is illegal due to a mental disorder (The American Law Institute, 1962). Clinical evidence might also be used to support other *mens rea* defences that argue even a sane defendant, who can distinguish right from wrong, still lacked intentionality at the time of the crime; in other words, lacked a complete awareness of the harms that would follow the criminal act (Cochrane, Grisso and Frederick, 2001).

These defences are sometimes granted to defendants whose offending is ‘obviously the product of forces that are ultimately beyond their control’ (Greene and

Cohen, 2004, p.1783). Such defendants include 'obviously' mentally ill offenders, whose criminal actions are associated with the delusions or hallucinations that characterise their illness. These so-called psychotic offenders are far more likely to secure an NGRI plea than psychopaths, whose deficit is one of emotional, rather than rational, understanding (Rogers and Zimbarg, 1987; Zonana et al., 1990; Warren et al., 2004). Generally, the success rate of NGRI pleas is extraordinary low both in mock trials (McGlynn and Dreilinger, 1981) and real trials (Callahan et al., 1991). Hence defence lawyers may be especially receptive to novel means of documenting deficits in rational choice, such as neuroscience. Here I review the neuroscience that those defence lawyers may use; specifically, the role of the brain in psychopathy, the role of the brain in self-control and the relative contribution of genetic and environmental factors to antisocial behaviour.

The role of the brain in psychopathy and offending

Psychopathy is a mental illness of particular relevance to this thesis, since one of the current studies examined the effects of communicating the neuroscience of psychopathy to lay people. Psychopaths do not experience the same degree of anxiety upon causing distress to other people; hence psychopaths are more likely to cause distress to other people, sometimes in the form of criminal acts (Aniskiewicz, 1979; Hare, 1991). Their reduced experience of empathy can be explained by their reduced activity in brain regions involved in emotional processing; these regions are known as the limbic structures and include the amygdala, which is implicated in the subjective experience of fear (Kiehl et al., 2001). This characterises a specific empirical finding, as opposed to a philosophical assumption, that could be communicated to lay people; indeed, this finding was communicated to participants in one of the current studies.

The role of the brain in self-control and offending

The remaining studies in this thesis communicated the neuroscience of deficits in rational, rather than emotional, choice: participants made judgements about an impulsive offender after learning about the neuroscience of self-control. In their general theory of crime, Gottfredson and Hirschi (1990) claim that criminal behaviour is generally the product of low self-control. By offending, an individual can attain the same rewarding goal without investing so much time, money and effort in the process; the legal route to attaining that same reward would be less risky in the long term, yet more costly in the short term. Hence Gottfredson and Hirschi reason that criminal behaviour results from a greater sensitivity to the short-term (than the long-term) consequences of behaviour; that is, low self-control. Accordingly, there is no shortage of the motivation to offend: theft, rape and violence are caused by our universal drive to attain 'money without work, sex without courtship [and] revenge without court delays' (Gottfredson and Hirschi, 1990, p.89). These criminal goals are naturally rewarding yet are only pursued by people who fail to recognise the long-term costs of offending; from legal and social costs to religious and physical costs. The theory proposes that everyone would offend if those costs were not sufficiently salient. The theory gains support from meta-analyses that suggest self-control is a strong predictor of offending (e.g., Pratt and Cullen, 2000).

Gottfredson and Hirschi (1990) claimed that people are born blind to the long-term costs of offending, which typically gain salience during the first eight years of life. During this period, parents bestow children with self-control – a fixed ability to recognise the long-term costs of deviant behaviours that give rise to immediate rewards. Not only does this theory posit the significance of parenting but 'the true genetic effect on the likelihood of criminal behaviour…is [claimed to be] minimal' (Gottfredson and

Hirschi, 1990, p.58) and ‘near zero’ (p.60). Such a claim directly contradicts the findings of longitudinal research conducted by Beaver et al. (2008): genetic differences between participants explained 52% to 64% of the variance in low self-control and 82% of the variance in the stability of self-control over two years. The remaining variance in self-control was explained by environmental influences that are not shared by members of the same family (e.g., friends and teachers). These findings were replicated by Boutwell and Beaver (2010), suggesting that self-control has a genetic component.

The primary neurobiological correlate of self-control is a region at the very front of the brain known as the prefrontal cortex (Raine, 2013). In particular, activity in a part of the fronto-median cortex responds more when participants cancel an intended motor action rather than translate that intention into action; in other words, when self-control is exerted (Brass and Haggard, 2007). Psychopathic offenders show an 11% reduction in the grey matter volume of their prefrontal cortex (Raine et al., 2000). Therefore, the reduced volume and activity of the prefrontal cortex might explain the impulsive offences committed by some psychopaths. In contrast, psychopathic liars have a greater white matter volume in inferior ventral areas of the prefrontal cortex, perhaps explaining their increased propensity for manipulation (Yang et al., 2007). Similarly, white-collar offenders have a greater grey matter volume in the same areas, perhaps reflecting the greater planning required to commit such offences (Raine et al., 2012). Grey matter contains the cell bodies, whereas white matter connects the cell bodies with each other. These studies document examples of empirical findings regarding the relationship between the prefrontal cortex and self-control in the criminal context.

The role of genes and the environment in offending

Some neurocriminologists question the legitimacy of positive retributivism on the basis that criminal behaviour results from a chain of events occurring in the brain (e.g.,

Greene and Cohen, 2004). Hence it is worth exploring in some depth how neurocriminologists think about some of the precursors of offending behaviour; with the aim of orientating the reader toward some of the key claims of neurocriminology and neuroscience more generally. The study of self-control demonstrates the neuroscientific approach to criminology: criminal behaviour is believed to result from deficits in cognitive and affective mechanisms, such as self-control and empathy, which are also brain mechanisms; in turn, the strength of these brain mechanisms is attributed to underlying differences in genetic and environmental influences (Raine, 2013).

Since the brain is the product of genes and the environment, neurocriminologists also study the genetic and environmental correlates of criminal behaviour. In the empirical research conducted for this thesis, participants normally learnt about the brain rather than genes and the environment. However, in one study, an explicitly biosocial explanation was presented, referring to genes and the environment instead of the brain. Given the relative rarity of criminal behaviours and the reluctance of people to disclose their own offending, neurocriminologists usually measure the precursors to criminal behaviour rather than offending itself. Collectively, these precursors are known as antisocial personality and behaviour (APB).

Broadly there are two approaches to measuring the genetic contribution to APB: molecular genetics and behavioural genetics. Molecular genetics involves identifying the specific genes that influence the probability of people displaying antisocial behaviours. In contrast, behavioural genetics involves calculating what proportion of the variance in offending can be collectively explained by hundreds of genes; the variance in offending is the differences in the degree of antisocial behaviour displayed by different people. Therefore, the black box approach of behavioural genetics cannot specify which genes are involved, only their relative net importance.

Particular versions of genes and their interactions with environmental factors have been associated with an increased probability of APB. For example, Thapar et al. (2005) studied a sample of children with attention deficit hyperactivity disorder (ADHD): carriers of the valine/methionine version of the catechol O-methyltransferase (COMT) gene were more likely to exhibit APB, potentially because of its impact on the development of the prefrontal cortex. Similarly, Retz et al. (2004) found that the deletion/insertion version of the serotonin transporter promoter gene (5-HTT) explained an impressive 5% of the variance in the violent behaviour of offenders. In their longitudinal research, Caspi et al. (2002) documented an example of the interaction effect: child carriers of the low (as opposed to high) activity version of the MAOA gene were more likely to exhibit APB and more likely to be arrested in adulthood only if the child had also experienced family abuse. Hence the expression of this genetic disposition to APB was contingent on a particular environmental trigger.

This approach of molecular genetics is limited: no geneticist would expect a single gene to explain any substantial portion of the variance in offending (Turkheimer, 2000). Therefore, molecular genetics cannot specify the scale of the genetic contribution to criminal behaviour; whereas behavioural genetics can. Behavioural geneticists calculate the proportion of variance in APB resulting from genetic variance by comparing the APB of monozygotic and dizygotic twins: whereas monozygotic twins share all of their DNA, dizygotic twins share only 50% of their DNA (Moffitt, Caspi and Rutter, 2005). If monozygotic twin members behave more similarly than dizygotic twin members, their greater behavioural similarity must be explained by their greater genetic similarity. The similarity of monozygotic twin members on measures of APB, in comparison to the similarity of dizygotic twin members, indicates the size of the genetic contribution to APB. For example, if APB was entirely caused by genes,

one would expect monozygotic twin members to score twice as similarly on APB measures as dizygotic twins. This approach does assume that monozygotic twin members share the same degree of environmental influences as dizygotic twin members. This assumption could be questioned on the basis that parents, for example, might treat monozygotic twin members more similarly than dizygotic twin members – an environmental factor that could explain their greater behavioural similarity. Nevertheless, behavioural genetics provides a reasonable means of calculating the heritability of APB.

Any variance unexplained by genes is attributed to the environment. Behavioural geneticists distinguish between two forms of environmental influence: the shared environment and the non-shared environment. The shared environment comprises features of the environment shared by children raised in the same home, such as parents and family experiences. The non-shared environment comprises features of the environment that are not shared by children raised in the same home, such as friends, teachers and head injuries. The proportion of APB explained by the non-shared environment can be calculated by comparing members of a monozygotic twin pair on measures of APB: any difference in the APB of monozygotic twin members must be explained by the non-shared environment, since monozygotic twin members share all their DNA and, by definition, their shared environment. For example, if there were no differences in the APB of monozygotic twin members, this would suggest there was no contribution of the non-shared environment to APB. Therefore, behavioural genetics can be used to calculate the relative contributions of genes, the shared environment and the non-shared environment.

A body of twin research has accrued on APB. In fact, meta-analyses have been conducted to estimate the overall proportion of variance in APB that is explained by

genes, the shared environment and the non-shared environment. The estimated proportion of variance in APB explained by genes ranges from 41% (Rhee and Waldman, 2002) to 50% (Miles and Carey, 1997) and 56% (Ferguson, 2010), depending on the inclusion criteria for the studies reviewed. In the same meta-analyses, the calculated contribution of the non-shared environment is 43% (Rhee and Waldman, 2002) and 31% (Ferguson, 2010), while the calculated contribution of the shared environment is 16% and 11% respectively. The conclusion of behavioural genetics, therefore, is twofold: first, genetic differences explain approximately 50% of the variance in APB (Raine, 2013), and second, the remaining 50% of the variance is predominantly explained by the non-shared environment rather than the shared environment. In fact, this same conclusion has also arisen for personality traits that are not antisocial (Polderman et al., 2015). Behavioural genetics, therefore, casts doubt on sociological theories of crime that place explanatory emphasis on the shared environment, such as parenting (e.g., Gottfredson and Hirschi, 1990).

Nevertheless, behavioural genetics does support sociological theories that place explanatory emphasis on the non-shared environment, such as peer influence. More generally, genetic and neuroscientific findings are entirely compatible with the social motivations that certain sociological theories attribute to offenders; for example, the patriarchal motives that feminist theorists attribute to perpetrators of interpersonal violence (IPV; e.g., Johnson, 2005). Still, the application of neurocriminology to IPV remains controversial:

> Many researchers have stubbornly refused to explore the idea that biological factors might partially account for why some individuals abuse their romantic partners (Boutwell and Lewis, 2015, p.271).

Two studies have assessed the heritability of IPV, generating similar conclusions to the heritability of APB more broadly. Hines and Saudino (2004) concluded that 15% to

25% of the variance in intimate partner aggression could be explained by genetic differences, while the shared environment explained no variance. More starkly, Barnes et al. (2013) found that genetic variance explained 24%, 54% and 51% of differences in the self-reported frequency of hitting a partner, injuring a partner and forcing a partner to have sex, respectively; once again, the shared environment was unable to explain any variance. While the reliance on self-report should be highlighted, this evidence reinforces the neurocriminological perspective: criminal behaviour primarily results from genes and features of the non-shared environment.

In contrast to most social scientists, 'plenty of criminologists…adhere to a purely nurture explanation of human variability' (Beaver, Barnes and Boutwell, 2015, p.IX). In part, this may reflect a conceptualisation of biology as only capable of explaining 'univariate, purposeless, and unprovoked' examples of aggression; in other words, acts that appear to have no visible social cause (Ferguson, 2010, p.164). Some sociologists might envisage the proposed genetic contribution to IPV as genes that suddenly generate random acts of IPV. Yet this is not the neurocriminological perspective on genes: a genetic 'disposition to commit IPV' is far less direct than the term implies; it is a genetic disposition to adopt a social position or cognitive perspective that increases the probability of behaviours such as IPV. For example, feminist theory attributes IPV to patriarchal motives: men use violence to control women to maintain the power imbalance between genders (e.g., Johnson, 2005). The male drive to control women facilitates the objectification of women and, once fulfilled, delivers men a sense of worth. This feminist explanation is, however, entirely compatible with a genetic explanation: some men may be more genetically driven to maintain the patriarchy than other men, ultimately to the extent of committing IPV. Hence the genetic account of criminal behaviour simply describes a genetic footing to

social or cognitive causes of offending. Consequently, it is a mistake to believe that neurocriminology invalidates the existence of damaging social motives and the need to challenge those motives.

In sum, although the exact proportions vary between studies, neurocriminological research converges on the conclusion that approximately equal proportions of variance in criminal behaviour can be attributed to genetic and environmental influences respectively (Raine, 2013); with a particular emphasis on the non-shared environment (Ferguson, 2010). The brain is routinely considered the intersection for genetic and environmental influences (Greene and Cohen, 2004). In turn, the brain is responsible for the cognitive precursors to offending, such as low self-control (Pratt and Cullen, 2000) and lacking empathy (Kiehl et al., 2001). Therefore, the philosophical assumptions commonly made by neurocriminologists are consistent with specific empirical findings regarding the brain, genes and the environment.

ii. How do neurocriminologists justify punishment?

The philosophical premise and empirical findings of neurocriminology suggest that criminal behaviour arises from the brain, which, in turn, is the product of genetic and environmental factors. If this model of behaviour holds merit, one might reconsider the idea that offenders entirely *deserve* to be punished for their criminal acts. Neurocriminologists take different perspectives on the retributive justification for punishment. In order to place these perspectives in context, this thesis first explores different forms of retributivism, with the aim of understanding which form of retributivism neuroscience has the potential to challenge (Greene and Cohen, 2004).

Retributivism generally advocates that offenders should be punished in proportion to the seriousness of the offence committed and their blameworthiness for

the act; this feature of retributivism is known as the principle of proportionality (von Hirsch, 1976). Most retributive theories comprise two components: positive retributivism and negative retributivism (Walen, 2016). Positive retributivism claims that offenders deserve to be punished for their wrongdoings. Negative retributivism claims only that wrongdoing is necessary for punishment (Mackie, 1982). Consequently, negative retributivism prohibits punishing the innocent or disproportionately punishing the guilty; the severity of punishment must not exceed the blameworthiness of the offender or the seriousness of the offence.

It is possible to be an entirely positive retributivist and hence claim that offenders deserve punishment, without prescribing any principle to limit the severity of the punishment. It is also possible to be an entirely negative retributivist and hence posit the principles by which the severity of the punishment should be limited, without providing the justification for punishment; negative retributivism alone only claims that, by having done wrong, the offender forfeits the right not to be punished (Quinton, 1954). The justification *for* punishment must derive from some other theory of punishment, such as consequentialism or positive retributivism. Hence most retributive theories comprise of both positive retributivism, a reason to punish, and negative retributivism, a reason to limit punishment (Walen, 2016). Accordingly, this thesis uses the terms ‘retributive theory’, ‘retributivism’ and ‘retributive punishment’ to refer to this common form of retributivism; one that encompasses both positive and negative components.

Positive retributivism can exist in strong or weak forms; with greater or lesser importance attributed to desert, as the justification for punishment, in comparison to other rationales (Walen, 2016). For example, whereas Moore (1997, p.87) describes retributivism as ‘the view that punishment is justified by the desert of the offender’,

Murphy (2007, p.11) describes retributivism as the view that 'the primary justification for punishing a criminal is that the criminal deserves it'. Since Murphy provides scope for alternative justifications, of which desert is simply 'primary', he describes a slightly weaker form of retributivism. Some forms of positive retributivism also posit a specific reason regarding why the offender deserves punishment; for example, Morris (1968) argues that the offender should pay back society for having gained an unfair advantage in the competition for resources, by violating the implicit moral contract.

Retributivism can be distinguished from vengeance by several factors (Nozick, 1981). First, the vengeful punisher takes pleasure in the act of administering punishment, whereas retributive justice does not require pleasure to be taken in the act of punishment. Second, vengeance lacks any form of negative retributivism and therefore, the principle of proportionality. In respect to this principle, some retributive theories specify how proportionality should be implemented; for example, by claiming that the offender should experience the same degree of harm during punishment as the victim did during the offence (Kant, 1797), or by stating simply that in the practice of real-life sentencing, it is difficult to ever realise a proportional punishment (Morris, 1982).

One theory, known as empirical desert theory, posits the need for criminal justice (CJ) policies to resonate with lay intuitions on the topic of justice; in turn, the CJS can maintain legitimacy and consequently, ensure compliance (Robinson and Darley, 1997). These theorists argue that the CJS must 'develop a reputation as a reliable statement of existing norms' (p.474) and specifically, must be 'seen as giving reliable statements of what is and what is not truly condemnable' (p.477). In respect to proportionality, therefore, the theory requires that the principle is applied in accordance with lay intuitions regarding which offences are the more serious and which offenders

are the more blameworthy. Empirical desert theory provides a strong justification for the focus of the current research into what, for lay people, intuitively constitutes a blameworthy offender. Specifically, if the CJS fails to respond to the neuroscientific questioning of blame, in the manner that lay people do, the CJS might lose legitimacy in the public eye (Greene and Cohen, 2004). Hence it is important to anticipate exactly how lay people respond to neuroscientific questioning of the concept of blame.

In theory, no form of retributivism is incompatible with empirical evidence of neurobiological correlates of criminal behaviour; neurocriminologists could endorse or reject positive and/or negative forms of retributivism. For example, Morse (2004a) endorses the legitimacy of both neurobiological causation and retributivism; specifically, he argues that offenders still deserve to be punished for behaviour that was caused by their brains. This thesis will address the extent to which lay people also endorse this view, known as compatibilism; since it finds (neurobiological) determinism to be compatible with free will, and consequently, desert.

However, there also exists a very different interpretation of the science; one that poses major implications for the legitimacy of positive retributivism. Captured in a seminal paper by Greene and Cohen (2004), this view is known as incompatibilism; it finds (neurobiological) determinism to be incompatible with free will, and consequently, desert. Similar to the compatibilist, Greene and Cohen consider the brain to be the proximal (i.e., the immediate) cause of all behaviour, including criminal behaviour. In turn, the brain is deemed the product of uncontrollable genetic and environmental influences. However, their reasoning subsequently deviates from the compatibilist view: since offenders cannot be held morally responsible for the onset of the factors that caused their criminal actions, offenders cannot be held morally responsible for their criminal actions. On this basis, Greene and Cohen (2004) would

excuse offenders from positively retributive punishment on the grounds that offenders *never deserve* punishment. Importantly, for these authors, the illegitimacy of blame alone invalidates the rationale for positive retributivism. It is assumed that the seriousness of the offence is insufficient to justify punishing the blameless offender on positively retributive grounds. Given the severity of its implications for the CJS, the incompatibilist interpretation of neuroscience has received particular attention in the public eye (Harris, 2012; Raine, 2013)

This thesis is concerned not with the validity of the incompatibilist interpretation, but with the question of whether (or not) lay people will make the incompatibilist interpretation, after learning about neurocriminological ideas. Greene and Cohen (2004) predict that lay people will reason in the same fashion as the incompatibilist following sufficient exposure to the concepts and findings of neuroscience. In other words, public engagement in neuroscience will undermine public support for positive retributivism. In turn, only consequentialist justifications for punishment will remain supported; those are justifications that aim to maximise the beneficial consequences of punishment for the offender and society by reducing the probability of future offending. Examples of consequentialist punishment include deterring, rehabilitating or simply incapacitating the offender or potential future offenders.

Greene and Cohen (2004) assume that a lay person, driven solely by consequentialist concerns, would be less punitive than a lay person driven by retributive (and consequentialist) concerns. First, this assumes that consequentialist justifications for punishment tend to generate less punitive outcomes than retributive justifications. However, this is certainly a questionable assumption, especially since consequentialism does not necessarily encompass any form of negative retributivism, as a means of

limiting punishment. Consequently, one might administer the death penalty for parking offences, if such a policy eliminated parking offences, with benefits to society that outweighed the cost of losing the offenders who had been executed (Greene and Cohen, p.1776). Similarly, a purely consequentialist theory could justify framing innocent people to deter other people from offending or justify faking punishment, since deterrence derives from the fear (rather than the reality) of punishment. Through these examples, it is clear that 'purely consequentialist' policies could be more punitive than retributive policies.

In response, Greene and Cohen (2004) state that none of these 'purely consequentialist' practices would benefit society in reality and hence, do not constitute the form of consequentialism that neurocriminology would leave intact. Great fear would arise from punishments that were visibly disproportionate or that were found to involve framing innocent people for utilitarian gains. Framing the innocent and faking punishments would also lead to corruption and when exposed, loss of the intended deterrence effects. By questioning the societal benefits of 'purely consequentialist' policies in practice, Greene and Cohen clearly endorse limits on consequentialist punishment. Therefore, the authors appear to argue that neurocriminology will replace public support for positive retributivism with public support for consequentialism, restrained by some form of negative retributivism. The authors make no specific predictions about how neurocriminology will impact public support for negative retributivism or consequentialism; only that consequentialism will remain unchallenged. Their focus is positive retributivism and its erosion through the communication of neurocriminology.

By predicting that the communication of neurocriminology will ultimately reduce punitiveness, Greene and Cohen (2004) also ignore the possibility that people

might infer an offender is dangerous simply from learning about neurocriminology (Snead, 2007). If lay people attribute an act of criminal behaviour to the brain and the brain is deemed uncontrollable, then that offender might be considered more likely to reoffend and less likely to respond to rehabilitation. Neurocriminologists would not endorse such an interpretation: the mere presence of neurobiological causation poses no implications for dangerousness (Wright and Cullen, 2012). Nevertheless, it remains possible that lay people make this misinterpretation of the science, and as a result, seek the indefinite detention of 'dangerous' offenders whose criminal actions result from their 'uncontrollable' brains; clearly, a punitive outcome of consequentialism.

Greene and Cohen (2004) claim that the communication of neuroscience could replace public support for positive retributivism with public support for consequentialism; this thesis aims to empirically address the plausibility of this claim. Before embarking on this task, however, it is important to establish the current sensitivity of lay people to retributive and consequentialist justifications for punishment. For neurocriminology to render positively retributive people consequentialist, two points must be true of the status quo: first, people must sufficiently pursue positive retributivism such that neuroscience has a philosophy to erode, and second, people must have, or be capable of acquiring, a sensitivity to consequentialist concerns to substitute their sensitivity to retributive concerns. Though unmentioned by Greene and Cohen (2004), this would seem to be an important premise of their argument.

In particular, one might only expect people to relinquish positive retributivism when a replacement justification for punishment is available. If consequentialist justifications for punishment lack salience, people may conceive there to be only two means of responding to neuroscience; those are defending retributive grounds for

punishment or relinquishing punishment altogether. By failing to consider the third option – the maintenance of punishment on consequentialist grounds –, people may be less willing to relinquish retributive justifications. In fact, if people currently justify punishment on retributive grounds *instead* of consequentialist grounds, there may also exist a confirmation bias for rejecting reasons to end positive retributivism. Hence this chapter also considers how people currently explain crime and justify punishment, without exposure to neuroscience, in order to predict the effect of exposing those beliefs to neuroscientific arguments.

iii. How do lay people explain crime?

In order to predict how lay people will respond to neurobiological attributions for offending, it is important to address how lay people currently explain crime. Lay people typically attribute criminal behaviour to social influences, especially those encountered during childhood (Erskine, 1974; Gajos, Beaver, Gertz and Bratton, 2014). People who have regular contact with offenders also adopt social and dispositional explanations of offending (Stalans and Lurigio, 1990). In contrast, people are far less likely to identify opportunistic causes (Erskine, 1974), even less so neurobiological causes (Gabbidon and Boisvert, 2012) and the least of all genetic causes (Erskine, 1974; Flanagan, 1987; Furnham and Henderson, 1983; Gabbidon and Boisvert, 2012). Therefore, most lay people clearly do not conceive of crime causation in the same manner as the typical neurocriminologist. Not only are social explanations preferred but these are considered *substitutes for* neurobiological explanations. The typical neurocriminologist, in contrast, would judge the brain to be the mechanism of transferring social influence into criminal behaviour; in other words, the mediator of social influence.

Given the clear preference for social explanations, one might question whether lay people could adopt a network perspective of crime causation. Importantly, such a network perspective would be a prerequisite to accurately understanding the role of the brain: most neuroscientists consider the brain to be a channel through which all causes of (criminal) behaviour must pass. In the words of Greene and Cohen (2004, p.1781), the brain is a '*transparent bottleneck*'. While no study has measured the scope for people to grasp this network perspective in the neuroscientific context, research has been conducted in the social sphere. By asking people about the relationships between different causes, Campbell and Muncer (1990) identified two models of crime causation. In one model, people attributed offending to feelings of anger and revenge, which were subsequently attributed to broken families and drug use; in the same model, drug use was attributed to peer pressure. This suggests that lay people do tie influences together into causal chains; in this case, attributing emotions to social factors. This finding therefore provides hope that people *could* tie neuroscientific influences together; in that case, attributing the brain to genes and vicious cycles of detrimental social influences.

The most comprehensive study of public receptivity to advancing biosocial research was conducted by Gajos et al. (2014). Among their representative sample, 64% of participants disagreed (52%) or strongly disagreed (12%) that 'genetic factors are involved in why someone becomes a criminal', while 93% believed that environmental factors were 'more important [than genetic factors] for determining whether someone becomes a criminal' (p.374). Hence the strong environmental bias persists to this day. If people consider neurobiological explanations to be substitutes for social explanations, the social bias might inhibit public receptivity to neurocriminology.

In contrast to the genetic explanation, Gajos et al. (2014, p.374) found that far more participants agreed that 'neurological problems with the brain contribute to someone becoming a criminal'. While this finding might be construed as support for neuroscience, the statement is ambiguous. The participant might infer that most offenders have 'neurological problems'; in other words, that 'neurological problems' are typically necessary for criminal behaviour to result. Alternatively, the participant might infer that 'neurological problems with the brain' would, if present, 'contribute to someone becoming a criminal'; in other words, that 'neurological problems' are typically sufficient to induce criminal behaviour, yet most offenders do *not* have 'neurological problems' and instead, offend for reasons other than the brain. If, by agreeing with this statement, participants expressed support for the latter interpretation, this would not constitute support for the claims of Greene and Cohen (2004): for positive retributivism to end, people would need to believe that every offender had a neurobiological frailty. Therefore, it is important that the empirical component of this thesis clearly and faithfully communicates the claims of neurocriminology.

iv. How do lay people justify punishment?

Researchers address this question by asking participants to sentence the same offender in the same case, yet still, varying some features of the case between participants. As a result, it becomes possible to determine the effect of varying those features on the sentence recommended by the participant. Those features regard either consequentialist concerns (e.g., the estimated probability of the offender reoffending) or retributive concerns (e.g., the extent to which the offender intended to offend). By consequence, the researchers can identify the sensitivity of lay people to the manipulation of consequentialist or retributive concerns. Through this method of experimental study,

one can infer the justifications for punishment that guide sentencing preferences in practice.

The first conclusion of such research is that people are not solely motivated by consequentialist concerns (e.g., Baron and Ritov, 1994). For example, people continue to punish offenders who are physically incapable of reoffending (Carlsmith, Darley and Robinson, 2002) or who are tried and sentenced outside of public view (Aharoni and Fridlund, 2012). Using a real-life example from the US, the Sexually Violent Predator Act should only be used for the consequentialist purpose of incapacitation (Kansas Statute, 1994). Yet, in hypothetical cases, the implementation of the Act remained sensitive to the retributive concern of whether the offender *deserved* a longer sentence (Carlsmith, Monahan and Evans, 2007). This finding suggests that the pursuit of positive retributivism may be sufficiently strong to surface even in contexts that should be strictly consequentialist.

In comparisons of these two major justifications for punishment, people appear to be far more sensitive to retributive concerns than consequentialist concerns (Darley, Carlsmith and Robinson, 2000; Carlsmith et al., 2002; Carlsmith 2008; Carlsmith, Wilson and Gilbert, 2008; Carlsmith and Sood, 2009; Aharoni and Fridlund, 2012). For example, in two particular studies, sentence recommendations *were* influenced by the seriousness of the described offence, yet not by the need to deter a particular offence (Carlsmith et al., 2002) or by the prior convictions of the offender (Darley et al., 2000). In the latter study, participants remained equally sensitive to the retributive concern (and insensitive to the consequentialist concern), regardless of whether participants were instructed to sentence in line with retributivism, in line with consequentialism or without guidance. Such findings demonstrate the persistence of the retributive drive for punishment.

However, this research is always limited by the unavoidable overlap between retributive and consequentialist concerns. For example, although offence seriousness is most obviously a retributive concern, it is still possible to make a consequentialist (rather than retributive) interpretation of this concern; specifically, the consequences of reoffending are greater for more serious offences (Warr, Meier and Erickson, 1983). Nevertheless, the empirical study of lay intuitions clearly favours the dominance of retributive concerns.

Once the retributive goal has been fulfilled, interest in consequentialist concerns may subsequently increase. For example, Carlsmith et al. (2007) found that participants were more likely to administer civil commitment to the more dangerous offender only if the original sentence had been sufficiently punitive. In other words, sensitivity to the consequentialist concern emerged only after the retributive goal had been met. Similarly, people tend to seek consequentialist information about a court case only after all available retributive information has been consulted (Carlsmith, 2006; Keller, Oswald, Stucki and Gollwitzer, 2010). People also withdraw support for a consequentialist policy after discovering that the policy violates the retributive principle of proportionality; that is the tailoring of sentences to offence seriousness and offender blameworthiness (Carlsmith, 2008). Therefore, their priority clearly lies with retributive concerns.

Together, this evidence suggests that lay interest in consequentialist concerns may be conditional on the prior and proper fulfilment of retributivism. This sheds further doubt on the plausibility that consequentialism could *replace* positive retributivism following exposure to neurocriminology. If lay people show interest in consequentialist concerns only after retributive goals have been met, it is difficult to conceive of a society where consequentialist punishment unfolds without retributivism

first in place. Therefore, while Greene and Cohen (2004) expect such a society to develop with the dissemination of neuroscience, the theoretical support for its plausibility is lacking.

v. Conclusion

Most lay people have different views from incompatibilist neurocriminologists about the causes of crime and appropriate justifications for punishing offenders. Most neurocriminologists attribute criminal intentions, desires and behaviours entirely to the brain and in turn, to the genetic and environmental factors that influence brain development across the lifespan. From this perspective, every decision to offend is caused by brain activity. The lay person, in contrast, tends to attribute criminal intentions, desires and behaviours to social influences in the environment. While the neurocriminologist also accommodates social causes, the brain is conceived to mediate their effects and genetic causes continue to be given a prominent position. Since lay people do not currently view the brain as a key cause of criminal behaviour, the questions arise of whether and how the lay person might adopt a neuroscientific attitude towards crime causation (Figure 1).

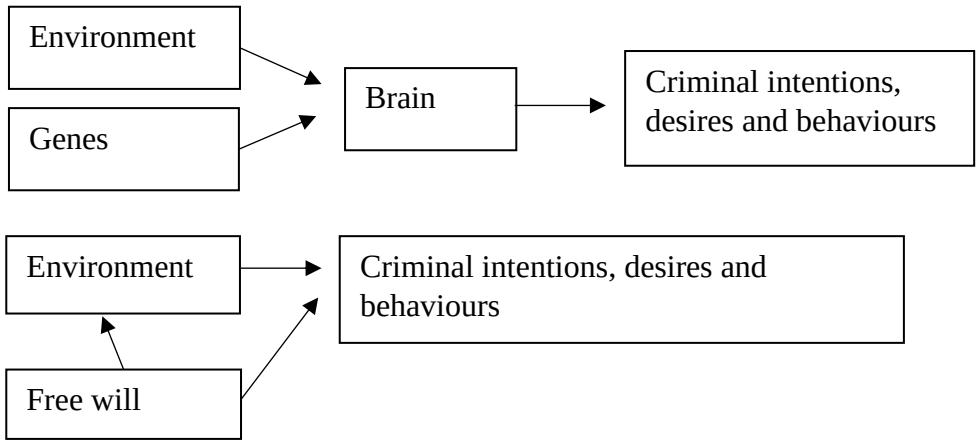


Figure 1. The neuroscientific (top) and lay (bottom) explanations of criminal behaviour.

The incompatibilist neurocriminologist believes that neurobiological chain reactions lead to every decision a person makes, including the decision to offend, and therefore, that no offender can *deserve* punishment. In this respect, positive retributivism is unjustified and only consequentialist grounds for punishment remain valid. In contrast, lay people currently seek to punish offenders primarily on retributive grounds, with little concern for consequentialist grounds. Most lay people and the incompatibilist neurocriminologist currently adopt opposing justifications for punishment, and therefore the questions arise of whether and how the lay person might adopt the neuroscientific attitude towards punishment.

This generates the research questions at the core of chapters five to nine:

- If lay people learnt about neurocriminological ideas, would their support for positive retributivism decrease?
- If lay people learnt about neurocriminological ideas, would their support for consequentialist punishment increase?

Affirmative answers to either of these questions would represent a shift in the attitudes of lay people towards neuroscientific opinion. It is possible, however, that even if people accept neurocriminological ideas about crime causation, there is no subsequent change in attitudes towards punishment. In this case, most lay people and incompatibilist neurocriminologists would still have different views, but about a different point; that is about the relevance of neurobiological causation for sentencing.

Chapter 5: The impact of neurocriminological ideas on attributions of blame and the unconscious: A conceptual framework

According to one dominant interpretation of neurocriminology, offenders do not freely choose to commit crime (Greene and Cohen, 2004). Instead, people offend as a result of unusual characteristics of their brains. While this argument may appear controversial, it is merely the claim that since all parts of the conscious mind are affected by genetic and environmental factors, no choice can be considered truly free. Put simply, scientific explanations of behaviour do not provide scope for a soul in causation (Kolber, 2016). Therefore, while 'there are many causes that impinge on behaviour' (Greene and Cohen, 2004, p.1781), 'all of them…must exert their influence through the brain' and hence, 'all behaviour is caused by our brains' (Raine and Glenn, 2014, p.161).

If people are not truly responsible for their brains, and therefore for their actions, offenders may not be truly responsible for the criminal behaviour that results from their brain. In this respect, neurocriminology may threaten a critical foundation of the retributive justification for punishment: that offenders should be punished in proportion to their blameworthiness for the criminal act (von Hirsch, 1976). The question of concern to this thesis is not whether this reasoning is valid but how lay people reason about neuroscience in this manner. Importantly, the idea that the brain cannot be controlled by some independent free will does not threaten the plausibility of changing behaviour that results from the brain. The probability that a person will offend can still decrease as a result of natural changes in genetic development and deliberate changes to the environmental inputs to brain function (Dar-Nimrod and Heine, 2011).

This chapter begins by discussing the evidence regarding the question of whether or not lay people hold offenders morally responsible for actions that were definitely caused by events beyond their control. In respect to this question, particular

attention will be given to the conditions that render lay people more or less likely to attribute blame to offenders whose actions were predetermined. The chapter then review four mechanisms through which neuroscience could reduce attributions of blame: neuroscience could lead participants to believe the offender was unconscious at the time of offending (the first mechanism), the offender had moral intent but this intent had no causal influence over behaviour (the second mechanism), the offender had the intent to offend but this intent did not cause the offending behaviour (the third mechanism), or that there was some abnormal cause of this offending behaviour (the fourth mechanism). Whereas the theatre study and the online study do not selectively test for one of these four mechanisms, the museum study focused on the fourth mechanism; the perception that there was some abnormal cause of the offending behaviour.

This thesis adopts the stance that neuroscience will not exert the degree of mitigating influence that Greene and Cohen (2004) anticipate. Instead, the communication of neurocriminology will exert weak effects on retributive attitudes towards crime and punishment. In this respect, Greene and Cohen overestimate the extent to which lay attributions of blame vary with the perceived *cause* of the motivation to offend in comparison to the perceived *degree* of motivation to offend. Where neurocriminology does hold influence, the primary mechanism of influence may also differ from the attribution of crime to uncontrollable causes (Greene and Cohen, 2004). Instead, neurocriminology may derive the greatest influence from changing moral judgements of the 'real self', or core character, of the offender.

i. Will neuroscience increase belief in determinism?

Greene and Cohen (2004) hypothesise that the communication of neuroscience will reduce attributions of blame by convincing lay people that determinism is true; that offenders are products of biological and environmental factors that, ultimately, are beyond the control of the individual. Various factors (e.g., the brain, genes and the environment) can be described as playing various roles (e.g., a causal role) in the development of criminal behaviour.

Different studies present lay people with each of these perspectives. I will use particular terms to indicate those perspectives. Some studies describe the role of genetic and environmental factors in criminal behaviour; for those studies, it will be stated that participants learnt about 'the biosocial offender'. Other studies describe the role of the brain; for those studies, it will be stated that participants learnt about 'the neurobiological offender'. Other studies describe the role of uncontrollable factors that are not necessarily specified, yet are explicitly described as having *causal* effects on criminal behaviour; for those studies, it will be stated that participants learnt about 'the caused offender'. According to Greene and Cohen (2004), the biosocial offender, the neurobiological offender and the caused offender are one and the same: any act of criminal behaviour can be traced back to the causal impact of the brain on behaviour, and in turn, the causal impact of genes and the environment on the brain. Nevertheless, the distinction between the biosocial, neurobiological and caused offenders is used to indicate which factors each study described to participants.

Not every study describes these factors as playing a 'causal' role in offending. Nevertheless, Greene and Cohen (2004) predict that people will shift their attributions for offending after learning about even the correlational findings of neuroscience.

Importantly, therefore, these authors predict that lay people will interpret the brain as having a causal impact on behaviour, even when causality is not explicitly stated. The authors also assume that lay people will consider the brain a factor that is, ultimately, beyond the control of the individual. In this respect, Greene and Cohen (2004) hypothesise that public engagement in neuroscience will increase belief in determinism.

Beliefs about causality range from the philosophy of indeterminism to the philosophy of determinism. Neither of these two extremes appears to represent intuitive beliefs about causality; one would not expect lay people to believe we act without reason – indeterminism – or that we act without the ability to choose an alternative course of action given the same sequence of prior events – determinism (Kane, 1999; Nichols, 2011). However, when forced to choose between the two, people are far more likely to believe that choices are entirely unaffected, rather than entirely caused, by prior events – at least in the US (Nichols and Knobe, 2007), China, India and Colombia (Sarkissian et al., 2010). Neuroscience, on the other hand, is premised on proximal determinism; the idea that the immediate cause of every decision, conscious or unconscious, is chemical reactions and electrical activity in the brain (Harris, 2012). Hence the premise of neuroscience poses a challenge to intuitive belief in (incompatibilist) free will.

Neuroscience might challenge two reasons why people intuitively reject determinism and, in this respect, might promote belief in determinism. Those two reasons are limits in introspection and the desire to believe in an uncontestably free form of free will. First, people do not intuitively have conscious access to a complete and accurate explanation of decision-making (Nisbett and Wilson, 1977). In contrast, most neuroscientists are acutely aware of findings that suggest brain mechanisms underlie every decision a person makes. The causal role of the brain has even been

supported by the replicated finding that conscious decisions can be predicted from activity in the frontopolar cortex up to seven seconds before people make the conscious decision to act (Soon et al., 2008; Bode et al., 2011). Therefore, the communication of neuroscience may simply increase public awareness of the brain and in turn, the mitigating implications of neurobiological causation (Shariff et al., 2014).

Second, lay people may believe it is more difficult to assert free will if they first accept determinism. Hence people may resist neuroscientific arguments for determinism, given their belief that determinism and free will are (somewhat) incompatible (Sarkissian et al., 2010) and their strong motivation to believe in free will (Clark et al., 2014). One can, however, envisage two alternative responses to neuroscience: the science could challenge belief in incompatibilism, such that people then accept neurobiological causation without relinquishing belief in free will. Finally, people could maintain incompatibilist beliefs and accept neuroscience, thereby concluding that neurobiological causation challenges free will (Greene and Cohen, 2004). In turn, the concept of free will would no longer be available to rationalise the instinctive desire for positive retributivism. In order to evaluate the plausibility of these three responses, the question will be addressed of how stubbornly people believe that free will cannot exist in a predetermined world; in other words, that determinism and free will are incompatible. In turn, it should be easier to predict whether this incompatibilist position is likely to influence, or be influenced by, the communication of neuroscience.

ii. When do people blame the caused offender?

Although there is evidence that lay people are incompatibilist (Sarkissian et al., 2010), people may recruit incompatibilist beliefs very selectively, such that their belief in

incompatibilism declines with the perceived plausibility of determinism (Shepherd, 2012). Therefore, if determinism appears more plausible in neuroscientific context, Greene and Cohen (2004) may be wrong to expect lay people to remain rigidly incompatibilist. Instead, people may seek to maintain their belief in free will more than their belief in the incompatibility of free will and determinism. In many contexts, therefore, lay people may believe that free will can exist in a predetermined world and hence be compatibilist.

The lay perspective on the debate between compatibilism and incompatibilism is certainly greatly influenced by context (Nichols, 2011). Incompatibilist intuitions dominate in response to *abstract* questions about free will *in general* in *hypothetically* predetermined worlds. In other words, participants tend to believe that 'people' do *not* have free will 'on an imaginary planet' where behaviour is described as an inevitable outcome of prior uncontrollable events. In contrast, compatibilist intuitions dominate when determinism is applied to the real world – rather than a hypothetical world (Nahmias et al., 2007; Roskies and Nichols, 2008) – and when participants judge a specific offender – rather than a specific morally neutral actor, people in general or offenders in general (Nahmias et al., 2007; Nichols and Knobe, 2007). In other words, participants tend to believe that an offender, whose crime has been described in detail, does have free will on Earth; even after being instructed to assume that all behaviour on Earth is the inevitable product of prior uncontrollable events. Therefore, people are likely to believe that free will is compatible with determinism when judging offenders (i.e., immoral actors) in the real world.

It is important to demonstrate the dominance of compatibilist intuitions with specific examples. Nahmias et al. (2005, p.566) found that 76% of participants continued to attribute free will to a specific offender in a world where a

‘supercomputer…can…predict everything [including the judged robbery] …with 100% accuracy…20 years before [the robber] is born’. In fact, Monroe, Brady and Malle (2017) observed no effects of challenging general belief in free will on attributions of wrongfulness, blame or punishment to a specific offender. The latter null effects persisted no matter the offence type, from sunbathing nude to selling drugs, and no matter how free will was challenged, from the assertion of genetic and environmental causation, determinism, a destiny or a supercomputer or brain scanner with infallible predictive powers. The null effects also cannot be attributed to a failure of the experimental manipulation: belief in free will for the self and people in general was successfully challenged, yet the attribution of free will to the specific offender persisted.

In fact, Monroe et al. (2017) conducted a third study that directly compared the sensitivity of lay people to compatibilist and incompatibilist concerns: participants read an explicit challenge to free will, before learning that an actor had decided to steal without coercion or because the outcome of a die roll required him to steal. Only the latter, compatibilist criterion for choice reduced attributions of blame and punishment to the actor. Hence the evidence consistently suggests that lay people are primarily concerned with the capacity of the actor to exercise rational (e.g., uncoerced) choice.

This compatibilist focus on rational choice may be especially prevalent among judgements of the most denigrated offenders. In one study, participants read a case of offending that resulted from a real or hypothetical psychological disorder or process, described either in neurobiological or cognitive terms (De Brigard et al., 2009). The offences were emotive: acts of uncontrollable aggression, watching a prepubescent boy shower or causing serious injury following impaired visual processing or face recognition. De Brigard et al. observed no effect of the neurobiological explanation (relative to the cognitive explanation), even when no disorder was mentioned, and the

causal role of the process was explicitly stated. In fact, the average participant scored the responsibility of *both* the cognitive and neurobiological offenders above the mid-point or, with the stronger manipulation, at the mid-point ('kind of responsible'). This finding casts doubt on the potential for neurocriminology to undermine blame in emotive offences. De Brigard et al. attribute their null effects to the fact that the judged acts were specific and harmful, and the science described as real – contexts that already promote compatibilist responding (e.g., Nahmias et al., 2007). Indeed, the authors did observe an effect of neuroscience when the judged act was ambiguous: participants attributed less moral responsibility to an actor whose neurobiological illness caused him to commit 'a certain behaviour' rather than 'rape' (p.11). This effect may reflect the generality rather than moral neutrality of 'a certain behaviour'.

Compatibilist intuitions have been observed even when there is no immoral outcome of the act and so no blaming function to fulfil. Nahmias et al. (2005) found that the majority of participants attributed responsibility to a morally neutral actor (79%) and even a moral actor (68%). Such findings have led researchers to hypothesise that compatibilist intuitions may be promoted by the specificity (and not just the morality) of the act. For example, when participants are asked to imagine that 'some people make more money than others solely because they have genetic advantages', participants believe that these people do not deserve their money – an incompatibilist response (Freiman and Nichols, 2011, p.127). However, when participants are asked to imagine that a specific musician or juggler makes more money entirely 'because of differences in their genetics' (p.128), people believe the musician and juggler *do* deserve their money – a compatibilist response. Therefore, even though this scenario features aspects that typically facilitate incompatibilist intuitions – namely, a hypothetical world and respectable actions –, the compatibilist response dominated; the

actor was still deemed responsible. Neuroscience derives the greatest scope for influence from appealing to incompatibilist intuitions (Greene and Cohen, 2004). Hence, if learning about a specific offender triggers compatibilist intuitions, the resultant influence of neuroscience may be limited.

While the research generally documents the dominance of compatibilist intuitions in the CJ context, there is one particularly striking inconsistency. Monroe et al. (2017, study 2) found that attributions of blame and punishment were unaffected by exposure to a deterministic passage from *The Astonishing Hypothesis* (Crick, 1994), thereby supporting compatibilism. In contrast, Shariff et al. (2014, study 2, p.3) found that a passage from the same book halved the prison sentenced recommended to an offender who 'beat a man to death', thereby supporting incompatibilism. There are three differences between these studies that might explain the inconsistency: first, although the recommended sentence was halved, this actually corresponded to a one-point difference on a seven-point scale of punishment. Second, unlike Monroe et al., Shariff et al. excluded all consequentialist concerns by stipulating that the offender had already been rehabilitated during two years of prior detention and that further imprisonment would pose no deterrence effect. Therefore, the participants of Monroe et al. retained more (consequentialist) reasons to punish the offender, such as the prosocial effects of communicating censure. These practical reasons to punish may have overridden the neuroscientific reason not to punish, such that neuroscience ultimately had no net effect. Third, although it is unclear from their papers, the authors actually presented different passages from Crick's (1994) book: the passage presented by Monroe et al. (2017, study 2) claimed that decisions were caused by the brain instead of 'beliefs, desires and values', whereas the passage presented by Shariff et al. (2014, study 2) explicitly rejected free will on the grounds that scientists believe free will to be

a neurobiological illusion. Hence the effects observed in the latter study may simply reflect a stronger and more believable manipulation.

The effects of neuroscience are not only sensitive to the type of science or philosophy presented but also to the type of crime judged. Within one study, Krueger et al. (2014) measured both belief in free will and belief in biosocial determinism and on this basis, split their sample into determinists and libertarians (i.e., people who believe in free will because they reject determinism). The determinists decided upon a 10% less severe punishment (than the libertarians) only for the offender who committed the less emotive of two crimes – a surprisingly small and contextually sensitive difference. Therefore, while people make some inference from deterministic beliefs, that inference is surprisingly small, especially when the act provides an incentive to blame. Even Kolber (2016, p.8), who believes lay people to be incompatibilist, recognises that the immorality and specificity of the crime may motivate compatibilist responding: 'It is possible that…judges, facing concrete cases, might find responsibility more durable even in the face of evidence of mechanism'. In combination, the research suggests that people are compatibilist in their judgments of offenders in real life – the context of most relevance to criminology and therefore, this thesis. People continue to attribute moral responsibility to the real offender whose offending is caused by a chain of uncontrollable prior events.

iii. How might neuroscience reduce attributions of blame?

Lay belief in free will is not robust to people learning of every form of determinism. Nahmias et al. (2007) compared the presentation of determinism in the cognitive and neurobiological contexts; all behaviour in the real world was described either as 'completely caused by the specific chemical reactions and neural processes occurring in

our brains' or 'completely caused by the specific thoughts, desires and plans occurring in our minds', which in turn were caused by 'earlier events' (p.224). In comparison to the cognitive description, the neurobiological description of determinism greatly reduced the proportion of people who continued to attribute free will to actors (from 83% to 38%) and blame to offenders (from 86% to 38%). In other words, participants were far less likely to attribute free will and blame to neurobiologically, than cognitively, caused actors. Therefore, neurobiological reductionism – that is the perception of the brain as an unconscious cause of behaviour – poses a particular threat to blame.

In contrast to Nahmias et al. (2007, p.224), Nahmias (2006) specified the nature of the ambiguous 'earlier events' that 'completely caused' the 'chemical reactions and neural processes' or 'thoughts, desires and plans'. Consequently, participants understood not only the proximal cause – 'neural processes' or 'thoughts' – but also the distal cause of behaviour; so the cause of the 'neural processes' or 'thoughts'. In a hypothetical world, 18% of participants attributed free will to people whose 'brain processes' were described as 'completely caused by…their particular genetic make-up and physical environment'. In contrast, 72% of participants attributed free will to people whose 'thoughts, desires and plans' were described as 'completely caused by…their particular genetic make-up and upbringing' (p.231). One interpretation of these results is that people consider the effects of 'upbringing' to implicate conscious thought more than the effects of the 'physical environment'. Therefore, neuroscience may effectively excuse offenders to the extent that the science suggests the distal and proximal causes of offending are unconscious.

Mechanism one: The unconscious offender

Unconscious determinism is the idea that every behaviour has a prior cause, yet none of these causes involves conscious 'thoughts, desires and plans' (Nahmias et al., 2007, p.215). Shepherd (2012) documented a strong effect of describing the philosophy of unconscious determinism by comparing lay people's judgements of worlds where behaviour was caused by neurobiological or cognitive factors, described either as conscious or unconscious processes. The description of these processes as conscious (rather than unconscious) greatly increased attributions of responsibility and free will (to a 66% consensus); both when the scenario emphasised causation and determinism (i.e., the inevitability of the causal chain). In contrast, the neurobiological (rather than cognitive) description reduced attributions of responsibility and free will only when the world was described as causal (rather than deterministic). Hence the severity of determinism (in comparison to causation) appeared to render participants less receptive to yet another reason to excuse actors: the neurobiological nature of causation. While this study suggests that people greatly excuse unconscious actors, it is unclear whether the small independent effect of neuroscience operated via the same mechanism or a different mechanism to the effect of describing the causal processes as unconscious.

Unconscious determinism clearly poses a major threat to belief in moral responsibility by suggesting that offenders have no conscious intent or desire to offend (and so no capacity for rational choice). However, the concept fails to represent how lay people or scientists explain normal human behaviour (Wegner, 2003). In fact, the clearly hypothetical nature of unconscious determinism may only encourage incompatibilist responses, since incompatibilist responses are more common when determinism is applied to clearly hypothetical contexts (Nahmias et al., 2007; Roskies and Nichols, 2008) or is conceived to be a hypothetical philosophy (Shepherd, 2012).

Therefore, while people would not blame the unconscious offender for his criminal actions, this finding is ecologically invalid. Neurocriminologists are not arguing that offenders commit crime unconsciously. It is, therefore, a misinterpretation to construe an argument for neurobiological causation as an argument for the absence of conscious desire or intent.

Mechanism two: The consciously helpless offender

Thus far, it has argued that when people are instructed to assume the brain causes criminal behaviour, they tend to perceive the neurobiological offender as unconscious and so as lacking moral responsibility. However, there also exists a more moderate perspective: people may perceive the neurobiological offender as conscious, yet perceive their conscious intent as causally irrelevant to their behaviour and therefore, as capable of becoming decoupled from their behaviour. In this model, criminal behaviour is caused by unconscious processes that operate independently of conscious desires and beliefs. Perhaps people are more likely to believe that the criminal behaviour of neurobiological offenders simply fails to reflect their moral intentions. In turn, this failure for behaviour and intent to correspond may account for the dramatic reduction in the degree of moral responsibility attributed to the neurobiological offender (Nahmias et al., 2007).

Indeed, Nahmias et al. (2007) found that participants believed the brain of the neurobiological actor bypassed his conscious mind and consequently, that he was less blameworthy for offending. When Nahmias et al. measured perceptions of bypassing, the scale probed the perceived *effect* of conscious mental states on behaviour. However, the mediating role played by these perceptions may actually derive from the fact that, when intent causes behaviour, the intent must be consistent with the behaviour.

Therefore, when people believe intentionality has been bypassed, it may be (falsely) inferred that the behaviour could have become inconsistent with the intent. The lay representation of the bypassed offender may be the offender who intends to act morally, yet whose body continues to commit criminal acts; intent and action are decoupled.

More generally, this bypassing hypothesis has been used to explain why certain descriptions of determinism undermine blame more effectively than others. For example, Murray and Nahmias (2014) argue that in generic judgements of hypothetical worlds, the primary cause of incompatibilist responding is not the unemotive nature of the judged act (Nichols and Knobe, 2007). Instead, the particular version of determinism presented may have led participants to infer bypassing; in other words, to infer that the offence would have been committed even if the beliefs and desires of the offender had been different – a compatibilist logic. Hence certain versions of determinism (e.g., neuroscientific versions) may challenge blame because they imply bypassing, not because people care about the unfree origin of behaviour. In turn, supposed evidence of incompatibilist responding might, in fact, result from a compatibilist concern; the capacity of the offender to exercise rational choice.

Murray and Nahmias (2014) support this theory of the bypassing error with an important finding: certain versions of determinism more strongly challenged the perceived causal force of the conscious mind, which predicted less belief in free will; in other words, perceived bypassing mediated the effects of determinism. In a second study, when the scenario explicitly rejected bypassing, the deterministic passage was less effective at reducing belief in free will. This suggests that reduced attributions of blame actually arise from misconceiving determinism as reductionism and so misconceiving the predetermined offender as consciously helpless.

Mechanism three: The unconsciously caused offender

Thus far this thesis has argued that lay people perceive the neurobiological offender as unconscious or consciously helpless. There remains, however, a third possibility: that the neurobiological offender is deemed conscious and acting in a way that is consistent with intent, yet still undeserving of punishment because that intent was caused by prior unconscious processes. It is possible for intent to be consistent with behaviour, yet causally irrelevant to generating that behaviour: the offender may only ever offend after intending to offend, yet that intent may not cause the criminal act. Instead, unconscious processes may cause both the intent to offend and the criminal behaviour, such that the two outcomes correlate yet are causally unrelated. Although this model of criminal behaviour may lack face validity, in fact, it is an established neuroscientific account of human behaviour (Wegner, 2003).

The account is empirically supported by a particular branch of predictive neuroscience originating with the innovative work of Libet (1985). Libet measured the temporal order of the conscious decision (to move one finger or another) and the readiness potential – the peak in brain activity before a voluntary action is made. Incredibly, Libet was able to predict which finger would be moved *before* the participant became consciously aware of having made the decision (e.g., to move the left finger). These predictions can now be made up to 10 seconds before participants report making their 'free decision' to move one finger or the other (Bode et al., 2011; Soon et al., 2008). While predictive accuracy is currently limited to a mere 59% (compared to chance accuracy of 50%), this figure should increase with developments in neuroimaging technology. The findings are used to support the model that action and the intent to act correlate, yet are causally unrelated, and instead, result from the same brain region (Wegner, 2003).

Predictive neuroscience does pose implications for punitiveness: people recommend a shorter retributive sentence for a violent killer after reading a news article about neuroscientific prediction (Shariff et al., 2014, study 3). Although Shariff et al. did not test the mechanism for this small (but significant) effect, evidence of neuroscientific prediction may imply that the neurobiological offender is either consciously helpless or unconsciously caused. Even when the temporal order of brain and mind is not specified, neurobiological explanations of conscious decisions may be interpreted as unconscious explanations of (subsequently) conscious decisions. Therefore, one mechanism of neuroscientific influence may be their (apparent) suggestion that the decision to offend arrives from unconscious (and so excusable) processes.

In contrast to the perception of the neurobiological offender as unconscious or consciously helpless, the perception of the neurobiological offender as unconsciously caused is a valid interpretation of the science. Since such an interpretation might therefore be communicated to the public in real life, it would be interesting to study its effects. However, it is not the most relevant neuroscience to considerations of crime. The use of neuroimaging technology to predict finger movements is most definitely an example of neuroscience rather than neurocriminology since the behavioural measure is – and foreseeably always will be – far removed from criminal behaviour. One might therefore doubt the plausibility for such neuroscience to exert any other or greater effect than that documented by Shariff et al. (2014, study 3).

Mechanism four: The abnormally caused offender

Even if people accurately perceive the neurobiological actor as unconsciously caused, this causal process cannot be demonstrated in the criminal context. While people may

perceive the neurobiological offender as unconscious or consciously helpless, this represents a misinterpretation of the science. This thesis is primarily concerned with *informed* public engagement in *neurocriminology* rather than basic misunderstandings of neuroscience. It is assumed that in real life, neurocriminology would be communicated in an informed context that aimed to avoid basic misunderstandings. Hence this thesis now considers a fourth mechanism of neuroscientific influence: the perception that the neurobiological offender was caused to offend by some abnormal process. While unconscious processes might be considered abnormal causes of behaviour, abnormal causation is not limited to such processes and also includes infinite causal chains that have no controllable endpoint (Greene and Cohen, 2004). This mechanism therefore requires people to believe that moral responsibility is incompatible with abnormal causes, even when those that implicate conscious intent and desire.

Since people are more likely to believe that the world is indeterministic than deterministic (Sarkissian et al., 2010), deterministic causes of criminal behaviour might be considered abnormal. Neurobiological descriptions of determinism may more effectively convey the philosophy of determinism than cognitive descriptions of determinism. People may more intuitively understand or believe that a decision would be inevitable if it was 'completely caused by specific chemical reactions and neural processes' rather than 'specific thoughts, desires and plans' (Nahmias et al., 2007, p.224). Harris (2012, p.11) claims that people only give compatibilist responses after failing to understand determinism – a failure that is avoided by placing determinism in a neurobiological context; since neuroscience encourages one to picture the causal mechanism as physical quantities of chemicals and electrical activity. Similarly, Greene and Cohen (2004, p.1780-1) argue that:

> Neuroscience can help people appreciate the mechanical nature of human action in a way that bypasses complicated [philosophical] arguments.

According to this perspective, people are incompatibilist when determinism is conveyed successfully (e.g., in the concrete terms of neurobiology). Hence the power of neurobiological terms might result from a stronger understanding of determinism rather than evidence of bypassing (Nahmias et al., 2007).

Indeed, there is evidence to challenge the theory that deterministic scenarios derive their influence from implying bypassing (Murray and Nahmias, 2014). Researchers measure perceived bypassing by asking participants to rate the causal effect of conscious mental states on behaviour (Murray and Nahmias, 2014; Nahmias et al., 2007). It is therefore possible that participants interpret these questions to ask about the *free* causal effect of conscious mental states on behaviour; that is their effect independent of prior causal events, which is hereon termed causal efficacy (Björnsson, 2014). This misinterpretation of the questions may have been especially prevalent among participants who cared more about determinism (when judging responsibility). These participants may have used the measures of bypassing to express their incompatibilist belief; the belief that conscious mental states lacked free causal power in a predetermined world. Put simply, the mediating effect of bypassing may not have been an effect of bypassing but a genuine effect of determinism. In this case, neuroscience may retain the scope to derive influence from increasing belief in determinism.

Björnsson (2014) not only criticises the measure but provides empirical evidence to challenge the bypassing hypothesis by comparing determinism as an explanation of behaviour in general or a specific act. First, 47% of the effect of act specificity was unexplained by causal efficacy ratings, suggesting that even if causal

efficacy did measure bypassing, bypassing was an incomplete explanation of free will judgements. Second, people who judged conscious mental states to *lack* causal efficacy were *more* likely to perceive the conscious mind as a bottleneck or transmitter for other influences:

> When earlier events cause an agent's action, they typically do so by affecting what the agent believes and wants, which in turn causes the agent to act in a certain way (Björnsson, 2014, p.106).

Since the bottleneck concept more accurately captures the notion of bypassing yet correlates *negatively* with the original (causal efficacy) probe of bypassing, perhaps the original probe does not actually measure bypassing. Given that attributions of responsibility explained 60% of the effect of act specificity on the original bypassing ratings, participants do appear to have misinterpreted the original bypassing probe as a reference to the *free* causal effect of conscious mental states. Third, 79% of participants (at least slightly) agreed with the conception of the mind as a bottleneck for other influences, suggesting that the average participant did *not* misinfer bypassing from the deterministic scenario. Fourth, there was no relationship between the refined (bottleneck) measure of bypassing and responsibility judgements, suggesting that bypassing poses no implications for responsibility judgement. In sum, Björnsson (2014) concludes that certain forms of determinism pose greater threats to responsibility not because people more easily mistake these forms of determinism for bypassing (Murray and Nahmias, 2014). Instead, these scenarios better convey the notion of determinism and determinism does carry moral weight in the public eye.

If the findings of Björnsson (2014) replicate in the comparison of neurobiological and cognitive (rather than generic and specific) forms of determinism, one could reinterpret the findings of Nahmias et al. (2007): people may not perceive the neurobiological offender as consciously helpless and even if they did, such a perception

of bypassing may not form the basis of responsibility judgements. Instead, neuroscience might simply prime people to better understand or attend to the causal chain that underlies every behaviour, including criminal behaviour.

A particular theory, known as the Explanation Hypothesis (EH), provides a strong rationale for why people attend to determinism more in the neuroscientific context (Björnsson and Persson, 2013). According to the EH, people attribute moral responsibility to actors when the most salient explanation for behaviour is a unique property of the self. After reading the cognitive description of determinism (Nahmias et al., 2007), two levels of explanation compete for the attention of the participant: the beliefs, desires and intentions of the offender and the prior causes of those conscious mental states. The EH predicts the most salient level to be the immediate conscious mental states, especially for explaining why an individual behaves differently to other people (e.g., why an offender offends). In contrast, since determinism applies to all actors, it may appear unable to explain why this person, and not anyone else, offended. People are also much more familiar with explaining behaviour in terms of immediate conscious mental states because these are easier to simulate and closer to the behavioural event. Hence, after reading the cognitive description of determinism, the most salient explanation for offending is a unique property of the self; consequently, the offender is blamed.

After reading the neurobiological description, in contrast, there is no mention of conscious mental states; therefore, people are not distracted by their default tendency to consider the immediate conscious mental states a better, or more significant, explanation of the behaviour. Consequently, people take more interest in the deterministic properties of the scenario (Björnsson and Persson, 2013). Importantly, although the focus is now the causal background rather than the conscious mind at the

time of the crime, people may still not reject the role of the conscious mind. By analogy, imagine explaining why a house struck by lightning is burning (Björnsson and Persson, 2013). Although a physicist might focus upon the descent of the negative charge to the bottom of the clouds, the lay person might focus upon a more proximal factor; the fact that lightning struck the house. Still, the physicist would *not* deny the role of lightning – the next step in the causal path. Similarly, the neuroscientifically informed person may not deny the role of the conscious mind but simply focus upon an earlier point in causal history.

An important outstanding question is what that earlier point is – a question of key concern to chapter seven (about the 'real self'). For now, it is clear that determinism directs attention to the absence of a factor that normally explains human decision making in particular rather than also human information processing, such as mathematical reasoning, or external physical events, such as a tree falling (Rose and Nichols, 2013; Knobe, 2014). This suggests that certain (e.g., neurobiological) versions of determinism more effectively direct attention towards the absence of mental states that uniquely contribute to human decision making; perhaps moral character or free will, for example. In turn, these states must be considered important justifications for blame.

More generally, the EH could explain why people tend to be compatibilist, yet also show some incompatibilist intuitions in hypothetical, general or morally neutral contexts (Nichols, 2011). In these contexts, people may have less need or motivation to distinguish one actor from another, such that the universal applicability of determinism has greater appeal. Hence the EH would also expect philosophical formations of neuroscience to have greater appeal in these contexts. No study has compared the EH and bypassing as alternative accounts of neuroscientific influence.

iv. Conclusion

I have proposed four mechanisms for the effect of presenting neurobiological determinism. The first two are that neuroscience paints the reductionist picture that offenders are unconscious or that their criminal behaviour betrays their moral intent. Not only do these perceptions represent basic misunderstandings of the science but also real offenders tend to be conscious and (at least somewhat) intentional. The third mechanism – the perception of the neurobiological actor as unconsciously caused – is scientifically valid yet impossible to demonstrate in the criminal context. Consequently, a key focus of this thesis is the fourth mechanism: the perception of the brain as an abnormal (rather than unconscious) cause of criminal intentions and desires.

This chapter also discussed a key reason to believe attributions of blame may be reduced by the perception of the brain as an abnormal cause of criminal behaviour. According to the Explanation Hypothesis, people attribute blame to offenders when the most salient explanation of the criminal behaviour is a unique property of the self – the 'normal' cause of human behaviour (Björnsson and Persson, 2013). If 'brain chemicals and electrical impulses' are not considered a unique property of the self, then (according to the Explanation Hypothesis) lay people will consider the neurobiological offender less blameworthy for his offending behaviour. Therefore, neurocriminology may derive mitigating influence from the perception that in comparison to 'conscious thoughts, desires and plans', 'brain chemicals and electrical impulses' are not a unique property of the self and so not a normal causal of human behaviour. This fourth mechanism of neuroscientific influence is of key concern to the current thesis (and is elaborated upon in chapter seven).

The next chapter proceeds with the findings of the theatre study, the online study and, in chapter eight, the museum study. Neither the theatre study nor the online study tested one particular mechanism of neuroscientific influence over any other. However, the online study did ask participants questions that sought to identify why the presented neuroscience had the particular effects it had. The museum study specifically examined the fourth mechanism of neuroscientific influence; that is the perception of the brain as an abnormal (rather than unconscious) cause of criminal intentions and desires. The focus upon this mechanism was achieved by correcting the misinterpretations of neuroscience that underpin the first two mechanisms of neuroscientific influence.

Chapter 6: The impact of neurocriminological ideas on attributions of blame and the unconscious: The empirical work

The first study conducted for this thesis is characteristic of the psychological approach to conducting public opinion research in this field: participants were exposed to a *correlational* explanation of adolescent behaviour, before judging a specific adolescent offender. Hence this chapter begins by considering the role of neuroscience in shaping attitudes towards adolescent offenders. In particular, the theatre study documents the receptivity of lay people to the message that adolescents are influenced by brain mechanisms that naturally develop, and so change, with age. The chapter then shifts to neurobiological explanations of character and their mitigating potential in the online study: in this study, participants were asked to judge an offender with a controlled or impulsive character, described in either neurobiological or cognitive terms.

In both the theatre study and the online study, neuroscience was presented as an explanation of poor impulse control and, in this respect, a trait that could be conceived to impair the capacity for rational choice. Although the theatre study is characteristic of the psychological approach to this field of research, the mock court approach has also been used to measure the effects of communicating the neuroscience of impaired rationality. Therefore, the findings of this approach will also be discussed here. Indeed, the current chapter begins by discussing prior findings from two relevant approaches – the psychological approach and the mock court approach –, before presenting the empirical findings of the theatre study and the online study.

i. The psychological approach

Under the approach of psychology, Shariff et al. (2014) report four studies documenting the effect of challenging the existence of free will on the sentences recommended by lay

people. The most striking effect was observed in study two (p.3): after reading a short neurobiological explanation of decision making that explicitly rejected the existence of free will, the average participant recommended a sentence of half the length of a control group. This substantial effect size was observed even though the described offender 'beat a man to death' – an emotive offence that would typically invite attributions of blame that were immune to causal arguments in mitigation (Nahmias et al., 2007; Nichols and Knobe, 2007). In the third study by Shariff et al., participants tended to recommend a shorter sentence after reading that conscious decisions could be predicted from prior unconscious brain activity (Soon et al., 2008; Bode et al., 2011).

In study four – the study that most closely represents the theatre study –, participants recommended shorter prison sentences after (compared to before) completing one university term of introductory cognitive neuroscience; in contrast, no change in punitiveness was observed across a term of geography. Shariff et al. asked participants to make sentence recommendations on the assumption that regardless of the recommended sentence, the offender would first be placed in a rehabilitative detention centre for two years, before serving the recommended sentence. Since the offender was described as no longer posing a threat to society after these first two years, sentence recommendations were interpreted as a measure of the desire for retributive punishment. Hence Shariff et al. conclude that arguments challenging the existence of free will reduce the desire for people to punish offenders on retributive grounds. Therefore, at least when presented as a challenge to free will, neuroscience may reduce the retributive drive for punishment.

However, such findings failed to replicate in a remarkably similar and more recent series of studies conducted by Monroe, Brady and Malle (2017). Monroe and colleagues observed no effects of challenging belief in free will on attributions of

wrongfulness, blame or punishment to a specific hypothetical offender. These null effects persisted no matter how free will was challenged, which varied from asserting the truth of genetic and environmental causation, the existence of a supercomputer or brain scanner with infallible predictive powers, or the reality of determinism or a destiny. The null effects persisted for norm violations varying from sunbathing nude to selling drugs.

In comparing Shariff et al. (2014) and Monroe et al. (2017), one might hypothesise that the experimental manipulations in the latter study simply failed to challenge belief in free will. However, this was not the case: belief in free will for the self and for people in general was successfully challenged, even though participants continued to attribute the same degree of free will to the specific offender. Therefore, these generic challenges to free will failed to influence attributions of blame and punishment; instead, these attributions were correlated with perceptions of the *specific offender* as free, intentional and capable of choice. Hence the findings of Monroe et al. suggest that how people ultimately explain behaviour may be irrelevant to their retributive pursuit of punishment.

ii. The mock court approach

In the psychological approach, neuroscience is used to explain a broad category of behaviour and sometimes, even to challenge the universal existence of free will. One might hypothesise that such a grand challenge is difficult for people to accept. In contrast, in the mock approach, neuroscience is used to explain a particular offender with a specific mental condition; typically, a severe mental illness, such as schizophrenia. In this context, therefore, the relevance of neurocriminology appears to

be restricted; for example, to schizophrenic offenders. Hence it may be easier for people to accept neuroscience that has more restricted applicability.

People could interpret evidence of mental disorder in one of two ways: either as suggesting that the offender is undeserving of blame or as suggesting that the offender is unchangeably dangerous, thereby mitigating or aggravating the resultant sentence (Aspinwall et al., 2012; Barth, 2007). The small body of mock court research suggests the mitigating edge of this double-edged sword is sharper. Indeed, when acting as judges in mock court scenarios, people attribute less moral responsibility to the offender whose schizophrenia or psychopathy is described in neurobiological, rather than cognitive, terms – both when the evidence is presented by the defence (Gurley and Marcus, 2008; Schweitzer et al., 2011; Schweitzer and Saks, 2011), by the defence but together with a prosecution who labels the defendant dangerous (Greene and Cahill 2012), and even by the prosecution (Aspinwall et al.; but see Fuss, Dressing and Briken, 2015). The net mitigating effect of neuroscience has also been observed with real judges engaged in mock sentencing (Aspinwall et al.) and an analysis of 800 real cases in which neuroscience has been presented (Denno, 2015). Such findings lend support to Greene and Cohen (2004): people may recognise neurobiological dispositions to offend as undermining the blameworthiness of offenders, unlike cognitive and social dispositions to offend (Dar-Nimrod, Heine, Cheung, and Schaller, 2011). In turn, one might expect neurocriminology to successfully undermine support for positive retributivism.

Effects on verdicts

Under the mock court approach, neuroscience is used to explain a deficit in the capacity to exercise rational choice. If this deficit is sufficiently severe, then it may be difficult

to consider the offender legally responsible for his actions; for example, one could judge the offender innocent by reason of insanity in cases where the offender lacks the brain mechanisms necessary to distinguish right from wrong. Indeed, lay people are more likely to reach an insanity verdict after reading a neuroscientific (rather than clinical or family) description of 'a severe mental disorder that causes [the offender] to act irrationally and be unable to control his behaviour' (Schweitzer and Saks, 2011, p.598).

Note, however, this effect of neuroscience may be limited to defences of insanity. Schweitzer et al. (2011) found that neuroscientific testimony had no effect on the probability of people supporting verdicts in favour of *mens rea* (mental state) defences that made no reference to insanity. In these cases, the offender could distinguish right from wrong, yet suffered a defect that 'prevented him from being able to premeditate and deliberate about, or form the intent to be guilty of, first- or second-degree murder' (p.368). Therefore, neuroscience may hold greater influence when documenting deficits in moral knowledge rather than deficits in intentionality. Since deficits in moral knowledge are rarer, these findings further the idea that neurocriminology holds greater influence in contexts where its application is limited.

Effects on sentencing

Even when neuroscientific testimony successfully mitigates attributions of blame, a reduction in punitiveness may not always follow. There exist other justifications for punishment, such as the desire to communicate censure, that could impede the translation of blame mitigation into sentence mitigation. Hence one might hypothesise that neurocriminology exerts greater influence when other motives for punishment are neither too high nor too low. Greene and Cahill (2012), for example, found that after

reading neuroscientific testimony of psychosis, participants were less likely to sentence a *dangerous* psychotic offender to death (as opposed to life without parole). Yet no such effect was observed when the same offender was deemed unlikely to reoffend. One might attribute the latter null effect to the possibility that there was insufficient prior motivation to sentence the offender to death; in other words, insufficient prior punitiveness for neuroscience to reduce.

The effects of neuroscience on lay sentencing appear to be highly sensitive to other case features. For example, Saks et al. (2014) observed a mitigating effect when the defence presented brain scans of psychopathic offenders; however, an aggravating effect was observed when the *prosecution* presented the same scans. Therefore, the mitigating and aggravating interpretations may be promoted by the defence and prosecution respectively. In contrast, Aspinwall et al. (2012) documented a mitigating effect of imageless neurogenetic evidence of psychopathy, *regardless* of whether the defence or prosecution presented the evidence. In Germany, however, the same evidence had no effect on sentencing (Fuss et al., 2015). One can only conclude that the effects of neuroscience on sentencing vary greatly with other features of the case.

iii. Beyond criminal justice

The mock court research falls against a backdrop of studies into the 'seductive allure' of neuroscience. Beyond the criminal justice context, people find more compelling cognitive explanations of behaviours that are supplemented with circular repetition of the original argument in neuroscientific jargon (Weisberg, Keil, Goodstein, Rawson, and Gray, 2008; Weisberg, Taylor, and Hopkins, 2015). Note this is an effect of neuroscientific description, not of brain images per se (Fernandez-Duque, Evans and Christian, 2015; Gruber and Dickerson, 2012; Hook and Farah, 2013; Schweitzer,

Baker and Risko, 2013; but see McCabe and Castel, 2008; but see Farah and Hook, 2013), as confirmed by studies of 1,971 participants (Michael et al., 2013). Importantly, the appeal of neuroscience cannot be attributed to the addition of jargon or the status of neuroscience (Fernandez-Duque et al., 2015), yet instead represents a broader effect of describing mechanisms in reference to their smaller components (Hopkins, Weisberg and Taylor, 2016). Therefore, the influence of neuroscientific defences in court may represent this broader 'seductive allure' of neuroscientific reductionism.

iv. Genetic and philosophical testimony

It is important to contrast the successful examples of neuroscientific influence documented by the mock court approach with the lacking influence of genetic or philosophical testimony. There appears to be less evidence for the mitigating influence of genetic testimony: despite reducing attributions of conscious control and intentionality (Cheung and Heine, 2015), in several studies, genetic testimony has exerted no effect on attributions of moral responsibility, insanity or punishment (Appelbaum and Scurich, 2014; Appelbaum, Scurich and Raad, 2015; Cheung and Heine, 2015; Scurich and Appelbaum, 2016). Therefore, genetic testimony certainly does not appear to be more effective in its capacity to change retributive attitudes, relative to neuroscientific testimony. Similarly, when philosophers ask participants to assume that offending is entirely caused by uncontrollable factors, strong attributions of blame persist (e.g., Nahmias et al., 2005).

Under the mock court approach, genetic testimony is typically used to explain examples of impulsive offending (e.g., Appelbaum and Scurich, 2014). In contrast, one might speculate that neuroscience is *relatively* successful at changing retributive attitudes because the target of attitudinal influence is typically the psychotic offender

(e.g., Greene and Cahill, 2012). In such cases, neuroscience may successfully cast doubt on the perceived capacity of the offender to make an intentional and desired choice (i.e., a rational choice) rather than to exercise an uncaused and so definitively free choice (i.e., a choice that cannot be attributed to any prior causal influence). Indeed, lay judgements of moral responsibility are highly sensitive to attributions of conscious intentionality (e.g., Alicke, 1992; Shepherd, 2012) and desire (Monroe, Brady and Malle, 2017; Woolfolk, Doris and Darley, 2006). People even tend to spontaneously define free will as the capacity to act upon intentions and desires (Monroe and Malle, 2010). Therefore, neurocriminology may derive the greatest influence from its capacity to document deficits in rational choice, as opposed to the causal origin of such deficits.

Generally, people may be more willing to accept implications of neuroscience that are restricted in scope. The neuroscience of severe deficits in rational choice applies only to offenders with severe deficits in rational choice. In contrast, the neuroscience of choice, in general, applies to every offender (and everyone). Therefore, people may be more receptive to the neuroscience of severe deficits in rational choice – the typical focus of the mock court approach. Indeed, people are more likely to surrender blame in light of deterministic causation when the causal argument is applied to a hypothetical world (Nahmias et al., 2007; Roskies and Nichols, 2008) or when people judge the causal argument to be inapplicable to the real world (Shepherd, 2012). Hence people may be less receptive to neuroscience that appears to have strong applicability. In this respect, the theatre study considers a middle ground; that is the neuroscience of a deficit in rationality that applies to every adolescent.

v. The role of neuroscience in youth justice

In the US, expert testimony of adolescent brain development has informed landmark decisions made by the Supreme Court in respect to youth justice (Feld, 2013; Steinberg, 2013); these include Roper v. Simmons (2005), Graham v. Florida (2010) and Miller v. Alabama (2012). Although the theatre study tested the effects of communicating neuroscience in a theatre rather than a courtroom, these seminal cases provide close parallels to the type of neuroscience and attitudinal measures used in the theatre study.

In the theatre study, neuroscience was communicated through a play that summarised the key conclusions of extensive research into adolescent brain development (Blakemore and Choudhury, 2006):

> The teenage brain is like a car with a great accelerator but terrible brakes. With powerful impulses under poor control, the likely result is a crash (Steinberg cited by Buchen, 2012, p.305).

In scientific terms, the brain region implicated in reward processing – the nucleus accumbens – responds more to rewards in adolescents (than in adults), yet the brain region implicated in self-control – the orbitofrontal cortex – is less responsive (Galvan et al., 2006). Hence adolescents have a greater need for a degree of self-control that in fact, they do not have (Harden and Tucker-Drob, 2011). The argument can therefore be made that youth offending results from a reduced capacity, rather than a reduced willingness, to exercise rational choice during adolescence (Steinberg, 2009, 2012; Steinberg and Scott, 2003). In turn, young offenders should be considered less blameworthy for their criminal acts.

This argument may have particular appeal for lay people. Indeed, people report more punitive attitudes towards young offenders when they consider adolescents and adults to be comparable in respect to moral responsibility (Allen, Trzcinski and Kubiak,

2012), intentionality and naivety (Metcalfe, Pickett and Mancini, 2015). Therefore, if people relate the conclusions of adolescent neuroscience to responsibility, intentionality and naivety, perhaps the science could change attitudes towards young offenders.

vi. The theatre study

This study measured the effects of learning about adolescent brain development on attitudes towards a hypothetical young or adult offender who committed a serious assault three times in three years. All participants judged the offender after the first offence and then after the third offence, and made those judgements either before or after watching Brainstorm – a play about adolescent brain development. Instead of measuring punitiveness, the study focused upon the mechanisms by which neuroscience could influence punitiveness; that is by changing attributions of moral responsibility and dangerousness to the described offender. The study extended prior research by three means: first, by using neuroscience to explain the behaviours of a particular age group – adolescents – rather than everyone or only psychotic offenders, second, by presenting the brain as changing across time, and third, by presenting neuroscience in a far more engaging format than the passages of text used in previous research.

Hypotheses

Three hypotheses were proposed on the basis of two assumptions:

A1: Participants would be more likely to attribute the criminal actions of young offenders to the progress of adolescent brain development after watching the play.

A2: Since the play was about the adolescent brain, the play would only change attitudes towards young, and not also adult, offenders.

The predicted outcomes of these two assumptions generated two hypotheses:

H1: Participants would recommend a higher age of criminal responsibility after watching the play.

H2: Participants would attribute less moral responsibility to a hypothetical young, but not adult, offender after watching the play.

Method

Design

This field experiment adopted a mixed model design, with the persistence of the offending manipulated within groups, and exposure to neuroscience and the age of the described offender manipulated between groups to avoid order effects (Table 2). The between groups manipulation was necessary since it would be impossible to counterbalance exposure to neuroscience and no exposure to neuroscience. All survey questions were answered by pencil and paper. Implied consent was assumed from the decision of the participants to return the survey.

Table 2.

Summary of Conditions for the Theatre Study (n = Number of Participants per Condition).

Participants made judgements:				
	Before:		After:	
	Watching the play			
	Regarding the offender's:			
Regarding:	First offence	Third offence	First offence	Third offence
The young offender	*n* = 200		*n* = 195	
The adult offender	*n* = 146		*n* = 187	

Participants

The 728 participants responded to four questions about crime either before or after watching Brainstorm at the National Theatre in London. The study was conducted around all six performances of the play in 2016; one matinee and five evening performances from Tuesday 29th March to Saturday 2nd April. Given that 1,320 people watched the play across this period, the survey response rate was 55%. It was necessary that participants could complete the survey quickly; hence the survey was very short, and no demographic data could be collected. Anyone aged 18 or older was eligible to participate.

The play

Brainstorm is a play about adolescent brain development, directed by Ned Glasier and Emily Lim. The play was produced by Company Three (formally known as Islington Community Theatre) in collaboration with two cognitive neuroscientists. It was presented at the Park Theatre in January 2015 and subsequently at the National Theatre in July 2015 and March 2016. The play was also adapted for BBC iPlayer. The play script conveys two main messages that the audience might consider relevant to judging young offenders:

- The adolescent brain is not a dysfunctional version of the adult brain. The adolescent brain has evolved to generate behaviours that facilitate development, such as risk-taking and self-consciousness.
- In adolescence, the limbic system – the brain region that renders risk-taking rewarding – develops far more quickly than the prefrontal cortex – the brain region that enables people to form intentions, anticipate consequences and exercise self-control. Hence adolescents can behave like a car without brakes, exhibiting

behavioural displays of emotion that adults would ordinarily inhibit, such as aggression.

Both the full script for the play and excerpts highlighting the key neuroscientific content can be found in Appendix A.

Procedure

For the 90 minutes before the play began, audience members were handed one of two types of envelope upon arrival at the theatre foyer. The envelopes differed in the text printed on the front: the text either instructed participants to open the envelope before watching the play or after watching the play; only the latter type of envelope was sealed. The front and back of an example envelope can be found in Figures A1 and A2 in Appendix A. Each envelope contained a pencil and a survey that described either a hypothetical young offender or a hypothetical adult offender. Although participants were instructed to respond privately, it was anticipated that people might talk to each other about the survey. Hence everyone who arrived at the theatre in the same group received the same type of survey. This was intended to minimise the likelihood of participants learning that there existed different types of survey and how those types differed.

One of each survey type was randomly allocated to participants by six research assistants. Each assistant held a box containing four bags of envelopes – one bag for each survey type – and systematically rotated between the bags in the process of distributing envelopes. One might fear that the process of random allocation was undermined by differences in the response rates between conditions. For example, only participants who were particularly interested in the survey may have completed it *after* the play (when the alternative to completing the survey was going home). By condition,

the response rates were 61% (the young offender before the play), 44% (the adult offender before the play), 59% (the young offender after the play) and 57% (the adult offender after the play). Therefore, the response rate was reduced only for judgements of the adult offender before the play; this alleviates concerns of a reduced response rate after the play.

Survey items

First, participants were asked: 'What do you think should be the age of criminal responsibility – the minimum age at which people can be arrested and charged with a crime?'. The subsequent three questions were oriented around a short vignette about a hypothetical offender called Adam: 'Adam commits a serious assault (a physical attack) against a stranger who has insulted him. Adam is [14 or 44] years old and has never committed a crime before'. Participants were asked 'To what extent do you think Adam is morally responsible for committing this crime?' (1 = *not at all responsible*, 9 = *entirely responsible*). Subsequently, the vignette continued: 'Adam commits another serious assault in similar circumstances at the age of [15 or 45] and then another at the age of [16 or 46]. From 1 to 9, to what extent do you think Adam is morally responsible for committing this crime on the third occasion?'. Finally, participants were instructed to 'imagine that Adam is never caught or punished, and nobody tries to stop his criminal behaviour. What do you think is the probability that Adam will commit another serious assault in his [30s or 60s]?'. The front and back of an example survey can be found in Figures A3 and A4 in Appendix A.

Results

Values that fell more than three times the interquartile range from the mean (Q1: $N = 3$) and missing or illegible responses (Q1: $N = 10$, Q2: $N = 3$, Q3: $N = 4$, Q4: $N = 11$) were excluded from the analyses. The deletion of such outliers had no effect on the results.

The recommended age of criminal responsibility

Participants recommended a significantly higher age of criminal responsibility after the play ($M = 15.89$, $SD = 2.61$) compared to before the play ($M = 15.14$, $SD = 2.51$), t (713) = 3.92, $p < .001$, $d = .29$ (H1). Note that the lowest of these mean recommendations – the mean before watching the play – differed significantly from the current age of criminal responsibility in England and Wales (10), as indicated by a one-sample t-test, t (339) = 37.73, $p < .001$, $d = 4.10$. This indicates a discrepancy between the actual age of criminal responsibility and the age that the public perceives to be most appropriate.

Attributions of moral responsibility

Attributions of moral responsibility were analysed using a mixed-model ANOVA, with the Age of the offender [Young, Adult] and the Play [Before, After] as between-subjects factors, and the offence Number [First, Third] as a within-subjects factor. Participants attributed significantly less moral responsibility to the young offender ($M = 6.52$, $SD = 1.55$) compared to the adult offender ($M = 7.97$, $SD = 1.19$), F (1, 720) = 196.53, $p < .001$, $\eta_p^2 = .21$, to the offender upon his first offence ($M = 6.57$, $SD = 1.93$) compared to his third offence ($M = 7.79$, $SD = 1.54$), F (1, 720) = 485.05, $p < .001$, and to the offender after the play ($M = 7.09$, $SD = 1.64$) compared to before the play ($M = 7.28$, $SD = 1.50$), F (1, 720) = 5.39, $p = .021$, $\eta_p^2 = .007$.

There was no significant three-way interaction (p = .147). However, there were significant two-way interactions, which were explored using simple effects analyses. The first interaction was between Age and Number, F (1, 720) = 93.48, p < .001, η_p^2 = .12. Participants attributed significantly more moral responsibility to the young and adult offender upon their third offence (young: M = 7.37, SD = 1.60, adult: M = 8.30, SD = 1.29) compared to their first offence (young: M = 5.68, SD = 1.78, adult: M = 7.64, SD = 1.49), young: t (393) = 23.87, p < .001, d = 1.00, adult: t (329) = 8.27, p < .001, d = .47. Although the increase in moral responsibility attributed for the third (compared to the first) offence was larger for the young (compared to the adult) offender (Figure 2), the increase remained significant for both the young and adult offender.

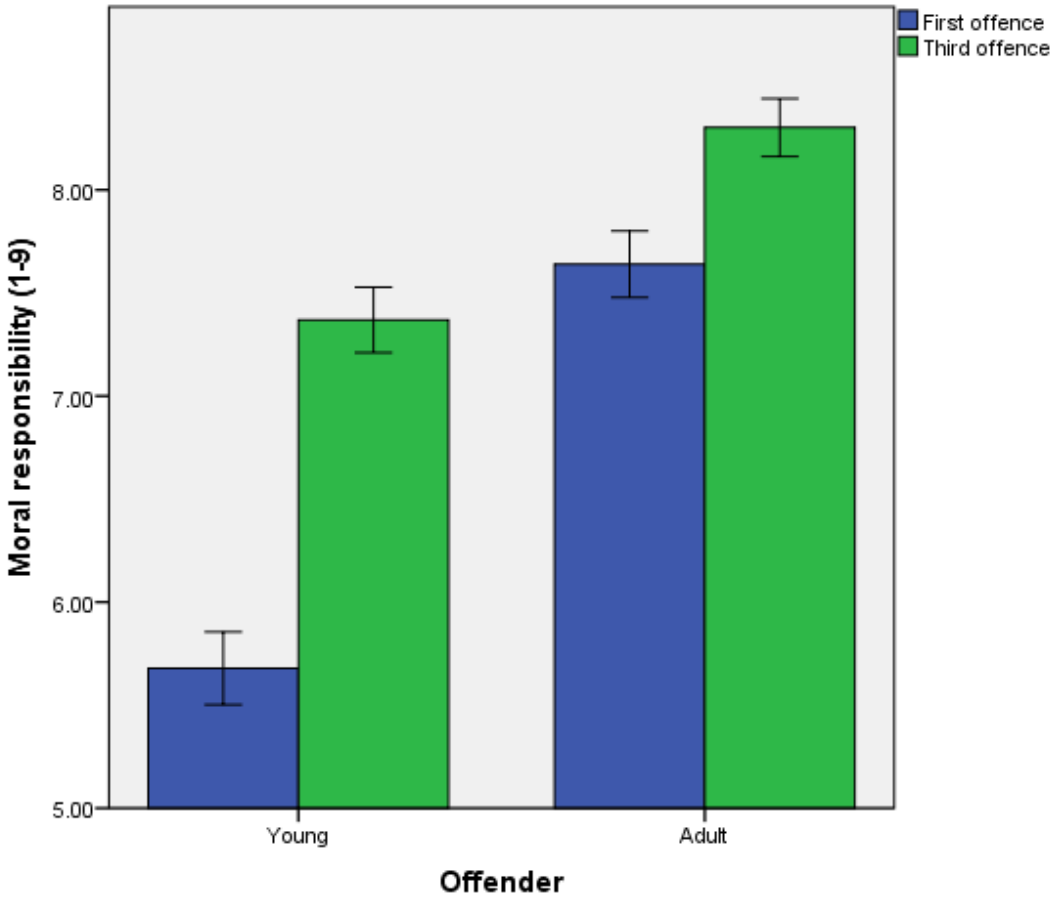


Figure 2. The effect of offence number on mean attributions of moral responsibility to the hypothetical young and adult offenders. Error bars represent the standard error.

The second interaction was between Age and Play, F (1, 720) = 20.49, $p < .001$, $\eta_p^2 = .03$ (Figure 3). Participants attributed significantly less moral responsibility to the young offender after the play (M = 6.15, SD = 1.54) compared to before the play (M = 6.87, SD = 1.48), t (393) = 4.73, $p < .001$, d = .48 (H2). In contrast, there was no significant difference between the attribution of moral responsibility to the adult offender before (M = 7.84, SD = 1.34) and after (M = 8.07, SD = 1.06) the play, t (266.76) = 1.67, p = .097 (equality of variance not assumed, p = .002).

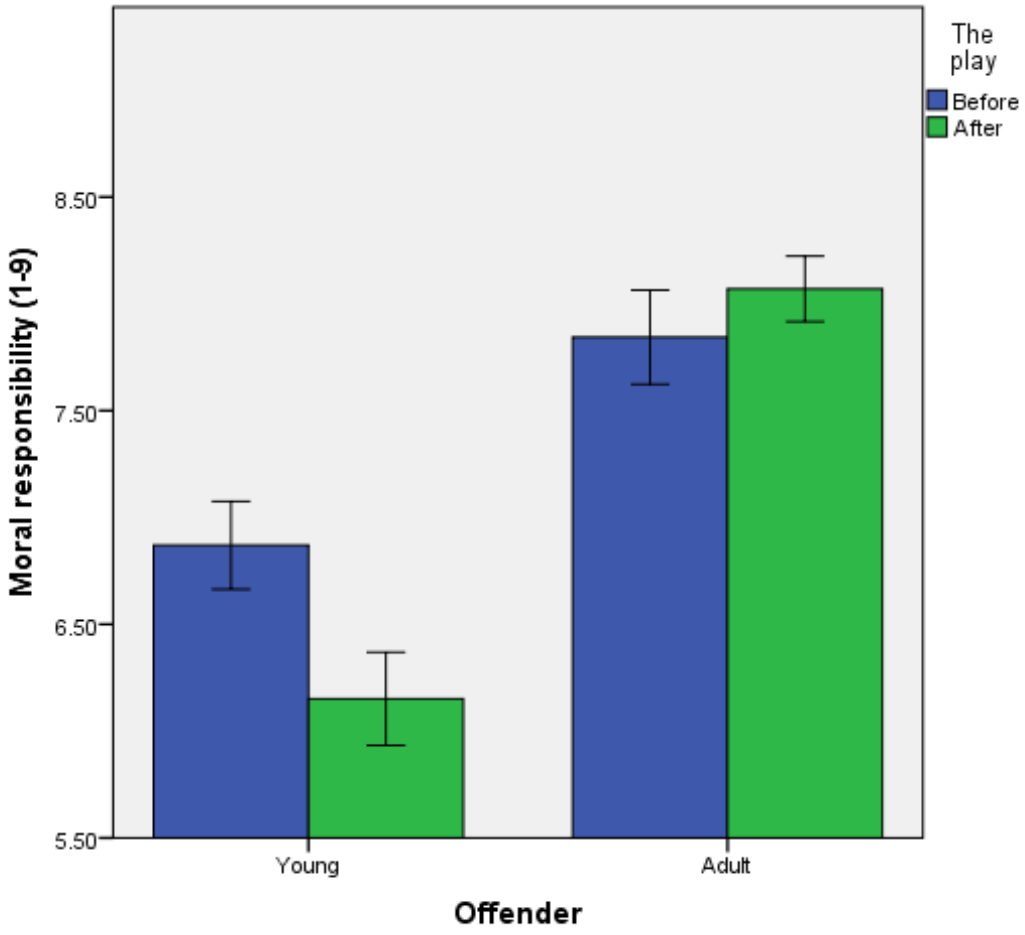


Figure 3. The effect of exposure to neuroscience on mean attributions of moral responsibility to the hypothetical young and adult offenders. Error bars represent the standard error.

The third interaction was between Number and Play, F (1, 720) = 3.91, p = .048, η_p^2 = .005. Participants attributed significantly more moral responsibility to the offender upon his third offence compared to his first offence both before the play (first offence: M = 6.71, SD = 1.81, third offence: M = 7.85, SD = 1.51) and after the play (first offence: M = 6.45, SD = 2.01, third offence: M = 7.75, SD = 1.57), before: t (343) = 14.49, p < .001, d = .68, after: t (379) = 16.13, p < .001, d = .72. Although the increase in moral responsibility attributed for the third (compared to the first) offence was larger after the play (compared to before; Figure 4), the increase remained significant both before and after the play.

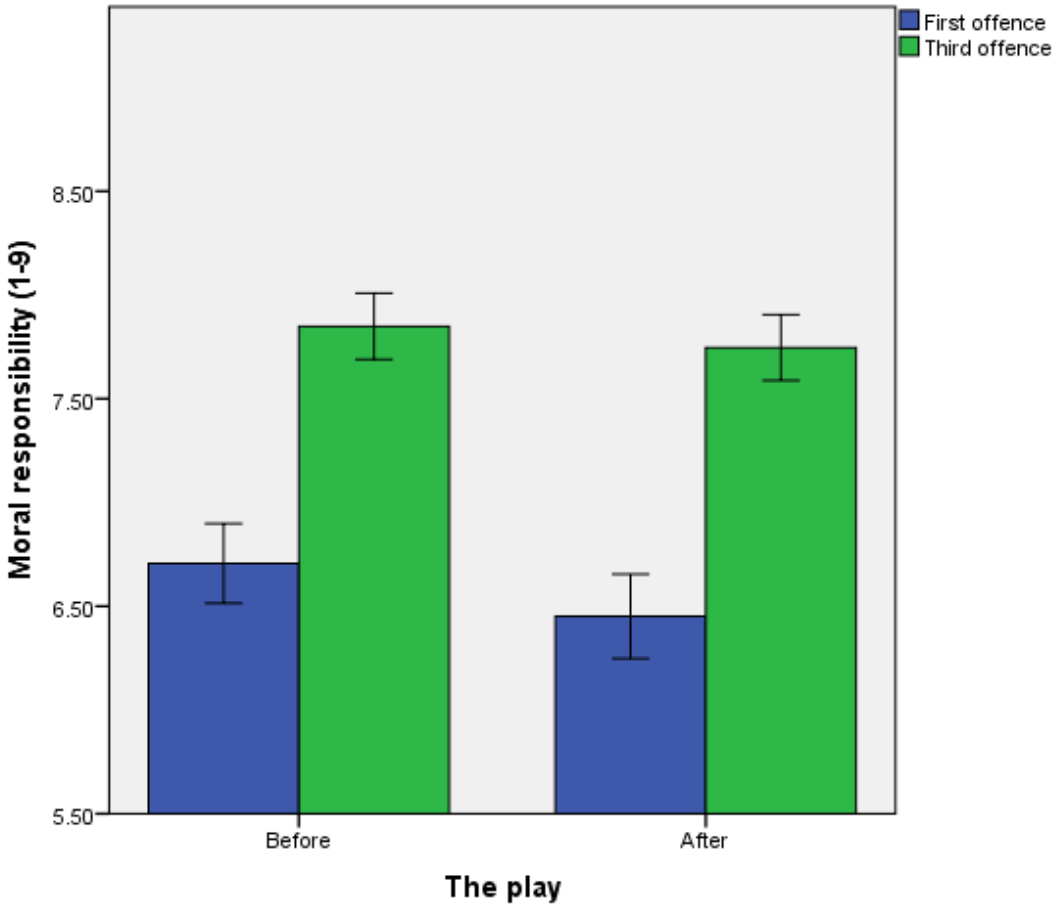


Figure 4. The effect of offence number on mean attributions of moral responsibility by exposure to neuroscience. Error bars represent the standard error.

Summary

The results supported the two hypotheses. Participants recommended a higher age of criminal responsibility after watching the play (H1). Participants attributed less moral responsibility to a hypothetical young, but not adult, offender after watching the play (H2).

Discussion

After watching a play about adolescent brain development, participants recommended a higher age of criminal responsibility (H1) and attributed less moral responsibility to a hypothetical young, but not adult, offender (H2). These findings suggest that the play

reduced the perceived blameworthiness of young offenders – a core component of the retributive justification for punishment (von Hirsch, 1976).

Prior research has tended to present neuroscience either to explain the behaviour of everyone (Nahmias et al., 2007) or to explain the behaviour of a particular offender in relation to a relevant and severe diagnosis of mental illness (Aspinwall et al., 2012; Greene and Cahill, 2012; Gurley and Marcus, 2008; Saks et al., 2014; Schweitzer et al., 2011; Schweitzer and Saks, 2011). This contrasts sharply with the means of presenting neuroscience in the current study: first, neuroscience was used to explain the behaviour of an entire group of people – adolescents – and therefore invited the participants to accept the relevance of neuroscience to this entire group rather than only a rare type of defendant. Second, the relevance of neuroscience was less salient than in the mock court or philosophical research: the neuroscience was not presented in respect to the described offender; instead, the participants spontaneously inferred the relevance of the play to the vignette, given the age of the offender. Third, the brain was explicitly presented as changing across development, thereby countering attributions of stability to brain-based traits (Fernandez-Duque and Schwartz, 2016) and more generally, biology (Dar-Nimrod and Heine, 2011; Miles, 2013). In these three respects, the current study most resembles study 4 of Shariff et al. (2014); the study that found a reduction in retributive sentences after participants completed a module in neuroscience.

Hence, together with Shariff et al. (2014), this study indicates the effects of public engagement in neuroscience beyond the context of mock court cases that fail to represent the average defendant. This study therefore supports Greene and Cohen's (2004) prediction that neuroscience will undermine the attributions of responsibility that justify positive retributivism – yet only to a minimal extent, as the neuroscience exerted only small (but significant) effects.

The participants also did not generalise the presented neuroscience; they did not apply their understanding of the adolescent brain to the brains of adult offenders. The fact that the 44-year-old had refrained from offending for 44 years may have undermined the perceived plausibility that the brain was responsible for the adult offending: if a seemingly stable factor, such as the brain, was to blame, one may have expected the 44-year-old to have offended earlier in his adult life. The outgroup ethnicity of the actors and actresses in the play may have also undermined the receptivity of the audience to neuroscience as an explanation of ingroup behaviour.

Independent of the play, participants attributed greater moral responsibility to the adult, compared to the young, offender (Scott, Reppucci, Antonishak, and DeGennaro, 2006), and for the third, compared to the first, offence. The large effect of offence number was observed for both the young and adult offender, though to a greater extent for the young offender. The latter finding may reflect a ceiling effect: for the repeat adult offender, participants may have reached the maximum responsibility that one would ever attribute to a perpetrator of serious assault (see Alicke, 1992). Repeat offending can be interpreted in one of two ways: either as suggesting that the tendency to act criminally is more automatic or more deliberate; as motivated by influences beyond the conscious control of the offender (von Hirsch, 2009) or as motivated by a stubborn intent to offend despite knowing the circumstances that generate the offending – and therefore the circumstances that could be avoided – and the harm inflicted as a result (Roberts, 2009). The distinction between these two interpretations is critical because the former corresponds to a mitigating and the latter an aggravating interpretation of repeat offending.

The aggravating interpretation prevailed in the current study, corroborating previous evidence that prior convictions increase lay attributions of guilt (Greene and

Dodge, 1995), intentionality and culpability (Kliemann, Young, Scholz and Saxe, 2008), and liability and dangerousness (Sanderson, Zanna and Darley, 2000). The current participants attributed a recidivist premium to repeat offenders, especially when the offender was expected to have matured with age. Indeed, the mitigating influence of youth declines quickly as the age of the offender increases (Scott et al., 2006), for example, from 16 (Hough and Roberts, 2004) to 18 (Roberts and Hough, 2011). Therefore, in the current study, participants may have expected the hypothetical offender to have matured, and so to have gained moral responsibility, from the ages of 14 to 16, but not from the ages of 44 to 46. Hence the repeat *young* offender was attributed greater responsibility.

It is important to note that the same participants judged the first and third offence and therefore, the manipulation of offence number was confounded by age (14 to 16 and 44 to 46). This combined manipulation accurately represents the fact that real judges consider the prior offences of the same offender at an earlier age. Nevertheless, future research is advised to use an independent groups design to manipulate each factor separately and thereby test the effects of prior convictions independent of age.

Since participants were informed that the offending persisted despite the absence of punishment or any other intervention, the recidivist premium documented in this study may reflect the greater perceived intentionality of the repeat offender. While an additional recidivist premium might be observed given evidence that the offender failed to respond to censure (Roberts, 2009), this cannot explain the current findings: the offender was described as receiving no censure – no punishment or other intervention.

While there was a large increase in the attribution of moral responsibility for the third (compared to the first) offence both before and after the play, the increase was greater after the play. According to attribution theory (Heider, 1958), one would expect people to attribute repeat offending to a stable feature of the offender that persisted with age (Carroll, Perkowitz, Lurigio, and Weaver, 1987; Feather and Souter, 2002). While people might ordinarily perceive brain-based traits as stable (Fernandez-Duque and Schwartz, 2016), akin to genes (Dar-Nimrod and Heine, 2011), the play explicitly presented the brain as developing across adolescence and therefore, as unstable. Hence participants may have sought a stable attribution for persistent offending, yet believed the brain failed to provide this type of attribution: when the offender continued to offend despite the development of his brain, it may have become less believable that his (now developed) brain caused his offending. People may also be less willing to excuse the repeat offender, instead seeking to attribute greater intentionality to account for the repeat harms caused (e.g., Alicke, 1992; Roberts, 2009), thereby justifying greater punishment. Such predictions may be of interest to future research.

It is also interesting to note that before watching the play, the recommended age of criminal responsibility (15.26) was far higher than the actual age of criminal responsibility in England and Wales (10). The discrepancy is consistent with academic calls to raise the age of criminal responsibility, as argued from the perspectives of developmental neuroscience (Delmage, 2013) and human rights (Farmer, 2011; Goldson, 2013). The discrepancy may reflect a lay belief in the value of diverting young offenders away from the formal CJS or the genuine belief that even adolescents lack sufficient maturity to be held blameworthy. If people hold the latter belief, future research may wish to establish what form of maturity is considered legally relevant to criminal responsibility (Steinberg and Scott, 2003; von Hirsch, 2001). According to

empirical desert theory (Robinson and Darley, 1997), the legitimacy of the youth justice system may be challenged by this discrepancy between the desired and actual age of criminal responsibility, thereby reducing the incentive to comply with the law. However, this discrepancy may fail to generalise beyond the subset of the population that attended this play – participants who likely sought to understand, rather than simply condemn, adolescent behaviour.

A clear limitation of this study is that there was no control condition; no condition in which participants watched a play about cognitive or social (rather than brain) development across adolescence. Hence one might attribute the observed effects to the mere act of watching a play; that is engaging in a positive social experience. However, it is unclear how this account could explain the specific direction of the observed effects, such as the reduction in responsibility attributed to the young, but not the adult, offender. A more substantive limitation might be the inability to specify which features of the play produced the observed effects; for example, it is possible that the same effects would have resulted from presenting any explanation of adolescent behaviour that induced greater empathy, perspective-taking, awareness or memory of adolescent experiences. Nevertheless, the plausibility of this account is also challenged by lacking evidence that punitiveness is eroded by social explanations of crime, even those that are communicated over many years (Selke, 1980; Giacopassi and Blankenship, 1991; Mackey and Courtright, 2000; Lambert, 2004; Shelley, Waid and Dobbs, 2011; Falco and Martin, 2012; Chen and Einat, 2015).

While this study necessarily lacked the control conditions and mediating measures that would be found in an online experiment, it gained ecological validity from presenting neuroscience in a far more engaging format – a play that participants had chosen to attend. Hence this study extended previous research by considering a

potentially more effective means of eroding punitive intuitions. Yet it is also possible that the attitudes of the current sample were more susceptible to neuroscientific change; that by purchasing tickets for the play, the sample had already expressed a desire to explain adolescent behaviour through neuroscience. Therefore, the current sample may have been unrepresentative on a theoretically important dimension.

In sum, exposure to neuroscience eroded a key basis of punitiveness: the perceived moral responsibility of a hypothetical offender (Carlsmith and Darley, 2008). While this suggests that neuroscience could exert a retributively mitigating influence (Aspinwall et al., 2012), the generalisation of this erosion process beyond young offenders and the first offence was limited. Such limits may reflect the strength of the punitive instinct (Crockett, Özdemir and Fehr, 2014; Fehr and Gächter, 2002; Nadelhoffer, Heshmati and Kaplan, 2013b) and the resultant resistance of punitive intuitions to reasoned change (Garland, 2001).

vii. The online study

The theatre study evidenced no effect of learning about the adolescent brain on attitudes towards adult offenders. Hence the next study – the online study – addressed whether attitudes towards adult offenders could be changed by learning about the neuroscience of character; or more specifically, a dispositional tendency to act impulsively. The offence types presented in the theatre study and the online study were the same – a violent offence committed upon emotive impulse. Yet, whereas the theatre study communicated the neuroscience of *adolescent* impulsivity, the online study communicated the neuroscience of *dispositional* impulsivity.

The online study can be considered part of the mock court approach: participants read a *correlational* explanation of character that directly concerned a

specific offender, before judging that offender. The study addressed a key limitation of previous research: in the mock court approach, participants typically read about the mental condition of an offender that is *consistent* with his offending behaviour. Yet some offenders commit crimes that are inconsistent with their normal characters. Since Greene and Cohen (2004) predict the widespread dissemination of neuroscience, it is important to consider the impact of neuroscience in these inconsistent cases; for example, when a normally controlled actor offends on impulse. The online study compared perceptions of the impulsive offender whose normal character was described either as controlled or as impulsive – and so, either as inconsistent or as consistent with his offending behaviour. In this respect, the online study measured whether a neurobiological explanation of character influenced judgements of offending behaviour.

Is the impulsive actor unable or unwilling to exercise self-control?

Some people's impulsive actions reflect their impulsive characters. When judging this type of person (hereon, 'the impulsive actor'), one could attribute their impulsive disposition to one of two factors: an incapacity to exert self-control for uncontrollable reasons or an unwillingness to exert self-control for controllable reasons (Steinberg, 2009). Given its persistence, the impulsivity either appears to be involuntary and so less deserving of punishment (von Hirsch, 2009), or to be the product of greater intent and so more deserving of punishment, since the offender continues to offend despite gaining knowledge of the causes and consequences of the behaviour (Roberts, 2009). Hence it is possible to adopt either a mitigating or an aggravating interpretation of the stable disposition to commit impulsive acts.

The neuroscientific response

Neuroscience may play a pivotal role in determining which interpretation of stable impulsivity people adopt. The exercise of self-control is also a brain mechanism and brain function is ultimately the product of uncontrollable influences (Steinberg, 2013). Therefore, the degree of self-control used in a particular moment reflects the neurobiological capacity for self-control in that particular moment. Hence the impulsive offender is simply a product of his capacities and for this reason, does not *deserve* punishment.

The legal response

It is legally important to distinguish an unwillingness to exercise rational choice in a particular moment from a general incapacity to exercise rational choice (Goldstein, Morse and Shapiro, 2003); only the latter justifies a reduction in the legal attribution of responsibility for offending (Steinberg, 2012). For example, whereas a psychotic offender could not have avoided hallucinating, an intoxicated offender could have avoided alcohol consumption. Therefore, the psychotic offender lacks the capacity to exercise rational choice, whereas the intoxicated offender only lacks the willingness. Hence there is a far stronger legal argument for excusing the psychotic offender. Accordingly, the neuroscience of impulsivity could derive legal relevance from the assumption that it is documenting the capacity, rather than the motivation, to exercise rational choice.

The intuitive response in theory

When asked to sentence a hypothetical offender, people tailor their sentence recommendations to the capacity of the offender to exercise rational choice, considering both his intentionality (e.g., Darley et al., 2000) and whether that intent was coerced or reflects a genuine desire to offend (e.g., Woolfolk, Doris and Darley, 2006). The online

study tested whether lay judgements also vary with a third, more contested, component of the capacity to exercise rational choice: the capacity to exercise self-control. In this study, the participants received no *direct* information about the *capacity* of the offender to exercise self-control, since such information is also unavailable in real life: self-report, observational, cognitive and neuroscientific measures only provide correlational, not causal, indicators of self-control. Hence, to some extent, it is always a matter of interpretation whether one considers those indicators to reveal the capacities or the motivations of the offender.

In the online study, participants read about two indicators of self-control. First, participants learnt whether the offender *normally* exercised self-control. People attribute dispositional causes to behaviours that persist across time and situations (Kelley, 1973). Hence the stability of the behaviour might be considered one cue that the offender lacked the capacity, rather than the motivation, to exercise self-control. Second, the vignette attributed the disposition of the offender either to himself or to his brain. Neuroscience is premised on physicalism; the idea that everything is represented in physical matter, including all aspects of the mind. From this perspective, there is no difference between the offender and his brain: 'all psychological states are also biological ones' (Monterosso and Schwartz, 2012). In contrast, lay people tend to intuitively adopt the conception that brain and mind are at least partially separate (Bloom, 2004; Forstmann and Burgmer, 2015). Therefore, people may distinguish the brain and the person as sources of self-control and infer that the brain more faithfully represents the *capacity* of the offender to exercise self-control.

There is currently only indirect evidence that people perceive the brain to be a physical manifestation of cognitive capacities rather than motivations. People are less likely to attribute moral responsibility to offenders when the brain is presented as an

explanation of every behaviour (Nahmias et al., 2007) or a specific mental disorder (e.g., Aspinwall et al., 2012). Hence these offenders are presumably less likely to be considered *capable* of rational choice. The traits that people believe to be more attributable to the brain are also the traits that people believe to be better indicators of moral *character* (Fernandez-Duque and Schwartz, 2016). Therefore, people conceive the brain to be the source of stable behaviours. In turn, this conception could result from, or lead to, the perception that the brain houses capacities rather than motivations. Consequently, in the courtroom, neuroscience might be used to strengthen the argument that the offender is *unable*, rather than simply unwilling, to exercise rational choice; the online study tested whether the science would do so successfully.

The intuitive response in practice

In the mock court approach, neuroscience is typically used to explain severe mental illness rather than an impulsive character. Nevertheless, two studies have addressed the impact of describing impulsivity in neurobiological terms, with further studies presenting *genetic* testimony of the trait. The two studies found no effect of neuroscience on sentencing in non-capital cases (Appelbaum et al., 2015; Scurich and Appelbaum, 2016). However, neuroscientific testimony did reduce the probability of participants deciding upon the death penalty in a capital case – an effect that was not found with neuro*genetic* testimony (Appelbaum et al., 2015). Although the authors did not explicitly refer to impulsivity, Monterosso et al. (2005) presented a neurochemical explanation of a violent murder that was disproportionate to the source of provocation. In comparison to child abuse, the neurochemical explanation reduced attributions of blame, voluntariness and punitiveness, also increasing sympathy for the offender; the offending was reconceived as ‘automatic and uncontrollable’ (p.143). Hence neurobiological accounts of impulsivity may have a mitigating effect.

Evidence for the influence of genetic testimony is more limited. Research has found no effect of genetic testimony on sentencing decisions and insanity verdicts in non-capital cases and capital cases (Appelbaum et al., 2015), including attributions of responsibility and punishment for second-degree murder (Appelbaum and Scurich, 2014), and for non-criminal deviance, where one might expect less emotional resistance to mitigation (Scurich and Appelbaum, 2016). Cheung and Heine (2015) compared genetic and environmental explanations of the dispositional tendency to respond violently to provocation. The genetic testimony reduced the perceived extent to which the offender 'had conscious control over his actions' and 'intended to kill the victim' (p.1725) while increasing the perceived applicability of the Insanity and Diminished Capacity defences. However, categorical decisions on the defences, attributions of 'criminal responsibility' and sentences were unaffected by the genetic testimony (studies one and two). Although reduced attributions of 'conscious control' predicted shorter prison sentences, the increased attribution of internal causes and reoffending predicted longer sentences; hence there was no net effect of genetic testimony on sentence length. In their study three, however, attributions of criminal responsibility and recommended sentences were reduced, with the use of a more sensitive sentencing measure.

The findings of Cheung and Heine (2015) indicate that the mitigating potential of biological criminology can be quashed by simultaneous increases in the attribution of dangerousness; in other words, by consequentialist responses to the science. Hence one cannot infer that neuroscience lacks any mitigating potential from the evidence that the ultimate sentence is unaffected by neuroscientific testimony (Appelbaum et al., 2015; Scurich and Appelbaum, 2016). Therefore, the online study measured various retributive concerns regarding the offender in the vignette.

Is a lapse from high self-control willed or caused?

Thus far this thesis has discussed the impulsive actor. However, not only impulsive people commit impulsive acts; people who are normally controlled commit impulsive offences too. The fact that such people normally exercise self-control clearly indicates their general capacity to exercise self-control. Hence one might consider this type of offender able, yet unwilling, to exercise self-control on this particular occasion. In turn, this offender might be attributed complete moral responsibility, regardless of their reason for deviating from self-control on this occasion.

The neuroscientific response

Most neuroscientists would attribute a lapse in controlled behaviour to a lapse in normal brain function; in turn, this deviation in brain activity would be attributed to chance variation or an unusually tempting feature of the situation faced by the actor (Greene and Cohen, 2004). The brain can still account for deviations in behaviour since brain activity also deviates. Therefore, from a neuroscientific perspective, the controlled actor who lapses into impulsive offending also does not *deserve* punishment.

The legal response

One day people may accept that every moral failure is a moral weakness: people fail to exercise self-control because of their brain, and in turn, because of genetic and environmental factors that affect the brain. Even if such a day manifested, it is typically argued that offenders would remain legally responsible for their unfree actions (Morse, 2004a). Offenders are *legally* responsible for their actions so long as their mind is generally capable of rational choice, even if every choice is caused and therefore unfree (Morse, 2006). Since evidence of causation is therefore compatible with legal

attributions of responsibility, the law can be described as compatibilist. Hence neurocriminology does not invalidate the legal justification for positive retributivism.

No statute or case law *explicitly* endorses this legal attribution of responsibility to the unfree, yet rational, offender (Kolber, 2016). On paper, to be held responsible, the law only requires offenders to have sufficient self-control to be capable of acting upon desired intentions (Morse, 2008a). However, law-makers may in the past have simply assumed that the degree of self-control a person exercises is always a free choice and therefore, never reducible to a flash of brain activity, caused by a history of genetic and environmental influences (Kolber, 2016). In other words, people may have traditionally considered self-control to always be a motivation, never a capacity. The founders of retributive punishment, therefore, 'plausibly never intended to punish people who make decisions in the mechanistic manner scientists now take to characterize human choice' (Kolber, 2016, p.3). If evidence of causation was incompatible with legal attributions of responsibility, the law would be described as incompatibilist. In this novel interpretation, neurocriminology does invalidate the legal justification for positive retributivism.

The intuitive response

When the impulsive offender was described as a normally controlled actor (hereon 'the controlled actor'), participants read that either the actor or his brain was responsible for his normally controlled behaviour. When participants read that his brain was responsible, they might also attribute the criminal *lapse* in self-control to the brain. Yet still, this may pose no implications for their attributions of blame and punishment, given the philosophical evidence that lay people consider causation and blame to be compatible (e.g., Nahmias et al., 2005). Hence, even if the lapse is conceived as caused

(rather than willed), the intuitive drive for positive retributivism may remain intact. In this respect, the legal and intuitive responses may be remarkably similar.

In sum, lay attributions of moral responsibility may persist for the controlled actor who *lapses into* impulsivity: his previously controlled behaviour indicates his general capacity to exercise self-control, rendering him (legally and intuitively) responsible for his lapse in self-control. Therefore, it was predicted that participants would blame the controlled actor for offending, regardless of the neuroscientific testimony. In contrast, when the offender is a persistently impulsive actor, participants may be primed to consider the possibility that he lacks the *capacity* to exercise self-control. In this context, people may be sensitised to further capacity cues, such as neuroscientific evidence of impulsivity.

Hypotheses

Participants read about a hypothetical offender, described as committing a violent offence on impulse. In the defence testimony for the case, participants read that the offender was normally impulsive or normally controlled, either in cognitive or neurobiological terms. First, it was hypothesised that the normal character of the offender would be considered relevant to retributive judgements:

H1: Participants would attribute more free choice, fault and punishment to the offender who was normally controlled (rather than normally impulsive).

Second, it was hypothesised that the effect of neuroscientific testimony would be restricted to judgements of the normally impulsive actor:

H2: Participants would be more likely to believe the impulsive offender lacked the capacity (rather than the motivation) to exercise self-control when self-control was described in neurobiological (rather than cognitive) terms.

H3: Participants would attribute less free choice, fault and punishment to the impulsive actor when self-control was described in neurobiological (rather than cognitive) terms.

These three hypotheses predict that, contrary to the claims of Greene and Cohen (2004), there is no gap between the intuitive and legal response to neuroscience. The cause of impulsive offending holds retributive significance only when the prior behaviour of the offender suggests he lacks a general capacity for self-control; in this context, the lay person may be sensitised to neuroscientific testimony.

Method

Design

This study adopted an independent groups design. Two independent variables were manipulated between groups: the normal behaviour of the actor, impulsive or controlled, and the source of this behaviour, the brain or the self (Table 3). Finally, a control condition was included that made no reference to self-control. All questions were answered through the online survey platform, Qualtrics.

Table 3.

Summary of Conditions for the Online Study (n = Number of Participants Per Condition).

Participants judged an offender:		
	Whose disposition was:	
Whose disposition was attributed to:	Impulsive	Controlled
The brain	*n* = 164	*n* = 164
The person	*n* = 162	*n* = 164
An additional control condition provided no information about the disposition of the offender (*n* = 167).		

Participants

Participants were recruited through Prolific Academic (n = 550), earning £0.85 for 10 minutes of participation, and through Viennese schools (n = 271), participating under supervision for free in IT classrooms; in these schools, only students competent in English and above the age of 16 were permitted to take part. Of these 821 participants, 62 (8%) were excluded for incorrectly answering one or more of the questions that checked their basic understanding of the vignette. Of the remaining 759 participants, 55% were female and 43% male (mean age = 30.88, SD = 13.33). The most common nationalities were British (47%), Austrian (25%) and American (18%). The most common highest level of completed education was the bachelor's degree (30%), pre-university college (23%) and high school (18%). Anyone aged 16 or older was eligible to participate.

Procedure

Participants imagined themselves to be members of a jury in court, yet under the instruction to assume that all the presented information was true. The described offender had already been found guilty; the remaining task was to decide upon an appropriate sentence. First, the participants read an overview of the case, testimony from the prosecutor and testimony from the defence. The offender had attacked his neighbour with a hammer, causing moderate brain injury, after the neighbour's lawnmower launched a stone into the offender's window, cracking it. The offence was described as a 'rage attack', also provoked by the offender discovering flirty text messages from his neighbour on his girlfriend's phone. A full description of the offence can be found in Appendix B, based on Stimulus C used by Schweitzer et al. (2011). To test their comprehension of the case, participants answered basic multiple-choice questions after every page of description.

Participants were randomly allocated to read one of five different defence testimonies: one control testimony and four experimental testimonies. The control testimony stated simply that the offender 'admitted his guilt at the earliest opportunity'. The four experimental testimonies additionally described the normal behaviour of the offender as either controlled or impulsive, in neurobiological or cognitive terms. The exact text read as follows, with the italicised words manipulated between the four experimental conditions:

> *Parts of the brain called the frontal brain lobes/We* normally exercise control over our impulses. An expert psychologist conducted a series of tests on [the offender] to measure how much control *his frontal brain lobes/he* normally *exercise/exercises* over his impulses. // The tests indicated that *[the offender]'s frontal brain lobes/ [the offender]* normally *use/uses as much/half as much oxygen/inhibition* as *the average person's frontal brain lobes/the average person*, meaning that [the offender] normally exercises *the same amount/half the amount* of control over his impulses as the average person.

These tests were described as 99% accurate and impossible for the offender to fake.

After reading all the case information, participants decided upon a sentence for the offender; the maximum possible sentence was 16 years, which was described as the actual maximum for this type of offence in England and Wales. This question was then repeated, with participants also given the opportunity to sentence the offender to an additional 16 years in secure accommodation; the purpose of the additional time was described as rehabilitation and incapacitation. Participants then indicated their agreement (1 = *strongly disagree*, 6 = *strongly agree*) that: a) the decision to offend was the offender's fault, b) the offender would likely reoffend, c) the offender could be reformed, and d) the offender could be deterred. Participants who read about the impulsive actor additionally indicated their agreement that: a) it was the offender's fault that he exerted only half the normal amount of self-control, b) he was consciously

aware that he exerted only half the normal amount of self-control, and c) it would take him a lot more effort than the average person to exert the normal degree of self-control.

Next, every participant indicated their agreement that: a) 'from the moment the stone cracked the window, [the offender] could have resisted the impulse to attack', b) the offender made a conscious decision to offend, c) 'the offender did not really want to attack [the victim]' and d) '[the offender]'s decision to attack [the victim] reflected his personality'. Subsequently, participants judged how important (1 = *very unimportant*, 6 = *very important*) each of the following factors was 'in causing' the offender to offend: his free choice, evil intent, his brain, his conscious mind, unconscious processes, his emotions, his social life experiences and his genes. Participants who read about the impulsive actor also indicated their agreement that the government should introduce a certain new policy; this policy would require anyone who used less than half the normal amount of self-control to live in secure accommodation until the detained person started to use more than half.

The final set of questions measured belief in biosocial (genetic and environmental) determinism and neurobiological determinism. Specifically, participants indicated their agreement that the offender would have offended, first, if he had the exact same genes and life experiences as the participant, and second, if he had the exact same brain as the participant. Participants also answered these two questions in the reverse direction; reporting their agreement that *they* would have offended, first, if they had the exact same genes and life experiences as the offender, and second, if they had the exact same brain as the offender. Lastly, participants indicated the country that represented their nationality, their gender, their age and their highest completed level of education.

Results

Perceptions of the offender were analysed using a between-subjects ANOVA, with the Disposition of the offender [Impulsive, Controlled] and its Source [the Self, the Brain] as between-subjects factors; the control condition was not included in this analysis. There were no significant effects on the recommended sentence; Table 4 shows the mean sentences by condition.

Table 4.

Mean Sentence by Experimental Condition (Years).

	Controlled Actor	Impulsive Actor
Cognitive Testimony	8.31	8.41
Neurobiological Testimony	8.62	8.50

Effects of the disposition

Table 5 shows the main effects of the Disposition. Participants were significantly more likely to: attribute fault to the controlled actor (Controlled: $M = 5.82$, $SD = 1.42$; Impulsive: $M = 5.51$, $SD = 1.46$), expect the controlled actor could have resisted the impulse to offend (Controlled: $M = 5.02$, $SD = 1.07$; Impulsive: $M = 4.58$, $SD = 1.07$), and expect the controlled actor could be reformed (Controlled: $M = 4.48$, $SD = 1.02$; Impulsive: $M = 4.19$, $SD = 1.07$). Participants were also less likely to expect the controlled actor to reoffend (Controlled: $M = 4.00$, $SD = 1.14$; Impulsive: $M = 4.49$, $SD = 1.12$).

Table 5.

ANOVA Results: Effects of the Disposition.

	F	*df*	*P*	η_p^2
Resist	25.09	1, 593	< .001	.042
Fault	6.63	1, 593	.010	.011
Reoffend	28.29	1, 593	< .001	.046
Reformed	11.57	1, 593	.001	.019

Causal attributions

Table 6 shows the main effects of the Disposition on causal attributions for the described offending. In respect to the controlled actor, participants were significantly more likely to attribute his offending to his free choice (Controlled: *M* = 5.65, *SD* = 1.05; Impulsive: *M* = 5.45, *SD* = 1.18) and social life experiences (Controlled: *M* = 5.74, *SD* = 1.10; Impulsive: *M* = 5.45, *SD* = 1.24), yet less likely to attribute his offending to his genes (Controlled: *M* = 3.90, *SD* = 1.29; Impulsive: *M* = 4.50, *SD* = 1.40) and brain (Controlled: *M* = 5.44, *SD* = 1.13; Impulsive: *M* = 5.80, *SD* = .98). This effect on the importance attributed to the brain also interacted with the Source of the disposition (Table 6): participants were more likely to attribute the offending of the impulsive actor to the brain only when self-control was attributed to the brain (Controlled: *M* = 5.30, *SD* = 1.15; Impulsive: *M* = 5.90, *SD* = .90; Figure 5). Finally, participants were less likely to expect to offend in the genetic and environmental position of the controlled actor (Controlled: *M* = 2.84, *SD* = 1.45) than that of the impulsive actor (Impulsive: *M* = 3.16, *SD* = 1.38).

Table 6.

ANOVA Results: Effects of the Disposition on Causal Attributions.

	F	*df*	*P*	η_p^2
Free choice	4.85	1, 593	.028	.008
Social life experiences	9.36	1, 593	.002	.016
Genes	29.21	1, 593	< .001	.047
The brain	16.91	1, 593	< .001	.028
The brain (source*disposition)	8.12	1, 593	.005	.014
The same genetic and environmental position	7.18	1, 593	.008	.012

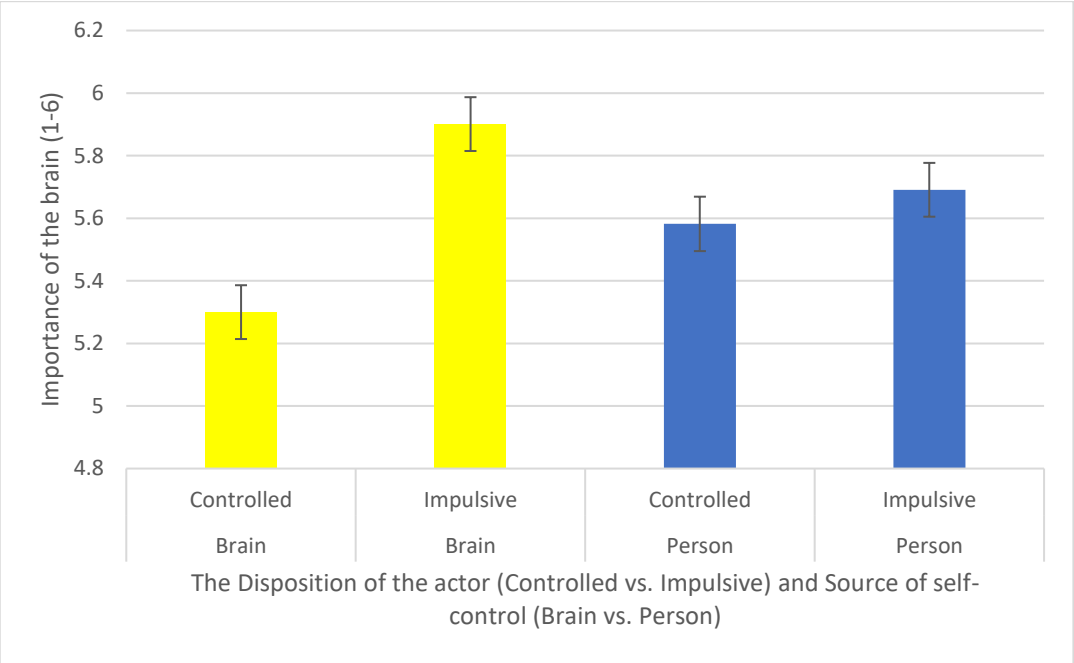


Figure 5. The effects of the Disposition of the actor (Controlled vs. Impulsive) and the Source of self-control (Brain vs. Person) on the perceived importance of the brain in causing the offending behaviour. Error bars represent the standard error.

Mediation

Mediation analysis was conducted using the PROCESS plug-in for SPSS (Hayes, 2013). There were no indirect effects of the Disposition on fault via the causal importance ascribed to social life experiences (p = .316), genes (p = .772) or the brain (p = .368); since none of these variables predicted fault. There were significant indirect

effects of the Disposition on fault via the participant's belief that he/she would offend in the same genetic and environmental position as the offender (β = .06, *SE* = .03, 95% CI [.02, .13]) and via the causal importance ascribed to free choice (β = .07, *SE* = .03, 95% CI [.01, .13]). There was no direct effect of the Disposition on fault (*p* = .144). Therefore, the greater attribution of fault to the controlled actor was mediated by the participant's belief that he/she would not offend in the same genetic and environmental position and that free choice was more important 'in causing' the controlled actor to offend.

Effects of the Source

Table 7 shows the main effects of the described Source of self-control. When self-control was attributed to the brain, participants were significantly more likely to attribute the offending to the conscious mind (Brain: *M* = 5.55, *SD* = .98; Self: *M* = 5.36, *SD* = 1.20) and less likely to believe the offender did not really want to offend (Brain: *M* = 2.48, *SD* = 1.02; Self: *M* = 2.77, *SD* = 1.21).

Table 7.

ANOVA Results: Effects of the Source.

	F	*df*	*P*	η_p^2
Conscious Mind	4.37	1, 593	.037	.007
Did Not Really Want	9.93	1, 593	.002	.016

Fault attributed for the Impulsive Disposition

Independent-groups t-tests were conducted on the four items that were only asked for the impulsive actor. Participants were less likely to believe that it was the offender's fault that he only used half the normal amount of self-control after reading the neurobiological description (*M* = 2.50, *SD* = 1.16) compared to the cognitive description (*M* = 3.23, *SD* = 1.29), *t* (299) = 5.15, *p* < .001, *d* = .60.

The control condition

Independent-groups t-tests were conducted to compare the control and experimental conditions. When described in neurobiological terms, people believed the impulsive actor was more likely to reoffend (t (305) = 3.73, $p < .001$; Exp: $M = 4.59$, $SD = 1.09$; Control: $M = 4.10$, $SD = 1.21$), to have offended because of his genes (t (305) = 4.15, $p < .001$; Exp: $M = 4.66$, $SD = 1.41$; Control: $M = 4.01$, $SD = 1.35$) and his brain (t (285.75) = 4.06, $p < .001$; Exp: $M = 5.90$, $SD = .90$; Control: $M = 5.41$, $SD = 1.19$), and less likely to have the capacity to resist offending (t (305) = 3.24, $p = .001$; Exp: $M = 4.52$, $SD = 1.07$; Control: $M = 4.92$, $SD = 1.11$).

Summary

The results provided some support for two of the three hypotheses. Participants attributed more free choice and fault, but not a longer sentence, to the offender who was normally controlled (rather than normally impulsive; H1). On one item, participants were more likely to believe the impulsive offender lacked the capacity (rather than the motivation) to exercise self-control when self-control was described in neurobiological (rather than cognitive) terms (H2). However, participants did *not* attribute less free choice, fault and punishment to the impulsive actor when self-control was described in neurobiological (rather than cognitive) terms (H3).

Discussion

People who normally exercise self-control also behave impulsively sometimes – and occasionally that impulsive behaviour is criminal. In this study, participants read about a violent offence, committed either by a normally impulsive actor or by a normally controlled actor who experienced a lapse in self-control. The participants were more likely to attribute the criminal actions of the controlled actor to free choice *and*

therefore attribute fault to this offender (H1). Perhaps these participants believed that the controlled actor had freely chosen to deviate from his clearly demonstrated *capacity* for self-control. This finding suggests that lay attributions of blame are informed by the (impulsive or controlled) character of offenders, not only their degree of intentionality (Darley et al., 2000; Carlsmith et al., 2002; Carlsmith 2008; Carlsmith et al., 2008; Carlsmith and Sood 2009; Aharoni and Fridlund 2012) and their desire to offend (Woolfolk et al., 2006).

How did people explain the offending of the impulsive and controlled actors?

After reading the neuroscientific testimony, participants were more likely to attribute the actions of the impulsive (compared to the controlled or unexplained) actor to his brain and (irrespective of neuroscience) to his genes. This suggests the participants more readily considered biology capable of explaining a stable disposition than a lapse from that disposition (Dar-Nimrod and Heine, 2011); this is perhaps unsurprising since the vignette only explicitly presented the brain as an explanation of stable disposition. In contrast, the participants were more likely to attribute the actions of the controlled (compared to the impulsive) actor to social life experiences. This suggests the participants more readily considered social influences capable of explaining a deviation from the actor's stable character.

From the perspective of neuroscience, social life experiences can only induce impulsive behaviour by first influencing brain activity – the 'bottleneck' through which every influence on behaviour must pass (Greene and Cohen, 2004, p.1781). In contrast, the current participants perceived the brain and social influences as explanations of character and a lapse in character respectively. Not only is there no scientific basis to this perception of the brain and social influences as in opposition, but the distinction also suggests that people believe in some other mediator of social influence, most likely

an independent mind. In this respect, the current study is consistent with prior evidence that people perceive the brain and mind to be separate (Bloom, 2004; Forstmann and Burgmer, 2015). The current study suggests that people also distinguish the factors that inform the basis of brain and mind.

However, the difference in attributions for offending was unable to explain why people attributed more fault to the controlled (than the impulsive) actor. In other words, participants did not derive their attributions of fault from their explanations for the offending. Instead, the difference in fault was explained by two factors: first, the belief that the offending was the inevitable outcome of causal factors (i.e., belief in determinism) and second, the attribution of 'free choice' to the offender. In respect to determinism, participants were less likely to perceive the lapse (compared to the disposition) as an *inevitable* product of genetic and environmental influences. Therefore, these participants must have been more likely to attribute the offending of the controlled actor to some third factor that could break the inevitability of causation – presumably an incompatibilist free will. In turn, this effect partially explained why the controlled actor was blamed more, independent of attributed 'free choice'. Hence participants recognised the relevance not of causation but of the *inevitability* of causation, independent of attributed 'free choice'. Since the participants therefore considered determinism to be somewhat relevant to blame, their attributions were somewhat incompatibilist.

The greater attribution of fault to the controlled actor was also explained by the greater attribution of 'free choice' independent of causal or deterministic beliefs. Hence the retributive significance of 'free choice' may not lie in causal or deterministic beliefs, but perhaps instead, in the perceived capacity of the offender to exercise *rational* choice. The participants did not consider the controlled actor more conscious

of his choice to offend or the consequences of that choice; therefore, it is unlikely that 'free choice' was equated to intentionality (Nahmias et al., 2007; Shepherd, 2012). Instead, the attribution of 'free choice' might regard *which part* of the conscious mind the impulse was believed to have bypassed:

> Participants must have thought that even if the agents' conscious mental states weren't bypassed, the agents themselves were; the causally relevant mental states [were] (De Brigard, Mandelbaum, and Ripley, 2009, p.10).

The current study proposes that one such 'causally relevant mental state' is self-control; more specifically, the capacity to be moved by, and therefore act upon, reason, where reason includes anticipated harm.

The participants may have considered the controlled actor to have a greater capacity to be moved by reason; indeed, he must have been sufficiently moved by reason in the past to exhibit controlled behaviour. Therefore, while both offenders wanted to offend, the controlled actor was able to place his desire in the context of far more experience of inhibiting primitive desires. Hence the controlled actor could better appreciate the difference between impulsive and considered action and thereby know that this act was excessively impulsive – knowledge that could move the offender to obey the law. Since the offender nevertheless offended, his response appears particularly disproportional to the source of provocation. Legal attributions of responsibility require the offender to know right from wrong (Morse, 2004b), yet the current study provides a new avenue for research: people might also consider the capacity of the offender to know impulsive from considered action and so, disproportional from proportional responses.

Is a lapse from high self-control willed or caused?

The previously controlled behaviour of the controlled actor documents his capacity to exercise rational choice. Since this type of offender clearly meets the legal criterion for responsibility (Morse, 2004b), it was expected that participants would no longer be concerned by the *cause* of the lapse in self-control and therefore the neuroscientific testimony (H3). Indeed, there was no effect of neuroscientific testimony. However, this finding cannot be used to suggest people consider neurobiological causation to be irrelevant to responsibility since the participants were no more likely to make a neurobiological attribution for the lapse after reading neuroscience. In other words, the participants did not consider the lapse in normally controlled behaviour to be a lapse caused by brain activity. Presumably, instead, participants believed the controlled actor overrode a default biological tendency to be controlled. It therefore remains a task for future research to assess how people would respond to evidence that the *lapse*, and not only the normal functioning, of self-control was caused by the brain.

Is the impulsive actor unable or unwilling to exercise self-control?

In contrast to the controlled actor, it was predicted that participants would care far more about *why* the impulsive actor offended. The stability of his impulsive behaviour may have primed participants to consider the possibility that the offender lacked the *capacity* to exercise self-control. In turn, neuroscientific testimony may have confirmed those suspicions (H2). Indeed, after reading the neuroscientific testimony, participants were less likely to believe it was the offender's fault that he used only half the normal amount of self-control. However, there was no subsequent reduction in attributions of choice, blame or punishment to the impulsive actor (H3). Therefore, although participants did consider brain activity to represent the capacity, rather than the motivation, to exercise self-control, this distinction between capacity and motivation

was deemed morally irrelevant. This perspective represents a highly compatibilist position: the reduced attribution of fault for an impulsive character failed to manifest in a reduced attribution of fault for impulsive offending. Though unexpected, this finding does generalise the results of Appelbaum and Scurich (2014) to neuroscience: their description of an impulsive character *in genetic terms* also failed to reduce attributions of moral responsibility to an impulsive offender.

In contrast to prior research, the current vignette explicitly described the stability of the impulsive behaviour. Since a stable behaviour indicates the possibility of a general incapacity (von Hirsch, 2009), there is a stronger legal argument to examine the source of a stable behaviour, as a means of confirming or disconfirming the possibility of incapacity. Although 'the distinction between didn't and couldn't is important under the law' (Steinberg, 2012, p.75) and the current participants made this distinction, it was not ascribed moral significance. This suggests that people may be even more compatibilist than the law, challenging the claim that there exists a 'gap between what the law officially cares about and what people really care about' (Greene and Cohen, 2004, p.1776). The current participants cared even more exclusively about 'what the law officially cares about'; that being the capacity to exercise rational choice.

The conclusion is also consistent with the work of experimental philosophy: lay people continue to attribute moral responsibility to actors who develop the desire to deviate from moral decision making for uncontrollable reasons (e.g., Nahmias et al., 2005). People are unconcerned by the ultimate cause of the desire to offend so long as the desire is conscious (Nahmias 2006; Nahmias et al., 2007; Shepherd 2012). In the philosophical research, participants are instructed to assume that determinism is true. Yet, even without this instruction, Shariff et al. (2014) found that challenges to free will, including neuroscience that made no reference to determinism, *did* reduce

attributions of moral responsibility. Note, however, this finding recently failed to replicate across four studies (Monroe, Brady and Malle, 2017). In fact, even when neuroscience is related to the specific offender, the mock court approach has documented null effects (Schweitzer et al., 2011), conditional effects (Greene and Cahill, 2012) and even reverse effects (Saks et al., 2014). Similarly, the current findings add further doubt to the hypothesised threat of neuroscience (Greene and Cohen, 2004).

This study is unable to specify why attributions of blame and punishment were uninformed by the described cause of the disposition towards impulsivity. In contrast to certain mock court studies (Gurley and Marcus, 2008; Saks et al., 2014), the current study stated the accuracy of the tests of impulsivity (99%). The null effects could therefore reflect the fact that the current participants were implicitly instructed not to infer the neuroscientific diagnoses were more accurate than the cognitive diagnoses. In other words, neuroscientific testimony might derive influence from its seeming objectivity. However, this interpretation could not explain the effects of presenting neuroscientific testimony when the behavioural method of diagnosis remains constant (Schweitzer et al., 2011; Schweitzer and Saks, 2011) or null effects when the method is changed (Greene and Cahill, 2012; Schweitzer et al., 2011). The seductive allure of neuroscience does not appear to be a more general effect of presenting evidence from a 'harder' science (Fernandez-Duque et al., 2015).

Instead, participants may have feared that if neuroscience could excuse the impulsive actor, every impulsive actor would be excused on the same rationale. Indeed, people are less likely to consider causation relevant to judgements of responsibility when such judgements pose consequences for judgements in real life (Nahmias et al., 2007; Roskies and Nichols, 2008; Shepherd, 2012). In contrast to more severe and so rarer deficits in rationality (e.g., psychosis), mitigation on the basis of impulsive

character presents a slipperier slope. Hence people may find the far greater applicability of this neuroscience threatening and consequently, in need of resisting.

How do lay people represent the brain?

Most unexpectedly, participants were *more* likely to attribute impulsive offending to the conscious mind and a 'genuine desire' to offend after reading the neuroscientific testimony. This finding could be explained by prior evidence that people believe the brain is responsible for the 'core self'; where the 'core self' is conceived to include self-control and stable desires (Fernandez-Duque and Schwartz, 2016). The Deep Self model predicts that people attribute greater intent – or in this study, a greater role for the 'conscious mind' and 'genuine desire' – to behaviours that result from stable dispositions – or in this study, the brain (Locke and Frankfurt, 1975; Sripada, 2010). Similar to the current finding, Saks et al. (2014) found that the presentation of brain scans increased attributions of moral responsibility to schizophrenic offenders.

Although neuroscientific testimony did not increase attributions of fault in the current study, it did increase attributions of the 'conscious mind' and 'genuine desire'. This finding directly contradicts evidence that people tend to believe the brain bypasses the conscious mind (Nahmias et al., 2007). The discrepancy may reflect the target of explanation: whereas in previous studies, neuroscience was used to explain behaviour directly (Nahmias et al.), the current vignette used neuroscience to explain the preceding and clearly conscious process; that is exercising self-control. Therefore, while people may not perceive the brain as conscious by default (Nahmias et al.), this study supports the conclusion of Shepherd (2012): people are willing to adopt this perception, perhaps given their 'theory-lite' conception of the relationship between the brain and conscious mind (Nahmias et al., 2014, p.504).

In sum, neuroscience effectively shifted lay conceptions of impulsive character from a failure in the willingness, to a failure in the capacity, to exercise self-control. However, this shift posed no implications for attributions of choice, blame or punishment to the impulsive offender. Yet still, participants attributed greater fault to the normally controlled (than the normally impulsive) offender as a result of compatibilist and incompatibilist concerns; specifically, his perceived degree of dispositional self-control and the perceived reduction in the inevitability of his criminal behaviour.

viii. Conclusion

Greene and Cohen (2004) hypothesised that unlike the law, lay people blame offenders for their *ultimately free* actions; by suggesting that their actions are not ultimately free, neurocriminology could reduce support for positive retributivism. The first two empirical studies of this thesis addressed the scope for attitude change to result from learning about the neurobiological *correlates* of offending; first, the correlates of teenage impulsivity and second, the correlates of impulsive character. The first study did not relate neuroscience to the judged offender and, in this respect, the study represented the psychological approach to this field of research. Under this approach, the communication of neuroscience significantly reduced the attribution of blame for offending; though only to a small extent.

The second study did relate neuroscience to the judged offender and, in this respect, the study represented the mock court approach. Under this approach, the communication of neuroscience failed to reduce attributions of blame for offending. Although participants considered the neurobiologically impulsive actor less responsible for his impulsive disposition (than the cognitively impulsive actor), there was no

subsequent reduction in the blame attributed to the manifestation of that disposition in offending behaviour. In fact, the neurobiological actor was believed to have offended more consciously and with more 'genuine desire' to offend. Moreover, the increased blame attributed to the normally controlled actor could not be explained by the reduced attribution of his offending to the brain; in other words, this change in blame was not guided by the shift in causal attributions.

Together, the two studies suggest that a small degree of attitude change sometimes results from the communication of correlational neuroscience. As part of the psychological approach, the theatre study required participants themselves to link the neuroscientific explanation to the judged offending to a greater extent than the online study, yet still, the theatre study resulted in greater attitude change. While theatre may simply be a more effective means of communicating science, an alternative possibility exists; there may be some empirically significant difference between the target of neuroscientific explanation in the two studies. In particular, people may be more receptive to neuroscience as an explanation of teenage impulsivity than dispositional impulsivity; since the latter poses implications for more offenders. This is consistent with the fourth mechanism of neuroscientific influence hypothesised in the previous chapter: neurocriminology may have more mitigating influence when the brain appears to be an atypical cause of atypical behaviour (e.g., the atypical impulsivity exhibited in adolescence). The next chapter elaborates upon this fourth mechanism of neuroscientific influence.

Chapter 7: The impact of neurocriminological ideas on attributions of blame and the 'real self': A conceptual framework

Thus far this thesis has failed to provide empirical evidence that neuroscience presents a major challenge to public support for positive retributivism. The first two studies, however, did not present the brain as an explicitly causal explanation of criminal behaviour. It therefore remains possible that a major change in attitudes would result from communicating the philosophical assumption of neuroscience; that being 'all behaviour is caused by our brains' (Raine and Glenn, 2014, p.161). By their very nature, these claims link the neuroscientific explanation to the judged offending *for* the participant. In this respect, the museum study shifts the direction of the thesis from the psychological and mock court approaches to the approach of experimental philosophy.

In the philosophical studies, participants are asked to assume that a universally applicable, complete and causal explanation of all behaviour, criminal or otherwise, is true (Nichols, 2011). The causal explanation is typically a passage of text describing the philosophy of determinism. Participants are instructed to assume this passage accurately describes behaviour in a hypothetical world or the real world, before judging an actor in that world. The aim of the next chapter is to discuss the findings of a study conducted for this thesis that took such an approach – the museum study. In order to contextualise those findings, however, some important theoretical groundwork must be done; that work is done in this chapter and concerns how neurocriminology might influence perceptions of the 'real self' of the offender.

Chapter five described four possible mechanisms of neuroscientific influence: neuroscience may create the impression that the offender was unconscious at the time of offending (the first mechanism), that the offence took place in spite of moral intent (the second mechanism), that the offending behaviour was caused by unconscious

processes (the third mechanism) or that there was some abnormal cause of this offending behaviour (the fourth mechanism). The current chapter (and the museum study) are particularly concerned with the fourth mechanism; that is the perception of the brain as an abnormal cause of offending, and therefore, as an obstruction to the expression of the 'real self' of the offender. For this mechanism to become the focus of enquiry in the museum study, it was necessary to minimise the effects of the remaining three mechanisms. Specifically, it was important that the offender did not appear unconscious (the first mechanism), consciously helpless (the second mechanism) or to be behaving in a way that was caused by some unconscious factor (the third mechanism).

The first two mechanisms of neuroscientific influence derive from misinterpretations of the science; namely, that the neurobiological offender is unconscious or acts in a way that betrays moral intent. In the museum study, these misinterpretations were corrected by describing the judged offender as making a conscious decision to offend (and therefore, as acting in a way that is consistent with intent). Participants also judged this decision to offend (rather than the offending behaviour itself). This design feature was intended to exclude the third mechanism of neuroscientific influence; that is the potential for neuroscience to lead participants to believe the criminal intent was a passive correlate (rather than an active cause) of the offending behaviour. This distinction between criminal intent as cause and correlate is not of concern to the neurocriminological argument against positive retributivism (Greene and Cohen, 2004; Raine, 2013). Hence none of the vignettes in this research comment on the distinction.

The current chapter examines the argument that attributions of moral responsibility for a behaviour depend on how much that behaviour is believed to reflect

the ‘real self’ of the person exhibiting that behaviour (Sripada, 2010). Since very little research has been conducted to address this question directly, this chapter draws upon a body of research into the effects of describing offenders as having been ‘manipulated’ into offending in some way – such that their *bad* decision to offend misrepresents their *good* underlying ‘real self’. This chapter begins by reviewing studies that assess whether lay people blame offenders who have been manipulated into offending by a supernatural agent, some special medication or technology, or under duress.

On the basis of research into the manipulated offender, this chapter makes predictions about the effects of positioning the brain as the cause of criminal behaviour (as done by neurocriminology). The key question is raised of whether the offender’s brain activity, at the point of the crime, is considered evidence of a *bad* underlying ‘real self’ or the result of an unusual situation obstructing a *good* underlying ‘real self’. Subsequently, the chapter addresses some potentially confounding correlates of perceptions of the ‘real self’ that could provide an alternative explanation of the influence of the perceived ‘real self’. Finally, this chapter culminates in a description of the Deep Self Model – a theory that places perceptions of the ‘real self’ (also known as the Deep Self) at the heart of attributions of moral responsibility. In these various respects, the chapter examines how neurocriminology might influence perceptions of the ‘real self’, and consequently, attributions of blame.

i. Do people blame the offender whose brain has been manipulated?

People may continue to blame the neurobiological offender so long as it is clear the offender still made a conscious decision to offend. Outside the criminal context, this finding has been documented by two studies (Nahmias, Shepard and Reuter, 2014; Shepard and Reuter, 2012). The researchers asked participants to imagine a voter whose

electoral decision was predicted by a neuroimaging cap. The average participant continued to attribute free will to an actor whose conscious decision could be perfectly predicted by the neuroimaging cap. The finding supports the null or weak effect of communicating predictive neuroscience (Bode et al., 2011; Soon et al., 2008), as previously documented by Shariff et al. (2014; study 3). This weak or absent effect could be attributed to the continued salience of the conscious intent and the resultant compatibilist fixation on the capacity for rational choice – which remained intact in the described voters. One must therefore consider *when* neuroscience might influence perceptions of the rational offender whose offending resulted from uncontrollable prior events.

One factor that might increase or decrease the influence of neuroscience is how people believe the brain causes offending in relation to other causal factors. Nahmias et al. (2014) presented the brain as a proximal cause of electoral decisions, without any reference to earlier (distal) causes of brain activity. In contrast, Nahmias (2006) and Nahmias et al. (2007) did refer to the distal causes of brain activity and, perhaps therefore, found a strong effect of communicating neuroscience. People may care less about the brain causing the proximal (immediate) intent to offend and more about the causal history of that intent; for example, whether that person has always wanted to offend and why. Hence, perhaps neuroscience could change perceptions of whether the distal (historical) intent of the offender was morally good or not.

In this vein, Woolfolk, Doris and Darley (2006) presented participants with a character, Bill. Hijackers demanded that Bill kill his friend, Fred – and, in the second study, injected Bill with a drug that made him compliant to their orders. In this scenario, therefore, the hijackers have manipulated the proximal intent of the actor, such that Bill cannot choose to do anything but commit homicide – a compatibilist concern that did

reduce attributions of moral responsibility. Still, participants attributed more responsibility to the coerced actions of Bill when, before the hijacking, Bill had already decided to kill Fred (after discovering that Fred had an affair with his wife). Therefore, the prior intent of the offender influenced responsibility judgements, even when this distal intent was causally irrelevant to the outcome. More broadly, there is a large body of evidence that people are more likely to blame offenders who premeditate their crimes, even in hypothetical scenarios where the premeditation has no causal link to the crime (Morewedge, Gray and Wegner, 2010). According to Greene and Cohen (2004), people will one day attribute the same causal power to the brain as to any other coercive force, such as the hijackers in this study. Nevertheless, the findings of Woolfolk et al. suggest that the distal intent of the offender will to continue to influence blame.

In the neuroscientific context, Shepard and Reuter (2012) demonstrated the effect of changing the distal intent on lay judgements of the neurobiological voter. This time, participants imagined a voter whose electoral decision *could* be manipulated by a brain implant (Frankfurt, 1969). In one scenario, the implant predicted the actor would vote Republican; consequently, his decision was manipulated to ensure he voted Democrat. In the other scenario, the implant predicted the actor would vote Democrat; so his decision was not manipulated. Neither voter had any more incompatibilist free will – or ability to vote Republican – than the other voter and the participants correctly understood this. Yet still, 86% of participants attributed free will to the unmanipulated actor, compared to just 23% for the manipulated actor. Similarly, Nahmias et al. (2014, p.514) found that the average participant no longer attributed free will to an actor whose neurobiological decision was surreptitiously changed by scientists ‘several seconds before people are consciously aware of those decisions’. Therefore, attributions of free will are greatly influenced by whether the eventual decision (i.e., proximal intent)

corresponds to distal intent, rather than whether the decision is ultimately controllable (which it never was in the first study). In the criminal context, this suggests people would not blame the offender whose moral (distal) intent was neurobiologically manipulated into a (proximal) intent to offend.

Mele (2012, p.430) replicated these findings by asking participants to imagine a 'compliance drug' that renders people 'irresistibly caused to decide to take that course of action…[suggested] through a tiny computer chip that [scientists] implant in a person's brain'. The average participant no longer attributed free will to this actor yet did attribute free will to an actor whose decisions were 'a brain process that was caused by other brain processes' (p.429). Since brain manipulation – and not mere brain prediction – is necessary to undermine lay belief in free will, collectively these findings challenge Greene and Cohen (2004): it is only realistic to expect advancements in neuroscience to document better predictors of offending, not a supernatural manipulator of brain activity (De Brigard, Mandelbaum, and Ripley, 2009). Hence Nahmias et al. (2014) conclude that neuroscience is unlikely to threaten belief in free will.

It is critical, however, to assess whether the perceived cause of brain activity does, or ever could, share the morally significant properties of a supernatural manipulator. For Greene and Cohen (2004, p.1780):

> The fact that these [neurobiological] forces are connected to the desires and intentions of evil scientists [in their hypothetical scenario] is really irrelevant…What matters is only that these forces are beyond Mr Puppet's control.

For Greene and Cohen, the perceived cause of brain activity is uncontrollable and therefore comparable to scientists acting as manipulators. While, in fact, people do distinguish between neurobiological causation and manipulation (e.g., Nahmias et al.,

2014), perhaps there are other forms of neuroscience that more clearly share the properties of manipulation, such as uncontrollability.

Do people conceive of the brain as the 'real self'?

Since lay people do not attribute brain activity to a supernatural manipulator, the question remains of how people do perceive brain activity: as unconsciously caused (but not manipulated), as consciously chosen (but still caused) or as freely chosen. Pitts-Taylor (2010, p.639-40) supports the latter two options: '[In] the neoliberal construction of subjects in neuroscience…[mental] health maintenance becomes a responsibility or a duty rather than a right'. In other words, neurobiological actors are morally responsible for the state of their brain. As a concrete example, Maguire et al. (2000) attribute the larger posterior hippocampi of London taxi drivers to their choice of career. By having larger posterior hippocampi, London taxi drivers have a superior capacity for spatial memory. Hence their proximal 'decision' to recall a road on the job can be attributed to the volume of their hippocampi. In turn, the volume of their hippocampi can be attributed to their choice of career; that is their distal intent to regularly expose themselves to an environment – the roads of London – that developed their neurobiological capacity for spatial memory. In this way, taxi drivers could be held responsible for the volume of their hippocampi.

The same neoliberal reasoning could be applied to neurocriminology. The lay person might attribute to neurobiological offenders some distal intent to expose themselves to (criminogenic) environments that developed the brain activity that caused their decision to offend. Accordingly, even if the proximal intent to offend was caused by the brain, the offender remains morally responsible for developing the (criminogenic) brain state. If one rewound the causal chain further, the 'intentional'

cause of this brain state (e.g., the intent to socialise with antisocial peers) could be attributed to even earlier uncontrollable events; thereby undermining the responsibility of the offender for developing this brain state. However, lay people may not have the motivation or cognitive capacity to rewind the causal chain so far that they consider the implausibility of an incompatibilist free will. Hence people may care more about the morality (than the ultimate uncontrollability) of distal choices.

Regardless of their normal criteria for moral responsibility, people may *respond* to evidence of (neurobiological) causation by placing the scope for choice before the cause. This motivated account of lay reasoning has been tested by one study, generating unsupportive results: Monroe and Malle (2010, p.217) presented participants with the neuroscientific claim that 'free will is a false impression…all of our behaviour is caused by our neural impulses'. Of the 49% of participants who rejected this claim, only 14% explained their rejection by asserting that choice preceded 'neural impulses'. This finding casts doubt on the intuitiveness of the neoliberal response to neuroscience: few people spontaneously attribute distal intent to brain activity. Perhaps, therefore, lay people do not believe offenders freely (or even consciously) choose their (criminogenic) brain activity. Instead, brain activity might be attributed to unconscious processes.

Do people consider situational influences obstructions to the expression of the 'real self'?

If the brain is not conceived to reflect distal intent, perhaps the brain is actually believed to obstruct the expression of distal intent. In respect to the hypothetical voter, one would not expect the voter to change his voting intent upon entering the voting situation. In other words, features of the voting situation should not derail the distal

intent to vote for a particular candidate. Therefore, the voting decision cannot be explained by proximal features of the voting event.

The same is not true of crime: proximal features of the crime event can explain the decision to offend, especially if that decision is impulsive. The answers to where, when and how a person could offend inform whether a person does offend (Cohen and Felson, 1979). The perceived opportunity to offend often constitutes a key reason for offending (Wikström and Treiber, 2009). In contrast, the perceived opportunity to vote for a particular candidate does not vary across time and space; hence it cannot explain the direction of the voting decision.

Given the causal power of the situation in crime, there is far greater scope for the proximal intent of the impulsive offender – the intent to offend – to differ from his distal intent – a perhaps more moral intent. Hence, perhaps participants assign the brain a causal role in offending by separating proximal and distal intentions in a dualistic manner: participants might attribute distal (moral) intent to the mind and proximal (criminal) intent to the brain (Bloom, 2004; Forstmann and Burgmer, 2015).

In philosophical experiments that assess the impact of communicating neuroscience, participants are instructed to assume the *entire* cause of offending is the brain, before judging a hypothetical offender (e.g., Nahmias et al., 2007). In this scenario, participants may preserve scope for the influence of a 'separate mind' by ascribing the brain and mind causal influence at different points in time. Whereas the brain might be conceived to dominate the immediate situation of the crime event – the 'heat of the moment' –, the mind might be conceived to dominate the (calmer and more considered) normal cognitive life of the offender. When the brain – and the abnormal situation or emotion that it represents – dictates action, it may be deemed an abnormal

cause of the behaviour. In turn, the neurobiological explanation of offending may gain mitigating influence.

The lay person might believe that neurobiological causes of abnormal behaviour misrepresent the normal cause of normal behaviour; that is the considered mind. Consequently, people may perceive criminal examples of proximal intent to misrepresent normal moral intentions and so, proximal brain activity to misrepresent the normal mind. This temporal means of sustaining belief in dualism is unavailable in the voting scenario, where the situation is causally irrelevant to the voting decision; there is no 'heat of the moment' that could justify the subsiding of the mind. Hence one would especially expect to observe this mechanism of neuroscientific influence in judgements of opportunistic crime. In these vignettes, people may perceive the opportunistic situation as inducing 'immoral brain activity' that decouples the actor from a normal 'moral mind'. Consequently, neuroscience may gain some of the mitigating influence of the supernatural manipulator.

Do people consider a manipulated brain an obstruction to the expression of the 'real self'?

From this perspective, the principal concern is not whether the distal intent to be good is free, but that the actor did have this intent and that it has been obstructed by an abnormal situation and resultant abnormal brain activity. Hence descriptions of manipulation may undermine a different criterion for responsibility than descriptions of determinism; where the latter criterion is the attribution of incompatibilist free will (Mele, 2012; Shepard and Reuter, 2012; Nahmias et al., 2014). Whereas people may believe the earlier (distal) intent of the manipulated offender was moral and consequently, excuse the offender, the incompatibilist could conceive the intent of the

predetermined offender to have always been immoral, yet unfree and so still exculpatory. In this case, the actions of the predetermined and manipulated offenders are excused for very different reasons.

In the study by Nahmias et al. (2014), it is unclear whether participants believed the brain implant manipulated the voting decision of the voter by, or without, changing her (proximal) voting intent. Participants either believed that the brain implant manipulated her behaviour directly, bypassing intent, or that the implant manipulated her (proximal) voting intent, thereby preventing the voter from acting upon prior (distal) intent (to vote for a different candidate). Supporting this latter interpretation, the vignette actually stated that the brain implant 'can change what [the voter] thinks' (Nahmias et al., p.514). However, when asked directly, participants were less likely to attribute the voting decision of the manipulated voter to her conscious mental states; moreover, this change in attribution partially explained the effect of describing manipulation. While this finding supports the bypassing interpretation of neuroscientific influence, it is unclear which conscious mental states the participants were judging: proximal states – the concern of the traditional bypassing hypothesis – or distal states – the concern of this chapter.

In fact, lay attributions of intentionality may comprise both the intent to offend formed immediately prior to offending (the proximal intent) and the lifetime accumulation of intent to act morally or not (distal intent). In turn, even people who have no concern with ultimate controllability (i.e., compatibilists) may still assess whether the (manipulated) intent of the offender immediately prior to offending represented his prior (unmanipulated) intent. In manipulation scenarios, the manipulated offender starts on a moral footing, yet develops the desire to offend

following manipulation (Mele, 2012; Shepard and Reuter, 2012; Nahmias et al., 2014). The compatibilist may excuse this offender on the basis that he once had moral intent.

In support of this theory, Sripada (2012b) presented participants with a hypothetical offender whose homicidal intent developed gradually across his life – either 'completely by chance' or following the deliberate attempt of an evil scientist to 'teach and reward' and 'nurture and cultivate' the intent to kill (p.12). Whereas a mere 36% of participants attributed free will to the nurtured offender, 91% attributed free will to the chance offender. Not only does this discrepancy evidence the large effect of manipulation, yet Sripada (2012b, p.21) concludes that lay people are compatibilist: 'Our intuitions…do not track the features that incompatibilists say they track (i.e., the agent's lack of ultimate control over his or her actions)'. Sripada's (2012b) conclusion assumes that participants perceived each type of homicide to be equally uncontrollable (Gorin, 2013), yet this assumption can be challenged. In the manipulated condition, criminal intent resulted from the upbringing provided by the evil scientist. In the unmanipulated condition, the source of intent was ambiguous, arising 'completely by chance', with no reference to upbringing (p.12). Therefore, it is unclear whether participants also believed the upbringing of the unmanipulated killer caused his criminal intent. While the type of upbringing one receives is a matter of chance, participants could also operationalise chance in many other ways (e.g., genetic differences) or simply fail to recognise the uncontrollability that 'chance' implies. This critical perception of chance as uncontrollable cannot be confirmed, since Sripada failed to measure 'the features that incompatibilists say [intuitions] track' (p.21). Hence it remains possible that participants perceived the manipulated offender as having less ultimate control (*and therefore* less free will) than the unmanipulated offender. If so,

this would constitute an incompatibilist explanation of the influence exerted by the manipulation description.

However, Nahmias et al. (2005) provide indirect evidence to challenge the incompatibilist explanation and thereby support Sripada (2012b). Nahmias et al. (p.570) asked participants to imagine a world in which 'one's genes and environment completely cause one's beliefs and values' and that 'after deliberation, [a thief in this world], because of his beliefs and values, keeps [some returnable] money'. While this account appears to represent a combined genetic and environmental (i.e., a biosocial) explanation of offending, the vignette also stated:

> If [the thief] had been adopted by the [same family that adopted his non-offending identical twin], he would have had the beliefs and values that would have caused him to [not offend like his identical twin] (p.570).

Hence upbringing was cited as the cause of offending.

The social explanation presented by Nahmias et al. (2005, p.570) differs in only one respect to the social explanation presented by Sripada (2012b): in the former study, the offender is unintentionally created by a 'selfish…family' rather than intentionally created by an evil scientist. Although Greene and Cohen (2004, p.1780) claim the 'intentions of evil scientists is really irrelevant' to responsibility judgements, both philosophers (Waller, 2014) and the empirical data suggest otherwise: whereas the majority of participants attributed free will to the *unintentionally* nurtured thief (Nahmias et al., 2005 - 76%) and the *unmanipulated* killer (Sripada, 2012b - 91%), only 36% of participants attributed free will to the *manipulated* killer (Sripada, 2012b). In fact, this effect of the offender having prior moral intent appears strong, since one would ordinarily expect blame to be greater for the far more emotive act of homicide than theft (Nichols and Knobe, 2007). Therefore, in the manipulation vignettes, attributions of free will appear to track the presence (or absence) of prior moral intent

rather than the ultimate controllability (or uncontrollability) of the immediate intent to offend.

Do people consider an abnormal brain an obstruction to the expression of the 'real self'?

In Sripada's (2012) study, the manipulated killer was not given a normal opportunity to develop moral intent. Although genes and social environments are uncontrollable products of chance, normal people at least gain entry to this biosocial lottery. While the outcomes of the lottery are not deserved (the lottery is a random process), normal people still receive the (randomly determined) opportunity to access genetic and environmental advantages. In this important sense, the manipulated offender is not normal. The evil scientist 'implemented his plan for Bill. He took Bill from an orphanage when Bill was an infant' (Sripada, 2012, p.574). Therefore, the 'quality' of parenting received by Bill was not random. Bill did not have access to the normal benefits of chance; that is the possibility of receiving 'good' parents.

It is normal to credit the Samaritan and blame the offender – life roles that are ultimately products of chance circumstance. When the actor is manipulated, however, normal chance processes are disrupted by the evil scientist (Sripada, 2012b) or the brain implant (Mele, 2012; Nahmias, Shepard and Reuter, 2014; Shepard and Reuter, 2012). As a result, the actor uses his capacity for rational choice to pursue criminal ends because of abnormal developmental processes and according to some theorists, this should mitigate the administered punishment (Haji and Cuypers, 2004). In contrast, the unmanipulated actor is responsible for the behavioural outcomes of chance – the normal origin of the causal chain.

From this perspective, the structure and function of the brain is the product of the normal biosocial lottery and therefore, a blameworthy product of random chance. If lay people recognised this fact, neurobiological explanations of offending would not threaten attributions of blame. While no study has tested this prediction in respect to moral responsibility, Fernandez-Duque and Schwartz (2016, p.4) measured judgements of moral character: the 'real self' (i.e., defining properties of the self) was considered a better indicator of moral character, more attributable to the brain, more enduring and less under 'wilful control' than the 'peripheral self'. This finding created 'the paradox of seeing the central self as most morally relevant and yet least changeable through wilful control' (Fernandez-Duque and Schwartz, 2016, p.11). In other words, moral judgements did not correlate with attributions of control. The paradox is, however, consistent with the current argument: people continue to blame actors for their uncontrollable traits so long as those traits reflect the real self – the self resulting from the normal processes of chance implicated in biosocial development. People may care more about whether this real self, also termed 'chance self' or 'distal self', is moral than controllable.

To predict the influence of neurocriminology, it therefore seems important to understand whether lay people attribute the real self to the brain; the findings of Fernandez-Duque and Schwartz (2016, p.4) suggest people do in a morally neutral context. In fact, if the brain is conceived to house the real self, neurobiological explanations of offending could even exacerbate attributions of blame by generating the impression that the real self is immoral. Yet this prediction assumes the brain is always perceived to be a product of the normal biosocial lottery and therefore, the real self. In contexts where the brain is not conceived as such, neurocriminology may have the

greatest mitigating influence. In these contexts, the brain may be deemed an obstruction to the normal origin of behaviour: chance.

Sripada's (2012) mediation analysis can be reinterpreted from this perspective of normal causation. The reduced attribution of free will to the manipulated (relative to chance) killer was entirely explained by ratings of two compatibilist concerns: the perceived 'corrupting of the information used by' the offender to form intent and the 'discordance' of the offending behaviour from the 'deep self' of the offender (Sripada, p.7). I argue that people define these compatibilist concerns in respect to chance. The 'corrupting of information' might represent the failure for the content of that information to be determined by random chance. Similarly, the 'discordance' of behaviour from the 'deep self' might represent the divergence of the proximal intent to offend from the distal intent that the offender (would have) exhibited by random chance. Hence, for the compatibilist lay person to consider a decision to offend 'free', it may be important not only that the offender acted in line with conscious intent and desire but also that this conscious intent and desire arose through normal developmental processes (i.e., by random chance).

There, however, remains a problem with this compatibilist interpretation of Sripada's (2012) findings. Sripada (2012, p.18) comments:

> Incompatibilists…believe that intuitions about free will are not influenced by deep self discordance and corrupted information, but rather are influenced by another factor (i.e., the absence of ultimate control).

However, this claim fails to account for the fact that participants may have inferred the 'absence of ultimate control' from 'deep self discordance' and 'corrupted information' (Björnsson, 2016). Participants may have perceived the unmanipulated killer to be less *controlled* by corrupted information and to display greater concordance with his deep self *and therefore his free self*. There are two pieces of evidence to suggest participants

extrapolated from the items in this manner: first, in a second study by Sripada (2012b, p.19), participants in both conditions were reassured that the offender was acting on his deep self (his 'desires include killing' and 'his core values permit killing') and not on corrupt information (he knows 'the difference between right and wrong [and…] exactly why he dislikes [the victim]'). Nevertheless, the manipulated actor was still conceived to be acting in a way that was inconsistent with his deep self to a similar extent as without the reassurances in the first study. Perhaps therefore, participants actually perceived the manipulated killer to be acting in a way that was inconsistent with his *free* self.

Second, Björnsson (2016) replicated the study without reference to control in the mediating items and by replacing the measure of free will with support for positive retributivism. Consequently, the effect of describing manipulation was largely unmediated by corrupted information and deep self discordance but was instead mediated by attributions of ultimate control (i.e., incompatibilist free will). In other words, describing manipulation had a mitigating effect by suggesting there is no free will at the beginning of the chain reaction that leads to a given behaviour. In fact, the effect of describing manipulation on corrupted information and deep self discordance was mediated by support for positive retributivism. Hence Björnsson concludes that people are concerned with deep self discordance only because it implies divergence from the free self, thereby providing some support for Greene and Cohen (2004, p.1780): 'rationality is just a presumed correlate of what most people really care about'; that being incompatibilist free will. However, although people may care about incompatibilist free will too, it is certainly not their chief concern; for example, on a scale of 1 to 7, mean support for the retributive punishment of the manipulated (and so unfree) killer was still above the mid-point, at 5.1 (Björnsson, 2016). Therefore, the

traditional compatibilist justification for blame – the attribution of proximal intent and desire – remains strong.

In sum, there is evidence that the effect of describing manipulation is one of casting doubt on the existence of incompatibilist free will (Björnsson, 2016). Yet still, this incompatibilist account assumes that lay people interpret items about 'ultimate control' and 'deserving punishment' to be unambiguous references to incompatibilist free will. I suggest that participants do not think so far back in the causal chain; instead, such items may be interpreted as references to the normal development of the causal chain along a randomly determined, rather than manipulated, path. Accordingly, the manipulation of brain activity undermines blame by suggesting the induced behaviour represents a deviation from the normal developmental trajectory of this person (i.e., the real self).

There are at least four reasons to favour the explanatory importance of the real self over the free self. First, there is a body of philosophical experiments to suggest blame withstands arguments against the existence of an incompatibilist free will (Nichols, 2011). Second, neuroscientific testimony can undermine blame without first reducing attributions of control; therefore, control is not a prerequisite to blame (Gromet et al., in preparation). Third, free will believers are more likely to punish an offender due to their greater propensity to attribute the offending behaviour to dispositional (rather than situational) factors; a tendency known as the correspondence bias (Genschow, Rigoni and Brass, 2017). In other words, people believe the offender to have a less moral self as a result of perceiving the offender to be freer, most likely interpreted as more intentional (e.g., Nahmias et al., 2005), thereby driving punitiveness.

Fourth, only the idea that people care about the real self can explain the asymmetrical effects of meta-desires. Meta-desires refer to whether the offender endorses or rejects his own desire to offend. People attribute less blame for immoral behaviour that results from 'uncontrollable' impulse as opposed to 'controllable' deliberation (Pizarro, Uhlmann and Salovey, 2003). However, people attribute no less praise for impulsive acts that are morally good. This selective effect of impulsivity for blame (and not praise) was explained by the perceived meta-desire of the impulsive actor: participants assumed that the impulsive *moral* actor wanted to experience the morally good impulse, yet that the impulsive *immoral* actor did not want to experience the morally bad impulse. Therefore, the impulsive offender was blamed less for appearing to have a disapproving (and thereby moral) meta-desire. If instead, the impulsive offender was blamed less for appearing to have less incompatibilist free will, one would also expect the impulsive *moral* actor to be praised less – which was not the case. This suggests that blame tracks the attributed meta-desire rather than the attributed degree of ultimate control.

This interpretation is strengthened by an additional finding of Pizarro et al. (2003): when the immoral actor was explicitly described as not wanting the immoral impulse, blame was reduced. Similarly, when the moral actor was described as not wanting the moral impulse, praise was reduced. Hence the typical effects of the assumed meta-desire can be strengthened by explicit assertions. Newman, Bloom and Knobe (2014, Study 4) replicated the selective effect of impulsivity on blame, finding the effect to be explained by perceptions of the 'true self' rather than the meta-desire. Perhaps therefore, meta-desires derive their moral relevance from their relation to the real self. Explicit descriptions of the 'real self' as bad (rather than good) do amplify blame (Newman et al., Study 5).

In fact, immoral desire is sufficient to invoke blame for actions that neither cause the immoral outcome nor intend to cause the immoral outcome. Specifically, Inbar, Pizarro and Cushman (2012) found that participants blamed a man for adopting a position that would cause him to benefit from harm occurring (e.g., betting on an earthquake occurring). It was clearly understood that betting on an earthquake could not increase the probability of the earthquake occurring. Nevertheless, the attribution of blame persisted even when the harm and resultant payoff for the man never did occur and was explained by the perception that the man *wanted* the harm to occur. People may therefore interpret moral desires as a strong indicator of the real self. Since the effect was observed independent of character judgements, the real self might be best represented as extending beyond character. Hence neurocriminology may gain mitigating influence when the brain is conceived as disrupting the normal expression of the real self, especially real desires; specifically, I have argued for the importance of the real self rather than the free self.

Do people consider the situational brain an obstruction to the expression of the 'real self'?

Thus far I have demonstrated that people attribute less moral responsibility to the actor whose behaviour results from a manipulated brain rather than an unmanipulated brain (Mele, 2012; Shepard and Reuter, 2012; Nahmias et al., 2014) or unspecified chance (Sripada, 2012b). When the brain is manipulated, it cannot be attributed to the normal biosocial lottery. When the brain is not manipulated, however, it can be attributed to the normal biosocial lottery and therefore, considered part of the blameworthy real self (Fernandez-Duque and Schwartz, 2016). Yet the criminal context complicates this perception: when the brain is described as the cause of immoral behaviour, this challenges the perception of the brain as the real self; since, by default, the real self is

believed to be morally good (Newman, Bloom and Knobe, 2014). Consequently, a certain cognitive dissonance is generated: in the morally neutral context, the brain is the real self, yet as a cause of crime, the brain cannot be the blameworthy real self because the real self is morally good.

After learning about neurocriminology, this dissonance must be resolved: the participant must perceive the real self of the offender as either endorsing or condemning, as either representing or fighting, the neurobiological desire to offend. This distinction resembles the difference between perceiving addicts, deep down, as either approving or disapproving their desire to engage in the addictive behaviour; termed the willing and unwilling addict (Frankfurt, 1971). People can either represent the unmanipulated brain as home to the immoral real self or as obstructing the expression of the moral real self, thereby generating an aggravating or mitigating response to neurocriminology respectively. The choice of representation and subsequent response may reflect the contextualised desire to punish this particular offender.

Given the default perception of the real self as morally good (Newman et al., 2014), people may be inclined to interpret criminogenic brain activity as obtrusive when there is a situational explanation of that brain activity. In the criminal context, therefore, brain activity may be perceived similarly to impulses; as mitigating obstructions to the normal capacity of the offender to act upon the reasoning real self. Indeed, people are less likely to blame the offender acting upon 'uncontrollable' impulse rather than 'controllable' deliberation (Pizarro, Uhlmann and Salovey, 2003). Importantly, this effect only arises in cases where one would expect the situation to disable the expression of normal moral character and therefore the real self; lay perceptions of the real self primarily comprise perceptions of moral character (Strohminger, Knobe and Newman, 2017). An example situation would be discovering

a partner having sex with another person. When offending appears to have a situational cause like this, evidence of the neurobiological impact of that situational cause might mitigate blame.

People do not always separate situational causes from the real self. When offending is impulsive and the impulse is attributed to the real self, people actually blame the impulsive offender more than the deliberating offender. This reversal of the usual mitigating effect is observed in situations that one would not typically consider capable of disabling moral character; for example, the situation of discovering a lost wallet (Critcher, Inbar and Pizarro, 2013). In these situations, participants believe the offender must have had a particularly immoral character to have experienced this impulse; since the situation alone cannot account for the impulse. When there is no situational reason to expect the real self to be disabled, yet the immoral impulse persists, the real self must be immoral; the neurobiological impact of the impulse only provides further evidence of the physical reality of this self. In turn, neurocriminology might not only fail to mitigate but in fact, aggravate blame.

I argue that people evaluate the brain similar to impulses, mitigating or aggravating blame depending on whether the brain is conceived to obstruct or represent the real self. Thus far, only two unpublished studies provide findings of direct relevance to this question. First, Gromet et al. (in preparation) found that participants attributed less moral responsibility, blame and punishment to the impulsive murderer whose 'dysfunction' in self-control was described as 'neurological' rather than 'psychological'. This effect was observed independently of whether the dysfunction was also attributed to genes or childhood abuse. Importantly, the effect of neurobiological description was explained by the perception that the neurobiological offender was 'less reflective of whom he is as a person and of his true character'

(Nadelhoffer et al., 2013a, p.202), yet not by attributions of control, credibility or the strength of the causal link between the dysfunction and the offending behaviour. This suggests that relative to cognitive dysfunction, people are more likely to perceive evidence of brain dysfunction as obstructing the expression of the blameworthy real self.

In this sense, Gromet et al. (in preparation) replicated the findings of Woolfolk et al. (2006) in the context of neuroscience: blame differs with the real self, even when the offending behaviour remains equally uncontrollable, in the former study, due to hijackers (Woolfolk) and in the second study, due to a dysfunction in self-control (Gromet et al.). Judgements of the real self, therefore, represent a compatibilist mechanism of neuroscientific influence. The neurobiological offender is perceived similar to the unwilling addict; as uncontrollably wishing to offend, yet also wishing he did not have that criminal desire (Frankfurt, 1971). Consequently, blame is reduced.

According to this theory, neurocriminology should lose its mitigating influence when the offender clearly embraces his criminal desire. Indeed, a second study by Gromet et al. (in preparation) found no effect of neurobiological description when participants learnt that the offender already had a strong motive to offend (against this victim) before encountering the situation that provoked the impulse to offend. Hence the neurobiological offender may not be excused when the neurobiological impulse is consistent with the distal intent, and so the real self, of the offender. In other words, people do not believe that criminogenic brain activity is obstructing a moral real self when there is direct evidence that the real self is not moral.

There are some contexts where the brain is clearly represented as the (immoral) real self rather than an obstruction; for example, when neuroscience is used to predict

reoffending. In a study by Nadelhoffer et al. (in preparation), participants read that the offender was highly likely to reoffend, as inferred from 'brain-based information' by a neuroscientist or from psychological or statistical information by a psychologist or statistician respectively. The participants were *more* likely to attribute blame, moral responsibility and punishment to the neurobiological offender *and* more likely to attribute his behaviour to 'his true underlying values and character' (Nadelhoffer et al., 2013a, p.206). Hence neurocriminology might actually increase attributions of offending to the blameworthy real self when the science is used to explain a stable disposition to offend. This finding raises the interesting possibility that the aggravating effect of neuroscience could actually arise from increased blame – a retributive (rather than consequentialist) concern. In this context, the neurobiological offender may be perceived similar to the willing addict; the offender who uncontrollably wishes to offend and most importantly, wishes to have that criminal desire (Frankfurt, 1971).

ii. Some potentially confounding correlates of perceptions of the 'real self'

Thus far I have argued that neurocriminology may reduce blame when the criminogenic brain activity is conceived as a situationally induced obstruction to an earlier moral self. However, one might question whether the two studies supporting this theory confidently measured perceptions of the real self independently of other correlated constructs. In particular, neither study specified the intentions and desires of the offender at the time of the crime. While Nadelhoffer et al. (2013a, p.207) claims 'there is no obvious reason for subjects to have supposed that [the offender's]…means-end beliefs, surface desires, and intentions, differed across the three conditions', neuroscientific descriptions can radically influence perceptions of these constructs (Nahmias et al., 2007). Moreover, if such perceptions were changed, they could subsequently influence character assessments (Pizarro and Tannenbaum, 2012). Hence

it remains possible that perceptions of the real self are changed only because perceptions of proximal intent and desire are changed – and that changes in the latter constructs explain variation in blame.

No study has separated these two potential mechanisms of neuroscientific influence. However, in their vignettes of moral behaviour, Gill and Cerce (2017) presented child abuse as the reason for an office bully developing the proximal intent and desire to bully his colleagues. The participants correctly understood that the abuse in no way changed the morality of his character or impeded his capacity to exercise rational choice (measured as intentionality and uncoerced choice). Nevertheless, the description of child abuse still mitigated attributions of blame and retributive punishment. Therefore, child abuse can mitigate without affecting the perceived state of the conscious mind at the time of the crime. Moreover, this effect of child abuse was mediated by a diminished perception that the bully, as opposed to some external influence (e.g., abuse), controlled the formation of the bullying disposition. Therefore, abuse may be conceived as tainting the prior moral self, such that the bully's 'character was "implanted" in him by his life experiences' (Gill and Cerce, 2017, p.3).

Importantly, this finding was not simply an effect of sympathy for the prior suffering of the offender; the effect was replicated when the child was spoilt rather than abused. Although the effect of childhood abuse was small, it does contrast with the null effects reported by prior vignette studies (Alicke, 1990; Najdowski and Bottoms, 2009; Pfeffer, Maxwell and Briggs, 2012; Weiner, Perry and Magnusson, 1988); real jurors are more likely to deem child abuse irrelevant (than mitigating) and controllable (than uncontrollable) in capital sentencing simulations (Stevenson, Bottoms and Diamond, 2010), while, in real cases, the perceived effects of child abuse on the dangerousness of the offender override its mitigating potential (Stevenson, 2009). Gill and Cerce (2017)

attribute the difference in their findings to their description of child abuse, in which a *clear narrative* explained the origins of their bullying intent. The Explanation Hypothesis might claim the narrative is necessary to render the abuse a sufficiently salient explanation of the behaviour (Björnsson and Persson, 2013).

The museum study considers whether this form of mitigation, where child abuse explains the proximal intent to offend, could be replicated with neuroscience; by presenting a neurobiological explanation of the proximal intent. Gill and Cerce (2017) compared the effects of describing child abuse and 'brain damage caused by a traumatic birth' (p.13). Describing brain damage diminished first, attributions of rational choice, thereby reducing blame, and second, control of character formation, thereby reducing support for retributive punishment. The latter finding could support the current argument: the participants may have engaged in counterfactual reasoning, imagining that the offender would have developed an alternative (moral) character *if only* the birth had not been traumatic and the brain not damaged. If so, brain damage would be perceived as a barrier to the expression of prior moral intent, thereby justifying sentence mitigation.

However, there also exists an alternative interpretation of the finding: unlike child abuse in the first three studies of their article, Gill and Cerce did not present brain damage as the reason why the *intact* capacity for rational choice was used to achieve immoral ends. In fact, the vignette contained 'no explicit description of [the bully's] actions as choices' (p.13). Hence it remains possible that participants believed the bully had a diminished capacity for rational choice and that this diminished capacity was caused by brain damage which lay beyond the bully's control. Indeed, recognised impairments to rational choice (e.g., seizures and delusions) lose mitigating influence when the offender appears to have controlled the formation of those impairments; for

example, through drug use (Alicke and Davis, 1990). Gill and Cerce explicitly described the bully as having 'brain damage caused by a traumatic birth' (p.13). Hence the participants may have identified the bully as having diminished rationality *and* as lacking control over the diminishment of his rationality. In turn, there remains a need to test the mitigating influence of presenting the brain as a cause of *how rational choice is pursued*, rather than a cause of the diminished capacity for rational choice; especially since this form of neurocriminology would apply to far more offenders (than simply those with clear deficits in the capacity for rationality). The museum study fulfilled that need by describing the brain as causing the *rational* choice to offend. The study therefore tested the scope for neuroscience to derive mitigating influence through the suggestion that even a rational brain could obstruct the real self in an unusual situation.

iii. The Deep Self Model

I first introduced the concept of the real self by referring to an example constituent of the real self: the dominant moral orientation of the intentions one has following normal developmental processes. However, the real self includes not only intentions but also desires and meta-desires. Hence I will hereon refer to the real self as the Deep Self, not only as a reference to the range of conscious mental states one can have and but also to the Deep Self Model (Sripada, 2010). In the Deep Self Model, Sripada (p.7) predicts that people blame offenders for their immoral actions when their criminally Acting Self – 'the narrow set of outcome-directed proximal desires, means-end beliefs, and intentions, that are the immediate causal source of the action' – concords with the stable traits of their Deep Self – 'a much larger set of more stable, enduring and fundamental attitudes'. Therefore, the Acting Self and the Deep Self are intended to be more encompassing terms than proximal intent and distal intent.

Do people consider genetic and environmental influences to constitute the Deep Self?

When people attribute criminal intent to events that occurred long before the crime, it is no longer plausible that this (distal) cause of offending could have obstructed the expression of the Deep Self – because the distal cause is the Deep Self (or at least entangled with it). The more distal the cause of offending, the less the plausibility of separating the Deep Self from the cause. Hence the lay person may blame the distally caused offender more – that is the offender whose criminal behaviour originates from factors in the distant past – because the Acting Self then represents the distal cause *and therefore* the Deep Self.

While these predictions are based on a descriptive account of how lay people attribute responsibility (Sripada, 2010), the same reasoning can be found in an earlier normative account of when the manipulated actor should and should not be held responsible for his actions. Locke and Frankfurt (1975, p.120) describe a 'Devil/neurologist', who manipulates the biology – and consequently, every behaviour – of his victim either continuously or only once at the outset of life. So the behaviour of the victim is always manipulated, either proximally (before every instance of behaviour) or distally (just once before a lifetime of manipulated behaviour).

Locke and Frankfurt (1975, p.120) argue that the *distally* manipulated actor should be held responsible for his uncontrollable actions:

> The other possibility is that the Devil/neurologist provides his subject with a stable character or program, which he does not thereafter alter…and that the subsequent mental and physical responses of the subject…are determined by this program rather than by further intervention on the part of the Devil/neurologist. In that case, there is no reason…against regarding him as capable of being morally responsible for what he does.

Since biosocial influences represent this more 'stable character or program', people are likely to blame the biosocial offender for his actions.

Do people consider situational brain activity to constitute the Acting Self?

In contrast to the distally manipulated actor, Locke and Frankfurt (1975, p.120) argue that the *proximately* manipulated actor should not be held responsible for his uncontrollable actions:

> Each of the subject's mental and physical states is the outcome of specific intervention on the part of the Devil/neurologist…the subject's instantaneous states of mind may be as rich as those of a person…but the person has no character or disposition of his own, and there is no reason to expect from him – …insofar as there is no reason to expect it from the Devil/neurologist – even the minimum of continuity…Instances of his behaviour…cannot legitimately be described as him and his actions.

I suggest that when neurocriminology is used to explain situational offending, lay people may perceive neurobiological causes of behaviour in this manner; as 'instantaneous states of mind' that are 'the outcome of specific intervention on the part of the [ever changing situation]' and therefore do not display 'character or disposition…[or] continuity'. Brain activity may therefore be deemed as fragile as the thoughts and situation it represents. In fact, the presentation of brain scans in court has been criticised on the very basis that the scans can only ever document brain states at the time of scanning: 'There is no known way to retroactively observe the state of the brain as it was during the commission of the offence' (Aharoni et al., 2008, p.151). Therefore, in some offending vignettes, the situation may be perceived as tugging the neurobiological Acting Self away from a different and moral Deep Self, thereby justifying sentence mitigation.

Brain states do fluctuate between situations and in this respect, it is normatively justified to consider the brain a proximal cause of offending. However, the brain also

'serves as a bottleneck for all the forces…of your past that affect who you are' (Greene and Cohen, 2004, p.1781). Over time, these 'forces' affect the shape of the bottle. Therefore, your brain is not only your Acting Self but also an imprint of 'who you are' – your past; your Deep Self. Consequently, neurobiological explanations are inherently ambiguous: one can make either a dispositional interpretation of brain activity as the Deep Self or a situational interpretation of brain activity as the Acting Self.

Participants may be primed to make the situational interpretation when the vignette contains other cues that a different and moral Deep Self exists. For example, in a previously discussed study, Mele (2012, p.430) describes the hypothetical offender as 'a very honest man' – an explicit assertion of his moral Deep Self – in the manipulated explanation of offending, yet not in the (merely) causal explanation. In turn, the causal explanation may have lost mitigating influence given the absence of any cue that the offender had an alternative, moral Deep Self. Hence, to conduct a more valid test of neuroscientific influence, in the museum study, participants judged a hypothetical offender who was always portrayed as having an alternative, moral Deep Self, regardless of whether neuroscience was used to explain his offending or not. It was predicted that this portrayal might promote a situational interpretation of the brain, thereby increasing the mitigating scope of neuroscience.

In the studies of brain manipulation, a vital and ecologically valid control condition is also missing; that is a condition in which the criminogenic brain activity is neither unexplained nor induced by an evil scientist but attributed to an unusual situation (Shepard and Reuter, 2012; Nahmias et al., 2014). In the museum study, participants judged a hypothetical offender whose offending behaviour was always described as the response to an unusual opportunity to offend; in turn, this gave participants the opportunity to link the neuroscientific explanation to an abnormal

situation. The Deep Self Model would predict that a proximal explanation of criminal intent could undermine blame by suggesting the moral Deep Self was derailed by a particular situation or tumour (rather than an evil scientist); for example, in cases of acquired paedophilia, paedophilic tendencies arise and cease with the growth and removal of a tumour (Burns and Swerdlow, 2003). When such brain abnormalities are attributed to a manipulator in hypothetical scenarios, blame might be especially reduced (Waller, 2014). Nevertheless, I hypothesise that blame can still be reduced (to a smaller extent) even without describing a manipulator to function as a substitute target for that blame.

Dualistic beliefs may also facilitate the situational interpretation of brain activity. Since people intuitively separate the brain from the conscious mind (Bloom, 2004; Forstmann and Burgmer, 2015), people may also separate neurobiological causes of offending from the Deep Self (but see Fernandez-Duque and Schwartz, 2016). The same reasoning, however, cannot apply to distal causes of behaviour; for example, genes and the social environment (i.e., biosocial influences). Unlike the brain, biosocial influences exert stable impressions on the self, with their effects unfolding gradually and relatively consistently across the lifespan. Since the Deep Self is, by definition, the collection of these stable impressions, it would be difficult to conceive of biosocial influences as obstructions to the Deep Self. Hence one would expect people to blame the biosocial offender; defined as the offender whose criminal actions can be entirely attributed to genetic and social influences.

In sum, I hypothesise that people rightly perceive brain activity at the point of the crime to be a proximal cause of offending and therefore, a potential obstruction to a different and moral Deep Self. Consequently, the neurobiological offender appears more likely to have a moral Deep Self than the biosocial offender or the unexplained

offender; recall that the unexplained offender is the offender whose actions are unexplained to participants and the biosocial offender is the offender whose actions are explained by genetic and environmental influences. In turn, the neurobiological offender attracts less blame. One unpublished study supports this prediction: participants attributed a less moral disposition to the offender whose actions were explained by child abuse rather than brain chemicals (Beall et al., 2013). This finding is consistent with the Deep Self Model: for the sufferer of child abuse, an enduring part of the self appears to have developed an immoral orientation, not just the self acting at the time of the crime. For the neurobiological offender, in contrast, the morality of the Deep Self may remain intact.

Note the difference in character judgements disappeared when scans of brain damage were presented in both conditions (Beall et al., 2013). Perhaps the brain scan led participants to perceive the neurochemical properties of the neurobiological offender as enduring, thereby removing the scope to believe he still had a moral Deep Self. Hence this study provides preliminary support for the Deep Self Model as an explanation of neuroscientific influence. The museum study extended this line of research by directly measuring perceptions of the 'real self' and their punitive effects, and by presenting a complete biosocial explanation of offending; rather than simply referring to child abuse.

iv. Conclusion

People may blame offenders for their offending behaviour in proportion to how much that behaviour is perceived to reflect the 'real self' of the offender (Sripada, 2010). This chapter presented evidence to support the influence of the perceived 'real self' from a body of research into judgements of the 'manipulated' offender; in other words, an

offender whose *bad* decision to offend is a clear deviation away from a *good* underlying 'real self', induced by some external agent, technology or substance. This research was used to generate a hypothesis about how lay people might interpret a neuroscientific explanation of offending; that being the offender's brain activity, at the point of the crime, reflected an unusual situation obstructing his *good* underlying 'real self'. An alternative interpretation also remains plausible: that the offender's brain activity, at the point of the crime, represented his *bad* underlying 'real self'. The chapter also considered some potentially confounding correlates of perceptions of the 'real self', such that these could be excluded from the design of the museum study.

Finally, this chapter presented the Deep Self Model – the theory that predicts people infer attributions of moral responsibility from perceptions of the Deep Self, a formal term for the 'real self' (Sripada, 2010). It was hypothesised that whereas genetic and environmental influences may be considered a distal, dispositional cause of offending, brain activity may be considered a proximal, situational cause of offending. This laid the theoretical groundwork for the empirical prediction of the museum study: that people will perceive the neurobiological offender's decision to offend as a *bad*, but situationally induced and temporary, deviation from his *good* underlying 'real self'. Consequently, the neurobiological offender may be blamed less for his offending behaviour.

Chapter 8: The impact of neurocriminological ideas on attributions of blame and the 'real self': The empirical work

People may sentence offenders in proportion to how far back the immoral intent is perceived to extend into the life history of the offender (Sripada, 2010). When that intent extends far back in time, the distal intent (or more broadly, the Deep Self) of the offender is conceived to be immoral and consequently, he may be blamed more. In this respect, the Deep Self Model provides an important conceptual backdrop to the museum study, which will be discussed in this chapter. The museum study documents the response of lay people to a universally applicable and causal explanation of every decision to give in to (or resist) impulse.

This chapter begins by comparing two mechanisms of neuroscientific influence: the mechanism discussed thus far – that being the perception of the brain as an obstruction to the expression of the 'real self' – and the mechanism proposed by Greene and Cohen (2004) – that being the perception of the brain as the product of ultimately uncontrollable influences. Should neurocriminology exert its influence through these mechanisms, the question is raised of whether each mechanism would lead people to reduce attributions of blame on legally relevant or legally irrelevant grounds. Subsequently, the chapter presents the findings of the museum study, which specifically sought to test whether, through one of these two mechanisms, the communication of neurocriminology would reduce attributions of blame to a hypothetical offender.

In the museum study, participants read either the proximal or distal level of explanation advanced by neurocriminology; that is either a neurobiological or biosocial explanation of offending. It was hypothesised that participants would seek a shorter prison sentence for the neurobiological offender as a result of perceiving him to have a more moral Deep Self. Note the hypothesised basis for this reduction in punitiveness

differs from the basis hypothesised by Greene and Cohen (2004): whereas they expect the neurobiological offender to be attributed less incompatibilist free will, I expect the neurobiological offender to be attributed a moral Deep Self. Only the former mechanism of neuroscientific influence requires people to perceive neurobiological causation as incompatible with free will; in other words, to have incompatibilist intuitions. Hence there are fewer cognitive prerequisites to the neuroscientific influence predicted by the Deep Self Model.

i. Two untested mechanisms of neuroscientific influence

I have argued that the most plausible mechanism of neuroscientific influence is offered by the Deep Self Model. Yet the question remains of whether the mechanism proposed by Greene and Cohen (2004) – that is the neuroscientific challenge to an incompatibilist free will – retains any influence in the criminal context. Although people still attribute responsibility to the conscious actor whose behaviour is explained by chance (Sripada, 2012b) or the brain (Mele, 2012; Shepard and Reuter, 2012; Nahmias et al., 2014), people may attribute *even more* responsibility to the actor whose behaviour arises from an incompatibilist free will. This prediction remains untested since none of these researchers included a baseline control condition – a condition in which the actions of the offender were unexplained or attributed to an incompatibilist free will. In contrast, the museum study did include an unexplained condition and items to test whether the effect of neurocriminology (relative to this baseline) reflected a change in attributions of incompatibilist free will; in other words, the survival of incompatibilist intuitions. Hence the museum study tested the relative strength of the mechanisms proposed by Greene and Cohen (2004) and the Deep Self Model.

Will there be a gap between the lay and legal response to neurocriminology?

In this chapter, I have raised two mechanisms by which a neurobiological description of the *conscious* decision to offend could mitigate blame: by shifting the perceived source of the decision, first, from an incompatibilist free will to an uncontrollable cause (Greene and Cohen, 2004) and second, from an enduring disposition to a temporary situation (Sripada, 2010). In turn, these two mechanisms represent the two grounds upon which the lay and legal responses to neurocriminology might differ. First, if the science mitigates blame by reducing attributions of an incompatibilist free will, this would represent a legally irrelevant, incompatibilist mechanism of neuroscientific influence (Greene and Cohen, 2004; Morse, 2004a; b; 2006; 2008; but see Kolber, 2016). Second, if the science mitigates blame by improving perceptions of the Deep Self, this would also represent a legally irrelevant (e.g., Fletcher, 1998), yet compatibilist, mechanism of neuroscientific influence: 'The law's criteria for determining culpability…do not include character as a salient factor' (Nadelhoffer et al., 2013a, p.208). Since I predict the second, compatibilist mechanism to bear greater influence, the basis of the gap between the legal and lay responses may also differ from the basis predicted by Greene and Cohen (2004).

One might doubt the longevity of this compatibilist mechanism. The mechanism requires people to perceive criminogenic brain activity to be distinct from the Deep Self; in other words, for people to make a situational attribution for the criminogenic brain activity. Yet, given evidence of dualistic beliefs (Bloom, 2004; Forstmann and Burgmer, 2015), one would expect future communications of neurocriminology to explicitly state the philosophical context of the empirical findings; for example, to state that '*every* mental difference is a function of some kind of difference in the brain' (Greene and Cohen, 2004, p.1779). These statements challenge the scope for a finding

about ‘brain activity’ to favour a situational (rather than dispositional) explanation of offending. Since ‘*every* mental difference’ is caused by the brain, the brain must reflect differences between people induced by both their situations and their dispositions. Consequently, people may be less likely to conclude that the brain obstructs the Deep Self and less willing to reach that conclusion since every (neurobiological) offender would then be eligible for neuroscientific mitigation. Ironically, therefore, if people accept the neurocriminological conception of the brain as a universal (i.e., normal) cause of offending, neurocriminology may lose mitigating influence. With the aim of modelling future communications of neurocriminology, the museum study encouraged participants to adopt this conception.

The overriding prediction remains that the punitive ‘gap’ between the lay and legal response to neurocriminology will be small: understood correctly, neurocriminology does not threaten the intentional and desiring stance of the offender at the time of the crime. In fact, by threatening the alternative, incompatibilist criterion for blame, neuroscience might actually motivate concern for these legally relevant, compatibilist justifications of blame (Nahmias et al., 2005). Therefore, the legally irrelevant, exculpatory responses to neurocriminology may be greatly limited by the stubborn instinct to punish wrongdoers (Crockett et al., 2014; Fehr and Gächter, 2002; Nadelhoffer et al., 2013b).

ii. The museum study

The museum study developed previous research by presenting neurocriminology in the manner that Greene and Cohen (2004) hypothesised to pose the greatest threat to positive retributivism. Specifically, neuroscience was presented not as an explanation of a specific mental illness (e.g., Aspinwall et al., 2012) or every decision ever made

(Nahmias et al., 2007) but, specifically, of every decision to resist or give into temptation. The museum study can be distinguished from previous research in eight respects:

1. For the caused offender (i.e., the offender who offended because of his brain or genes and environment), the relevant factors were described as the *complete* cause of the *conscious* decision to exercise *self-control* or not. This target of the explanation was chosen deliberately to challenge the lay conception, first, of self-control as free (Monroe and Malle, 2010) and, second, of brain processes as unconscious or consciously helpless (Nahmias et al., 2007). Participants also judged the *decision* to offend rather than the offending behaviour itself. In these respects, neurocriminology was presented in an ecologically valid context where the offending could not be mistaken as unconscious, unintended or undesired.
2. For the caused offender, the presented explanation was causal rather than correlational, as in the mock court approach, or deterministic, as in the approach of experimental philosophy. The participants were left to decide how they interpreted the notion of 'causal'; for example, as determining the offence to take place or probabilistically causing the offence to take place (i.e., increasing its likelihood of occurring). In this respect, the museum study tested the effects of communicating the causal premise of neurocriminology, without reference to determinism; since some neurocriminologists do not refer to determinism in their causal claims (e.g., Raine and Glenn, 2014), and neuroscience has been found to have greater effects in causal than deterministic scenarios (Shepherd, 2012).
3. In the vignette, the offender was described as a university graduate, with no mention of suffering mental difficulties. Since the offender therefore appeared mentally

normal, the study used neuroscience to explain the average offender rather than a psychotic offender; the focus of the mock court approach.

4. The vignette described a specific offender (rather than offenders generally) in the real world (rather than only a hypothetical world). This context was chosen to represent courtroom judgements and because Nahmias et al. (2007) found neuroscience to mitigate blame more in a real (than hypothetical) context. Judgements of real, specific, immoral actors tend to invoke the drive to blame (Nichols, 2011). Hence an unemotive offence type – opportunistic theft – was chosen to avoid providing any further incentive to resist neuroscientific mitigation. Opportunistic theft is also more common than the most serious offence types, and in turn, the scope for neuroscientific impact is broader.
5. The offender was described as committing *opportunistic* theft. The opportunistic nature of the crime implied that, in a different situation, the offence would never have been committed. Hence the vignette offered scope for participants to attribute the brain to an unusual situation and not only a disposition.
6. In the neurobiological explanation, brain activity – rather than the volume of a particular brain region – was the cause of the decision to offend. This represented a more universally applicable cause of offending that could also be attributed to an unusual situation (rather than only a disposition).
7. The explanation of the decision to offend never involved a human actor. Hence one could not attribute a reduction in blame to the displacement of blame from the offender to another human actor, such as an evil scientist (Waller, 2014).
8. The study tested whether lay people correctly perceive the brain, and genes and the environment, to be scientifically equivalent explanations of offending.

Hypotheses

This study tested the effects of lay people reading a neurobiological, biosocial or unexplained account of self-control on their attitudes towards an impulsive offender, first in a hypothetical context, and second, in a real-world context. Two hypotheses were proposed:

H1: Participants would be less punitive towards the neurobiological offender than the biosocial offender, and towards the biosocial offender than the unexplained offender.

H2: These effects would be mediated by moral perceptions of the 'person' who offended: the neurobiological offender would be considered a more moral person than the biosocial offender, and the biosocial offender would be considered a more moral person than the unexplained offender.

Method

Design

This experiment adopted a mixed model design. Three variables were manipulated between groups: first, the account of self-control presented in the hypothetical world (neurobiological, biosocial or unexplained), second, the account presented in the real world (neurobiological, biosocial or unexplained), and third, whether the public or prisoners comprised the intended audience of the judgements made about the real world (Table 8). Comparisons were also made between the hypothetical and real worlds – a variable that was manipulated within groups. At the start of the study, all participating visitors were asked to name the hypothetical offender in the scenario; only people who correctly named the offender were considered participants.

The study was conducted as part of the Live Science programme at the Science Museum, London. This programme enables researchers to collect data from museum visitors in a dedicated space of the 'Who Am I?' gallery in the Contemporary Science wing. All questions were answered through the online survey platform, Qualtrics. Participants responded using one of the three computers or three iPad Minis located on the central desk or using one of the three iPads distributed on the chairs around the perimeter of the space. A maximum of nine participants could take part at any one time between 11am and 6:30pm seven days a week from 15th August to 25th September 2016.

Table 8.

Summary of Conditions for the Museum Study (n = Number of Participants Within Each Real-World Condition).

Participants read that:					
The offence and the science were:			The decision to offend was explained by:		Participants read no explanation of the decision to offend
			The brain	Genes and the environment	
Hypothetical					
Real	Their judgements would be communicated to:	The general population	*n* = 836	*n* = 875	*n* = 842
		The prison population	*n* = 863	*n* = 813	*n* = 855

Participants

The 5,084 participants were museum visitors (mean age = 26.85, SD = 12.68). Of the participants who indicated their gender, 55% were female, 41% were male and 1% identified as other. While 53% of the sample were British, 122 different nationalities were represented among the respondents. Anyone aged 12 or older was eligible to

participate. The study was publicised on the Science Museum website, information screens positioned around the museum and flyers distributed around the gallery. Upon entering the 'Who Am I?' gallery, the majority of visitors were approached by the researcher or one of 10 undergraduate research assistants, volunteering at different times across the six weeks. Visitors received no payment for their participation.

Procedure

Participants read about a hypothetical planet called planet 'Erta' and its imaginary inhabitants called 'Ertans'. The text used to describe this planet was identical to the description presented by Nahmias et al. (2007):

> On planet Erta, the landscape is very similar to Earth, and there are advanced life forms called Ertans who look, talk, and behave very much like we do. For example, the Ertans have families, schools, various jobs, parties, arguments, etc. Science on planet Erta has advanced far beyond science on planet Earth, so Ertan scientists know far more about how their planet works.

A box labelled 'Ertans' temptations' was positioned on the central desk. Participants were instructed 'to imagine that we asked Ertans to write about a time when they faced a temptation and either resisted or gave into that temptation. We asked the Ertans to write down: 1. The temptation, 2. How they responded to that temptation. Then the Ertans placed their responses in the box labelled 'Ertans' temptations'. So the 'Ertans' temptations' box contains the made-up temptations of pretend people living on an imaginary planet called planet Erta'.

Next, participants took one folded piece of paper from the box. Although the following was not stated to participants, every piece of paper described the same temptation faced by the same hypothetical offender named Adam. Adam, a university graduate, wrote about his failure to resist the temptation to take and use a lost credit card to buy a luxury holiday; a full description of the offence can be found in Appendix C. After reading this full description, participants entered the name of the offender

onscreen supposedly so that the subsequent questions could be tailored to the 'unique' temptation printed on their piece of paper. Participants then returned the piece of paper to the researcher.

Next, the science was presented in the hypothetical context. Participants were randomly allocated to read one of three different accounts of 'why some Ertans have better self-control than others': an unexplained account, a biosocial account or a neurobiological account. In the neurobiological and biosocial accounts respectively, self-control was attributed to the brain or in equal proportions to genetic and environmental influences. These accounts were presented as *complete causal* explanations of 'why some Ertans have better self-control than others'. The full description of this science can be found in Appendix C.

After learning about the science, participants read: 'We would now like to ask you some questions about Adam the Ertan. When answering these questions, please assume that everything you've read about Ertans is true'. First, participants indicated their agreement (1 = *strongly disagree*, 6 = *strongly agree*) that it was Adam's fault that he decided to give in to the temptation. Participants also estimated the percent of museum visitors who would agree that the decision to give into temptation was Adam's fault. Subsequently, participants indicated their agreement that: a) Adam wanted to give in to this temptation, b) Adam's decision to give into temptation reflects him as a person, c) Adam could have decided to resist the temptation, d) Adam is likely to give into similar temptations in the future if nobody tries to stop him, e) Adam can be changed so that he stops giving into similar temptations in the future (which was displayed only when participants at least 'somewhat agreed' with the preceding question), and f) the victim was capable of preventing Adam from giving into this temptation. Finally, participants recommended an appropriate punishment for Adam;

the only available option was a whole number of weeks in prison from zero to 36. Before making their decision, participants were informed that in England and Wales, the maximum punishment for this type of theft was 36 weeks (i.e., 9 months) – a figure based on the sentencing guidelines of England and Wales.

After completing these questions, participants were transitioned from the hypothetical world to the real world. They read: 'Until now we told you that Adam is an imaginary Ertan living on an imaginary planet called planet Erta. Adam is actually a real human being living on planet Earth who visited the Science Museum in July this year. In July this year, our research team left a box in the Science Museum. Next to the box were a set of instructions. The instructions asked visitors to the Museum to write about a time when they faced a temptation and either resisted or gave into that temptation. We asked the visitors to write down: 1. The temptation, 2. How they responded to that temptation. Then the visitors placed their responses in a box. That box is the same box that is here today labelled 'Ertans' temptations'. We typed up and printed the responses to ensure the quality of the handwriting was the same across the different pieces of paper. So, the response you picked out of the box was written by a visitor to the Science Museum in July this year. All we know about the visitor is that he called himself Adam'.

In fact, Adam was a hypothetical character, not a real offender or a real visitor to the museum. However, real visitors were asked to write down their temptations for the first two days of the study and 18 of the 120 visitors who completed this task reported committing opportunistic theft. This finding minimised the ethical implications of deceiving participants into believing that Adam was a real visitor who had actually committed opportunistic theft; the seriousness of Adam's offence corresponded to the seriousness of some of the offences disclosed by real museum visitors.

Following this transition to the real offender, the science also became real. Participants read:

> Until now we told you that scientists on planet Erta have discovered why some Ertans have better self-control than others. Real scientists in the real world (planet Earth) have also discovered why some humans have better self-control than others.

Participants were then randomly allocated to read one of the three accounts of self-control that were also randomly presented in the hypothetical world: the unexplained account, the biosocial account or the neurobiological account. However, this time, the account was presented as a *real*, complete and causal explanation of why some *real* people have better self-control than others and why Adam, the *real* offender, gave into temptation. Since the type of science presented in the hypothetical and real worlds was chosen randomly, the science presented in the real world was different from the science presented in the hypothetical world for two-thirds of participants.

Following this transition to the real offender and real science, participants answered the same questions that had previously been posed in respect to the hypothetical offender (Adam the Ertan). This time, participants answered about the real offender (Adam the museum visitor) without any instruction to assume the science was true. Before answering the questions, participants were randomly allocated to read one of the two variations of the following message: 'It is important that you answer carefully because we hope to publish the results of this experiment' either 'in a *national* newspaper read by *people* across the country' or 'in a *prison* newspaper read by *prisoners* across the country'. Therefore, in the real-world condition, participants used real science to make judgements about a real offender that would be communicated to real people – three features that were absent from the hypothetical world condition.

After completing these questions, the first 120 of the 5,084 participants completed the temptation task that Adam had supposedly completed. These participants wrote down their response to a temptation faced in the past, then selected whether they gave into or resisted that temptation and accordingly, indicated how much their 'response to the temptation was a morally *good* [or *bad*] thing to do' (1 = *Not at all good/bad*, 6 = *Extremely good/bad*). It was originally intended that all participants would complete this task, but, on the third day, the museum stopped the task: visitors were reporting temptations that were more severe than the museum expected, raising ethical and legal issues for the museum.

Nevertheless, another measure of deviance remained in the study: all participants reported how often they had committed eight acts of minor deviance in the previous five years (*Never*, *Once*, *Twice*, *3 or 4 times* or *5 times or more*). These eight items were taken from the measure of everyday crime used by Karstedt and Farrall (2006). The acts were: illegal downloading, agreeing to pay or be paid in cash to avoid taxes, failing to disclose faults when selling a second-hand item, wardrobing (returning worn clothes), making an unjustified insurance claim, making other unjustified compensation claims, receiving speeding fines and travelling by train without a valid ticket. The scale of everyday deviance showed moderate internal consistency (Cronbach's $\alpha = .56$).

In the final bank of questions, participants rated the believability of the real science (1 = *strongly disbelieved*, 6 = *strongly believed*). Participants also indicated the country that represented their nationality, their age, their gender, their willingness to complete a five-minute survey in a few months' time, their interest in receiving a summary of the results and if so, their email address. Finally, participants were

debriefed verbally and handed a debrief card, with attention directed towards the hypothetical nature of Adam the offender.

The follow-up survey

At 5pm on Monday 28th November 2016, a link was sent to all participants who agreed to receive the follow-up survey. Of the original 5,084 participants, 2,572 (51%) agreed to receive the follow-up survey. Participants were given a week (until 9am on Tuesday 6th December) to respond. Therefore, participants could complete the follow-up survey between 2 months 11 days and 3 months 21 days after their original participation date. The average participant completed the follow-up 2 months 28 days after their original participation date (*SD* = 11.50 days).

With the aim of maximising the response rate, the follow-up survey contained only six randomly ordered items: three items from the Free Will Inventory (Nadelhoffer, Shepard, Nahmias, Sripada, and Ross, 2014) and three items from a measure of punitiveness (Maruna and King, 2009). Punitiveness (Cronbach's $\alpha = .54$) and belief in free will (Cronbach's $\alpha = .66$) showed sufficient internal consistency for such short (three-item) scales.

Participants were sent unique links to this survey; this ensured their follow-up responses could be linked to their original participant number – and so to their experimental condition and date of participation at the museum. Participants who did not complete the follow-up survey received a reminder email at 6am on Saturday 3rd December. Of the 2,572 participants who received the follow-up survey, 1,344 (52%) responded; in other words, 26% of the original sample.

Results

The significance level of $p < .05$ was adjusted using the Bonferroni correction for all pairwise comparisons.

The hypothetical world

Perceptions of the hypothetical offender were analysed using a between-subjects MANOVA, with the Hypothetical Explanation [Unexplained, Biosocial, Neurobiological] as the between-subjects factor. Table 9 shows the main effects of the Hypothetical Explanation. In comparison to the Biosocial and Neurobiological offenders, participants were significantly more likely to attribute fault to the Unexplained offender (Unexplained: $M = 5.34$, $SD = 1.53$; Biosocial: $M = 4.87$, $SD = 1.55$; $p < .001$; Neurobiological: $M = 4.75$, $SD = 1.69$; $p < .001$) and expect other visitors to attribute fault (Unexplained: $M = 77.37$, $SD = 16.80$; Biosocial: $M = 71.90$, $SD = 18.30$; $p < .001$; Neurobiological: $M = 71.02$, $SD = 19.69$; $p < .001$). Participants were significantly more likely to believe the Unexplained offender ($M = 5.95$, $SD = 1.21$) could have decided to resist temptation than the Biosocial offender ($M = 5.54$, $SD = 1.10$; $p < .001$), and significantly more likely to believe the Biosocial offender could have decided to resist than the Neurobiological offender ($M = 5.33$, $SD = 1.59$; $p < .001$). In comparison to the Unexplained and Biosocial offenders, participants were significantly less likely to believe that the decision reflected the Neurobiological offender as a person (Unexplained: $M = 4.54$, $SD = 1.32$; $p < .001$; Biosocial: $M = 4.51$, $SD = 1.35$; $p = .003$; Neurobiological: $M = 4.35$, $SD = 1.40$) and participants punished the Neurobiological offender less (Unexplained: $M = 15.45$, $SD = 11.78$; $p = .023$; Biosocial: $M = 15.51$, $SD = 11.63$; $p = .016$; Neurobiological: $M = 14.45$, $SD = 11.45$).

Table 9.

ANOVA Results: Effects of the Hypothetical Explanation.

Response variables	*F*	*df*	*p*	η_p^2
His fault	52.36	2, 4180	< .001	.02
Others would agree his fault	52.36	2, 4180	< .001	.02
Reflects him as a person	7.61	2, 4180	.001	< .01
Could have resisted	70.00	2, 4180	< .001	.03
Sentence	3.64	2, 4180	.026	< .01

Mediation analysis

The MANOVA identified two perceptions that were uniquely affected by the Neurobiological Explanation (and not also the Biosocial Explanation). Mediation analysis was conducted to test for the relationship between these two perceptions. The Neurobiological Explanation was recoded as a dummy variable (1 = Neurobiological Explanation, 0 = some other Hypothetical Explanation), such that it could be entered as a predictor of the recommended sentence in a series of linear regressions (Table 10). In the first model, the Neurobiological Explanation was a significant negative predictor of the sentence (β = -.04, *p* = .006). In the second model, the perceived probability of reoffending (β = .08, *p* = < .001) and perceived changeability of the offender (β = -.04, *p* = .007) were entered, and acted as significant predictors, while the Neurobiological Explanation remained equally significant (β = -.04, *p* = .004). In the third model, perceived intentionality was entered, and also acted as a significant predictor (β = .08, *p* < .001), while the significance of the Neurobiological Explanation remained the same (β = -.041, *p* = .007). In the fourth model, the capacity to decide to resist was entered and acted as a significant predictor (β = .08, *p* < .001) that reduced the significance of the Neurobiological Explanation (β = -.03, *p* = .04). In the fifth model, moral perceptions of the person were entered and acted as a significant predictor (β = .12, *p* < .001); consequently, neither the Neurobiological Explanation (β = -.03, *p* = .069) nor

the perceived probability of reoffending (β = .03, *p* = .063) predicted the sentence. This suggests that moral perceptions of the person mediated the effect of the Neurobiological Explanation (and the perceived probability of reoffending) on the sentence.

Table 10.

Mediation Analysis: Predictors of the Sentence in the Hypothetical World (Significant Predictors Emboldened).

	Model 1		Model 2		Model 3		Model 4		Model 5	
Response variables	β	*p*	β	*p*	β	*P*	β	*P*	β	*p*
Neurobiological Explanation	**-.04**	.006	**-.04**	.004	**-.04**	.007	**-.03**	.04	-.03	.069
Is likely to reoffend			**.08**	<.001	**.06**	<.001	**.06**	<.001	.03	.063
Can be changed			**-.04**	.007	**-.04**	.004	**-.06**	<.001	**-.06**	<.001
Wanted to offend					**.08**	<.001	**.06**	<.001	**.04**	.015
Could have resisted							**.08**	<.001	**.06**	<.001
Reflects him as a person									**.12**	<.001

The real world

Perceptions of the real offender were analysed using a between-subjects MANOVA, with the Real Explanation [Unexplained, Biosocial, Neurobiological], the preceding Hypothetical Explanation [Unexplained, Biosocial, Neurobiological] and the Target Audience [General Population, Prison Population] as between-subjects factors. Table 11 shows the main effects of the Real Explanation. In comparison to the Biosocial and Neurobiological offenders, participants were significantly more likely to attribute fault to the Unexplained offender (Unexplained: *M* = 5.45, *SD* = 1.28; Biosocial: *M* = 5.09, *SD* = 1.36; *p* < .001; Neurobiological: *M* = 5.00, *SD* = 1.39; *p* < .001), expect other visitors to attribute fault (Unexplained: *M* = 78.73, *SD* = 17.17; Biosocial: *M* = 73.37, *SD* = 19.39; *p* < .001; Neurobiological: *M* = 73.24, *SD* = 20.69; *p* < .001), believe the

Unexplained offender wanted to offend (Unexplained: *M* = 4.55, *SD* = 1.36; Biosocial: *M* = 4.28, *SD* = 1.32; *p* < .001; Neurobiological: *M* = 4.26, *SD* = 1.34; *p* < .001), believe that the decision reflected him as a person (Unexplained: *M* = 4.74, *SD* = 1.32; Biosocial: *M* = 4.52, *SD* = 1.31; *p* < .001; Neurobiological: *M* = 4.48, *SD* = 1.38; *p* < .001) and believe that the Unexplained offender should be punished (Unexplained: *M* = 16.65, *SD* = 12.27; Biosocial: *M* = 15.53, *SD* = 11.68; *p* = .016; Neurobiological: *M* = 14.85, *SD* = 11.63; *p* < .001). Participants were significantly more likely to believe the Unexplained offender could have decided to resist temptation (Unexplained: *M* = 5.66, *SD* = 1.25) than the Biosocial offender (Biosocial: *M* = 5.30, *SD* = 1.32; *p* < .001), and more likely to believe the Biosocial offender could have decided to resist than the Neurobiological offender (Neurobiological: *M* = 5.17, *SD* = 1.40; *p* = .009). Finally, participants were significantly less likely to believe the Neurobiological Explanation (*M* = 4.18, *SD* = 1.07) than the Biosocial Explanation (*M* = 4.34, *SD* = .97; *p* < .001) or the Unexplained Account (*M* = 4.34, *SD* = .91; *p* < .001).

Table 11.

ANOVA Results: Effects of the Real Explanation.

Response variables	*F*	*Df*	*P*	η_p^2
His fault	42.64	2, 4100	< .001	.02
Others would agree his fault	36.38	2, 4100	< .001	.02
Wanted to offend	19.05	2, 4100	< .001	< .01
Reflects him as a person	15.39	2, 4100	< .001	< .01
Could have resisted	51.68	2, 4100	< .001	.03
Sentence	7.57	2, 4100	.001	< .01
Explanation was believable	12.37	2, 4100	< .001	< .01

Mediation analysis

The MANOVA identified five perceptions that were equally affected by the Biosocial and Neurobiological Explanations. Mediation analysis was conducted to test for

relationships between these perceptions. The Unexplained Account was recoded as a dummy variable (1 = present, 0 = absent), such that it could be entered as a predictor of the recommended sentence in a hierarchical linear regression (Table 12). In the first model, the Unexplained Account was a significant positive predictor of the sentence (β = .06, $p < .001$). In the second model, the perceived probability of reoffending (β = .12, $p < .001$) and the perceived changeability of the offender (β = -.02, $p = .14$) were entered; the former and the Unexplained Account (β = .05, $p < .001$) acted as significant predictors. In the third model, perceived intentionality was entered, and also acted as a significant predictor (β = .11, $p < .001$), while the significance of the Unexplained Account remained the same (β = .04, $p = .004$). In the fourth model, the perceived capacity to decide to resist was entered and acted as a significant predictor (β = .10, $p < .001$) that reduced the significance of the Unexplained Account (β = .03, $p = .034$). In the fifth model, moral perceptions of the person were entered and acted as a significant predictor (β = .13, $p < .001$). Consequently, the Unexplained Account no longer predicted the sentence (β = .03, $p = .07$). This suggests that moral perceptions of the person mediated the effect of either causal Explanation on the sentence.

Table 12.

Mediation Analysis: Predictors of the Sentence in the Real World (Significant Predictors Emboldened).

Response variables	Model 1		Model 2		Model 3		Model 4		Model 5	
	β	*p*	β	*p*	β	*P*	β	*P*	β	*p*
Unexplained Account	**.06**	<.001	**.05**	<.001	**.04**	.004	**.03**	.034	.03	.07
Is likely to reoffend			**.12**	<.001	**.10**	<.001	**.09**	<.001	**.05**	.001
Can be changed			-.02	.14	-.02	.125	**-.04**	.005	**-.04**	.007
Wanted to offend					**.11**	<.001	**.08**	<.001	**.05**	.006
Could have resisted							**.10**	<.001	**.08**	<.001
Reflects him as a person									**.13**	<.001

Summary

The results provided some support for both hypotheses. In the hypothetical world, participants recommended a shorter sentence for the neurobiological offender than the biosocial offender, and in the real world, participants recommended a shorter sentence for the biosocial offender than the unexplained offender (H1). These effects were mediated by moral perceptions of the 'person' who offended (H2).

Discussion

In the hypothetical world, participants recommended a shorter sentence for the neurobiological offender than for the biosocial offender or the unexplained offender. In the real world, participants recommended a shorter sentence for the biosocial offender and the neurobiological offender than for the unexplained offender. These findings do not support the first hypothesis exactly yet do support the more general notion that neurocriminology has greater influence than biosocial criminology. In spite of being less believable, the neurobiological explanation always mitigated the sentence, whereas the biosocial explanation only mitigated in the real world. This finding poses policy

implications for science communicators, such as expert witnesses in court. The challenge for communicating neurocriminology may be persuading people that the brain actually causes offending. In contrast, the challenge for communicating biosocial criminology may be persuading people that biosocial causation justifies sentence mitigation.

It is surprising that the biosocial explanation exerted a greater effect in the real world rather than the hypothetical world; given that participants were only instructed to assume the science was true in the hypothetical world. However, perhaps there is greater scope for science to change attitudes towards real people; in a previous study, for example, the neurobiological explanation reduced blame more in the real world than the hypothetical world (Nahmias et al., 2007). Since people have greater incentive to blame real offenders (Nahmias et al., 2007; Roskies and Nichols, 2008), there may also be greater scope for the science to undermine blame in the real world (i.e., less risk of floor effects).

While previous research has found people still blame the neurobiological offender, this research cannot conclude there was *no* effect of describing the neuroscientific argument for determinism (Nahmias, 2006; Nahmias et al., 2007) or the neuroscientific capacity to predict behaviour (Mele, 2012; Shepard and Reuter, 2012; Nahmias et al., 2014). These studies did not include the control condition that would be necessary to assert there were no effects of the science. In contrast, the current study did, finding that, while the effects were small, the science did promote less punitive perceptions of the offender. When either explanation of criminal intent was presented, participants were less likely to: attribute fault to the offender, believe that other visitors would attribute fault, believe that the offender could have decided to resist offending and in the real world, believe the offender wanted to offend. For these measures,

therefore, participants inferred the same implications from the scientifically equivalent, neurobiological and biosocial, explanations.

Although the effects were small, participants did respond differently after reading the science; this suggests they did not adopt the neurobiological or biosocial perspectives by default. This finding is consistent with prior evidence that lay people perceive genetic and neurobiological factors to play a minor role in crime causation (e.g., Gabbidon and Boisvert, 2012; Gajos et al., 2014). It is also clear that the participants believed their criteria for blame were shared by the population at large.

Nahmias et al. (2007) found that neuroscience could derive mitigating influence from the misperception that the neurobiological offender does not make a *conscious* decision to offend. This misperception should be relatively easy to correct. In the future, therefore, one hopes that responsible communications of neuroscience do not let this misperception go unchecked. The current study aimed to correct this misperception by explicitly and repeatedly stating, in every condition, that the offender made a conscious decision to offend. Participants were asked to assume this was a fact about the hypothetical offender, which might explain a discrepancy between the hypothetical and real judgements. In the hypothetical world, participants believed the neurobiological and biosocial offenders wanted to offend just as much as the unexplained offender. In contrast, in the real world, participants believed the unexplained offender wanted to offend more than the neurobiological or biosocial offenders; in this context, participants could decide whether or not to believe the assertion that the decision was conscious. Perhaps, therefore, the biosocial and neurobiological offenders were believed to have made a less conscious decision and consequently, were attributed less desire to offend. Nevertheless, while the science affected attributions of desire, this effect could not explain why the science also reduced the sentence.

How did neuroscience affect moral perceptions of the person?

The neurobiological explanation reduced the sentence by improving moral perceptions of the person in both the hypothetical and real worlds. In contrast, the biosocial explanation reduced the sentence by improving moral perceptions of the person only in the real world. The mediating role of person perceptions supports the Deep Self Model (Locke and Frankfurt, 1975; Sripada, 2010). Since biosocial factors – genes and the social environment – exert their influence across the lifespan, such distal factors are most suited to explaining the disposition of the offender. In contrast, brain activity is ambiguous: 'chemical reactions and electrical activity' could be attributed to the disposition or the situation in which the offender finds himself.

In the real world, it seems that participants believed the criminogenic influences – either the brain or genes and the social environment – had obstructed a more moral Deep Self. In turn, this perception justified a reduction in blame, corroborating prior evidence that perceptions of the real self influence blame (Newman et al., 2015). In the hypothetical world, however, it seems that participants believed only the brain could have obstructed a more moral Deep Self. This finding suggests that people more consistently separate the 'person' from his brain than the 'person' from his genes and social environment. Instead of the criminogenic brain activity being attributed to the Deep Self, it may be attributed to the temporary and unusual situation in which the offender found himself. In turn, the neurobiological imprint of that situation may act as a convincing mitigating factor. While future research is required to test this interpretation directly, the current study finds that neurocriminology robustly improves perceptions of the offending person, with mitigating consequences.

The vignette provided specific details that may have facilitated a situational interpretation of the brain. First, the criminal intent was attributed to brain *activity*, not

the volume of particular brain regions or pathways: people may more intuitively perceive brain activity as continuously changing with the varying input of the external situation – and therefore as distinct from the Deep Self. Second, the offender was described as a fellow visitor to the Science Museum and a university graduate, who had offended whilst returning from vacation with friends and for the intention of taking his girlfriend on a holiday he could not afford. These details may have acted as cues to his ingroup membership, prosocial disposition and normal mental functioning, and by inference, the morality of his Deep Self. Therefore, the particular vignette may have enabled a situational interpretation of neurocriminology.

How did neuroscience affect attributions of self-control?

The science mitigated sentences not only by improving moral perceptions of the person offending but also by reducing his perceived capacity to decide to resist temptation. In both worlds, the neurobiological offender was deemed less capable of deciding to resist temptation than the biosocial offender, who, in turn, was deemed less capable of deciding to resist temptation than the unexplained offender. Therefore, a neurobiological source of self-control was considered less capable of delivering self-control than a biosocial source, which, in turn, was considered less capable than an unexplained source. This reluctance to perceive the brain as a strong source of self-control has no scientific basis: 'For the neuroscientist, controlled processes (that may go "against" the initial impulse) are just as much neural states determined by their preceding states' (Monroe and Malle, 2010, p.222). Therefore, the participants appear to adopt the dualistic belief that self-control is not a material construct.

It is unclear how exactly participants interpreted the capacity to decide to resist temptation. On one hand, participants may have judged whether the offender had a general capacity to resist this temptation, for example, in a different situation; this

represents a notion of choice that is compatible with determinism (i.e., caused self-control). On the other hand, participants may have judged whether the offender could have decided to resist this temptation, even in the exact same situation with the exact same preceding circumstances; this represents a notion of choice that is incompatible with determinism (i.e., incompatibilist free will). Consequently, one cannot determine whether participants made a compatibilist or incompatibilist interpretation of this question, and therefore, which interpretation explains the mitigation of the sentence that appears to be premised on some understanding of the science.

If the science had reduced the sentence by reducing belief in incompatibilist free will, this would support the mechanism of neuroscientific influence hypothesised by Greene and Cohen (2004). However, there are two strong reasons to suggest the participants did not make this incompatibilist interpretation of the self-control question. First, the vignette repeatedly described the neurobiological and biosocial offenders as having no *uncaused* capacity to resist temptation; their (lacking) capacity to resist temptation was entirely caused by prior causal influences. Therefore, if the participants had interpreted the capacity to resist temptation as the *uncaused* capacity to resist temptation, one would expect a stark difference between the self-control attributed to the caused and unexplained offenders. In fact, the difference was very small: in the unexplained, biosocial and neurobiological conditions, 95.8%, 92.7% and 90.9% of participants (at least somewhat) agreed that the offender could have decided to resist temptation. These high percentages suggest that when answering the self-control question, participants were not judging the degree of *uncaused* self-control; which was zero.

Second, I have discussed much evidence that lay attributions of responsibility tend to persist despite causal arguments (e.g., Nahmias et al., 2005), especially in the

criminal context (Nichols and Knobe, 2007). Hence it was not hypothesised that sentence mitigation would result from the perceived loss of incompatibilist free will. In fact, Monroe and Malle (2010) found not a single participant who spontaneously defined free will as a first cause or an uncaused cause; perhaps because people do not conceive that self-control could possibly be caused by another factor (Kolber, 2016). It is therefore unlikely that the current participants spontaneously interpreted the self-control question as a reference to *uncaused* self-control in particular.

For these two reasons, I suspect that participants intuitively made the compatibilist interpretation of the self-control question. In other words, it is most plausible that participants judged the general capacity to resist this type of temptation rather than the free capacity to act against the exact same factors that caused the actor to give in to temptation on this occasion. Accordingly, the science mitigated the sentence by reducing the perceived capacity for the offender to exercise rational choice.

How limited were the effects of neuroscience?

Morse (2004a) refers to the 'fundamental psycho-legal error' as the tendency to perceive evidence of ultimate causation as exculpatory. The current findings do not entirely rule out the possibility that the science mitigated the sentence by suggesting the offender lacked incompatibilist free will. However, even if the participants did mitigate the sentence as a result of believing in ultimate causation, it would be misleading to term this basis for mitigation a 'fundamental' error. Every difference between perceptions of the caused and unexplained offenders was small considering that such a complete and explicitly causal explanation of offending was presented. Given the large sample size, it is important to note not just the significance, but also the size, of the reported effects (Table 13). For example, in the unexplained, biosocial and neurobiological conditions respectively, 92.9%, 89.7% and 88.1% of participants

agreed that the decision to offend was the offender's fault; clearly, therefore, attributions of fault were only slightly affected by the causal argument. Hence this study challenges the claim that people have a 'fundamental' tendency to perceive causation as exculpatory.

Table 13.

The percent (%) of participants who (at least somewhat) agreed that the following traits could be attributed to the (real) unexplained, biosocial and neurobiological offenders. Only participants who read the same explanation in the hypothetical and real worlds are included here.

Response variables	The unexplained	The biosocial	The neurobiological
		…offender	
His fault	92.9	89.7	88.1
Others would agree his fault	77.1	74.1	73.0
Wanted to offend	78.7	73.2	72.8
Reflects him as a person	80.3	79.6	76.0
Could have resisted	95.8	92.7	90.9
Mean sentence (weeks)	15.22	14.96	13.76
Explanation was believable	87.4	85.4	80.6

The science exerted only small effects on attributions of fault, immoral desire, an immoral self, self-control and punishment. Hence this study supports neither the mechanism nor the scale of the prediction made by Greene and Cohen (2004, p.1784); that being 'retributivist notions of criminal responsibility will give way…thus radically transforming our approach to criminal justice'. Perhaps the authors only expected such results if neuroscience was communicated in a more engaging format than several paragraphs of text in a short online survey. However, larger effects were not observed in the theatre study, which did present the science in a more engaging format and

dispersed across time. Nevertheless, perhaps the science would have larger effects if it had been presented over a more sustained period than several paragraphs of text in a short online survey. For example, one would expect larger effects if neurocriminology was taught as part of the national curriculum to every child at a more impressionable age.

However, I would contest the idea that the current intervention was weak: participants read a *complete causal* explanation of offending and, in the hypothetical world, were instructed to assume the explanation was true. In this respect, the intervention linked the neuroscientific explanation to the judged offending for the participant, communicating the common philosophical assumptions of neuroscience – specifically, physicalism and universal causation, which could not possibly be documented by current neuroimaging technologies (Morse, 2008a). Current technologies are limited to documenting weak correlations between specific brain networks and particular antisocial behaviours (Raine, 2013). The current intervention therefore progressed far beyond the limits of current neurocriminological data and in this respect, was not weak. The key ideas were also repeated across four pages of the survey.

The target of attitude change was also deliberately modest: participants were asked to judge the theft committed by an educated fellow visitor to the museum and first in a hypothetical world, not the emotive actions of a repeat violent offender immediately in the real world. These details of the crime were specifically designed to invoke incompatibilist intuitions (Nahmias et al., 2007, 2006; Nichols and Knobe, 2007; Roskies and Nichols, 2008), and thereby facilitate the mechanism of neuroscientific mitigation hypothesised by Greene and Cohen (2004). Yet still, the mitigating influence remained small.

The limited effects of scientific mitigation did not result from participants simply refusing to believe the science. In fact, participants were as likely to believe the biosocial account as the unexplained account and less (but still highly) likely to believe the neurobiological account. Specifically, 87.4%, 85.4% and 80.6% of participants (at least somewhat) believed the unexplained, biosocial and neurobiological accounts respectively. Similar percentages were observed by Nahmias et al. (2014, p.505): 80-82% of participants believed 'it is possible for [neuropredictive] technology to exist in the future'. Monroe and Malle (2010, p.217) found that 74% of participants objected to a neuroscientific argument against free will, yet only 24% of those participants objected by rejecting neurobiological causation; specifically, the part of the argument that claimed 'all of our behaviour is caused by our neural impulses'. This suggests that the reluctance to relinquish belief in free will is not typically a reluctance to believe in neurobiological causation. In fact, people are receptive to the causal role of the brain; perhaps because the implications for blame are deemed so minor. The compatibilist stance of lay people may facilitate their acceptance of neuroscience (Shepherd, 2012).

Compatibilism would also explain the failure for neuroscience to exert any major mitigating effects. The participants correctly understood the neurobiological offender to have the capacity to exert rational choice (i.e., to form and act upon intent using self-control), thereby justifying blame. Indeed, this compatibilist style of reasoning is more prevalent in judgements of a specific offender as opposed to behaviour in general (e.g., De Brigard et al., 2009). The nature of the communicated science may have also promoted compatibilist responding: the science was used to explain *every* conscious decision to resist or give into temptation. Therefore, if the offender in the vignette was excused on the basis of causation per se (i.e., on incompatibilist grounds), this would logically require *every* impulsive offender to be

excused; since the vignette described every decision to act on impulse as caused. Hence the sheer applicability of this excuse might motivate its resistance (Shepherd, 2012) and the adoption of alternative compatibilist criteria for blame (Nahmias et al., 2005).

Compatibilist responding may have also been motivated by the inability to displace blame onto prior human actors in the causal chain; since the scientific explanation also applied to their decisions. This absence of the opportunity to displace blame may have reduced the willingness for people to make a mitigating interpretation. In contrast, narratives of child abuse do permit displacement: narratives of the abused offender increase disdain for humanity more generally and as a result of causing a small (but significant) reduction in the attribution of blame to the offender (Gill and Getty, 2016). In this study, it is unclear whether disdain represents the belief that the abuser can be blamed or an acceptance that nobody can be blamed. If mitigation is facilitated by the perceived presence of a scapegoat (Waller, 2014), the absence of one in the museum study may explain the limited effects observed.

What were the effects of sequentially presenting the hypothetical and real worlds?

Thus far I have considered the mechanisms by which neurocriminology could influence the positively retributive punishment of offenders. I have already hypothesised that the mechanism of neuroscientific influence may change with the contextual features of the crime and science described in the vignette (e.g., Shepherd, 2012). Those features can change the motivation to blame and consequently, the willingness to engage criteria for blame that would invite a mitigating response to causal explanation (Nichols, 2011). People have more motivation to blame and punish a real offender than a hypothetical offender (Roskies and Nichols, 2008), given the punitive emotions stirred by the harm resulting from real crime (e.g., Nichols and Knobe, 2007). In the current study, such

punitiveness may have been amplified by the fact that participants were instructed their judgements would be publicised (to either prison or general populations) only in the real-world condition.

The museum study adopted a novel approach to reducing the probability that instinctive punitiveness would impede receptivity to scientific explanation. Participants first read that the offender and the causal explanation were hypothetical. By first presenting the causal explanation in a context known to increase its mitigating influence (Nahmias et al., 2007; Roskies and Nichols, 2008), participants may have subsequently been more likely to sustain that mitigating interpretation in the real world. This prediction assumes people seek to avoid cognitive dissonance; the conflict between opposing attitudes (Festinger, 1957). The sequential presentation of the hypothetical and real worlds may discourage people from switching sides of the philosophical debate, thereby harnessing the initial receptivity to causation in the hypothetical world.

Although no previous research has sequentially presented causal arguments, a sequential presentation style has been used to effectively change attitudes towards concepts other than moral responsibility. For example, two studies found that participants reported less prejudice towards real homosexuals after having imagined living in a hypothetical world where relationships were characterised by the features of homosexual relationships; such as stigma regarding cohabitation and public displays of affection (Hillman and Martin, 2002; Hodson, Choma and Costello, 2009). This attitudinal intervention, a simulation task known as ‘Alien Nation’, increased empathy for homosexuals and their common identities with heterosexuals, thereby reducing prejudice (Hodson et al., 2009).

In the hypothetical world of Alien Nation, participants do not recognise the characteristics of homosexual relationships to be the characteristics of *homosexual* relationships; consequently, people do not perceive their judgements of Alien Nation to represent their judgements of homosexual communities. Similarly, in the museum study, participants may not have perceived their judgements of the hypothetical offender to represent their judgements of real offenders. After participants have expressed sympathy for the characters of Alien Nation, research assistants highlight the parallels between these characters and real homosexuals, asking participants to report their opinions of the latter. Similarly, in the museum study, the hypothetical offender becomes real, with attitudes then reported. If this format of sequential presentation can improve attitudes towards sexual minorities (Hillman and Martin, 2002; Hodson et al., 2009), perhaps it can also improve attitudes towards other minorities, namely, offenders.

Hodson et al. (2009, p.977) explain:

> The [Alien Nation] simulation apparently frees participants to explore a hypothetical threat-free outgroup-relevant context, gaining perspective about situational constraints imposed on lifestyle choices and behaviours. [The researcher] then orients this perspective-taking experience onto…a real-life target unobtrusively relevant to the simulation.

This rationale inspired the use of sequential presentation in the current study: participants who first encountered neuroscience in a hypothetical – and so threat free – context may have been freer to recognise neuroscientific reasons to mitigate punishment in the real world.

The current study

In respect to sequential presentation, the museum study can be distinguished from previous research in two further respects:

1. Participants responded to the vignette, first, believing that the science and the offender were hypothetical, and second, believing that both were real. This approach aimed to introduce the science in a less emotive, and so more receptive, context.
2. In contrast to the approach of experimental philosophy, the current participants were led to believe that the offender – and not only the science – was real in the real-world condition. This condition was therefore a more accurate representation of real-life judgements in which the offender is a real person.

Hypotheses

These novel aspects of the study generated two further hypotheses:

H3: Participants would be more punitive towards the real offender than the hypothetical offender.

H4: Participants would be less punitive towards the real caused offender after first reading the same causal explanation in the hypothetical world (as opposed to a different explanation or no explanation).

The real caused offender refers to the offender whose real-life criminal behaviour was described as the causal effect of the brain or genes and the environment.

Results

In contrast to previous research (e.g., Nahmias et al., 2007), this study manipulated hypothetical and real worlds using a repeated measures design. Consequently, the manipulation of the hypothetical and real worlds was confounded by whether the crime and the scientific explanation were new (the hypothetical world) or potentially familiar (the real world). The repeated measures design did provide the opportunity to assess

whether people changed their responses in the transition from the hypothetical world to the real world. To remove the confounding nature of how novel the scientific explanation was, only participants who read the same explanation in the hypothetical and real worlds were included in this analysis: perceptions of the offender were analysed using a mixed-model MANOVA, with the Explanation Type [Unexplained, Biosocial, Neurobiological] as the between-subjects factor and the World Type [Hypothetical, Real] as the repeated-measures factor.

Table 14 shows the main effects of the World Type. In comparison to the Hypothetical offender, participants were significantly more likely to attribute fault to the Real offender (Hypothetical: M = 5.00, SD = 1.64; Real: M = 5.20, SD = 1.36, d = .13), expect other visitors to attribute fault to the Real offender (Hypothetical: M = 72.78, SD = 18.90, d = .14; Real: M = 75.46, SD = 19.28), believe that the Real offender wanted to give into temptation (Hypothetical: M = 4.23, SD = 1.43; Real: M = 4.39, SD = 1.35, d = .12), believe that his decision to offend reflected him as a person (Hypothetical: M = 4.53, SD = 1.39; Real: M = 4.65, SD = 1.34, d = .09) and believe that he should be punished (Hypothetical: M = 14.45, SD = 11.50; Real: M = 15.05, SD = 11.81, d = .05). Yet participants were also significantly less likely to believe the Real offender could have decided to resist temptation (Hypothetical: M = 5.64, SD = 1.45; Real: M = 5.41, SD = 1.33, d = .17) and that the Real offender is likely to reoffend (Hypothetical: M = 5.30, SD = 1.12; Real: M = 5.12, SD = 1.04, d = .17).

Table 14.

ANOVA Results: Effects of the World Type.

Response variables	*F*	*df*	*P*
His fault	1, 1331	27.70	< .001
Others would agree his fault	1, 1331	27.23	< .001
Could have resisted	1, 1331	37.31	< .001
Wanted to offend	1, 1331	20.21	< .001
Reflects him as a person	1, 1331	16.75	< .001
Is likely to reoffend	1, 1331	46.06	< .001
Sentence	1, 1331	18.76	< .001

Perceptions of the real offender were analysed using a between-subjects MANOVA, with the Real Explanation [Unexplained, Biosocial, Neurobiological], the preceding Hypothetical Explanation [Unexplained, Biosocial, Neurobiological] and the Target Audience [General Population, Prison Population] as between-subjects factors. There was an interaction effect of the Hypothetical Explanation and the Real Explanation on the perceived extent to which the offender should be punished (F (4, 4100) = 2.51, p = .040, η_p^2 < .01). Participants recommended a significantly shorter sentence for the Neurobiological offender (M = 14.50, SD = 11.56) than for the Unexplained offender (M = 16.62, SD = 12.34) only if the preceding Hypothetical Explanation was Neurobiological (p = .005; Figure 6). Similarly, participants recommended a significantly shorter sentence for the Biosocial offender (M = 15.20, SD = 11.80) than the Unexplained offender (M = 17.69, SD = 12.26) only if the preceding Hypothetical Explanation was Biosocial (p = .010; Figure 7). Finally, participants recommended a significantly shorter sentence for the Neurobiological offender (M =

14.10, *SD* = 11.30) than for the Biosocial offender (*M* = 16.11, *SD* = 11.70) only if the preceding Hypothetical Explanation was Unexplained (*p* = .011).

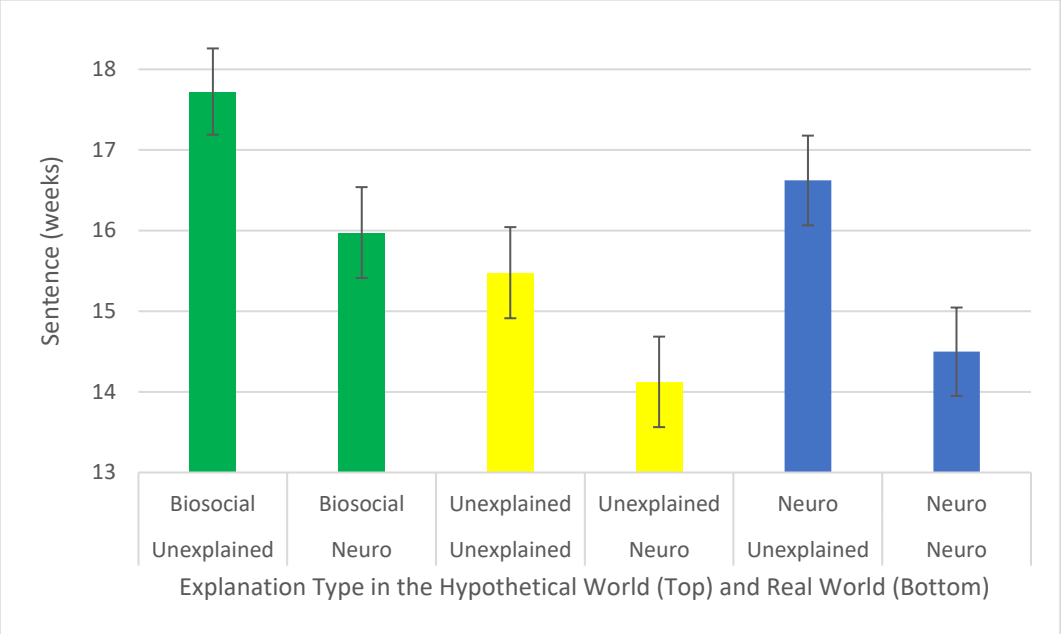


Figure 6. The effect of the Explanation Type (Neuro vs. Unexplained in the Real World) on the sentence recommended for the real offender. Error bars represent the standard error.

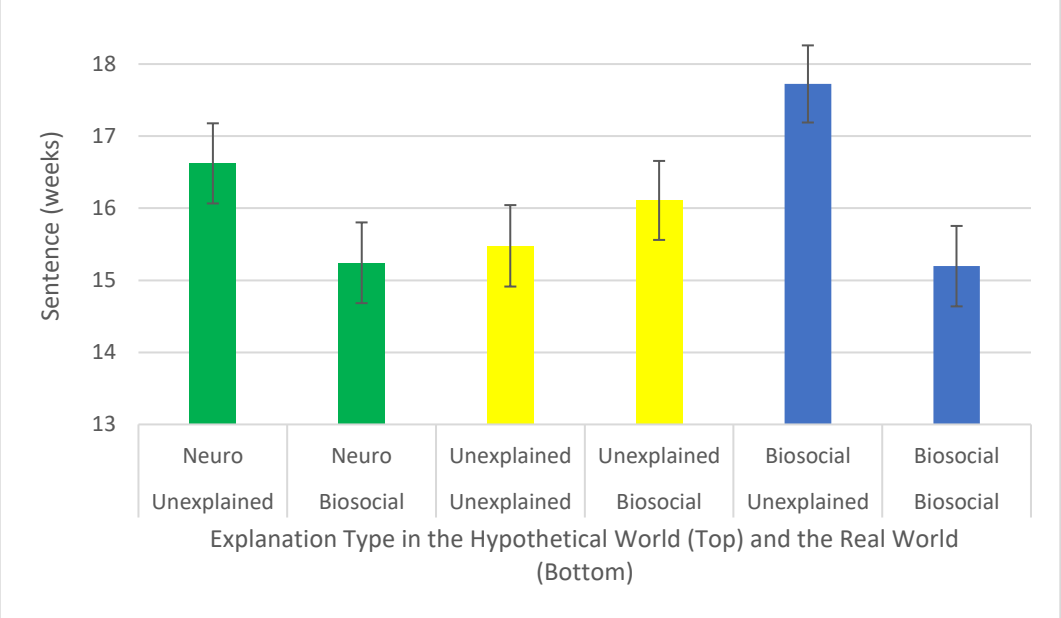


Figure 7. The effect of the Explanation Type (Biosocial vs. Unexplained in the Real World) on the sentence recommended for the real offender. Error bars represent the standard error.

The follow-up survey

There was no significant effect of the Hypothetical Explanation, the Real Explanation or the Target Audience on the probability of participants agreeing to receive the follow-up survey or completing the follow-up survey ($p > .05$). Therefore, the random allocation of participants to conditions was preserved across time. After the follow-up survey was closed, a MANOVA was conducted on the two follow-up variables. There was one interaction effect of the Hypothetical Explanation and the Real Explanation on punitiveness (F (4, 1094) = 2.73, p = .028, η_p^2 = .01). Participants were significantly less punitive three months after reading the Biosocial Explanation if the preceding Hypothetical account had also been Biosocial (M = 3.31, SD = .82) rather than Unexplained (M = 3.67, SD = .86; $p < .001$; Figure 8).

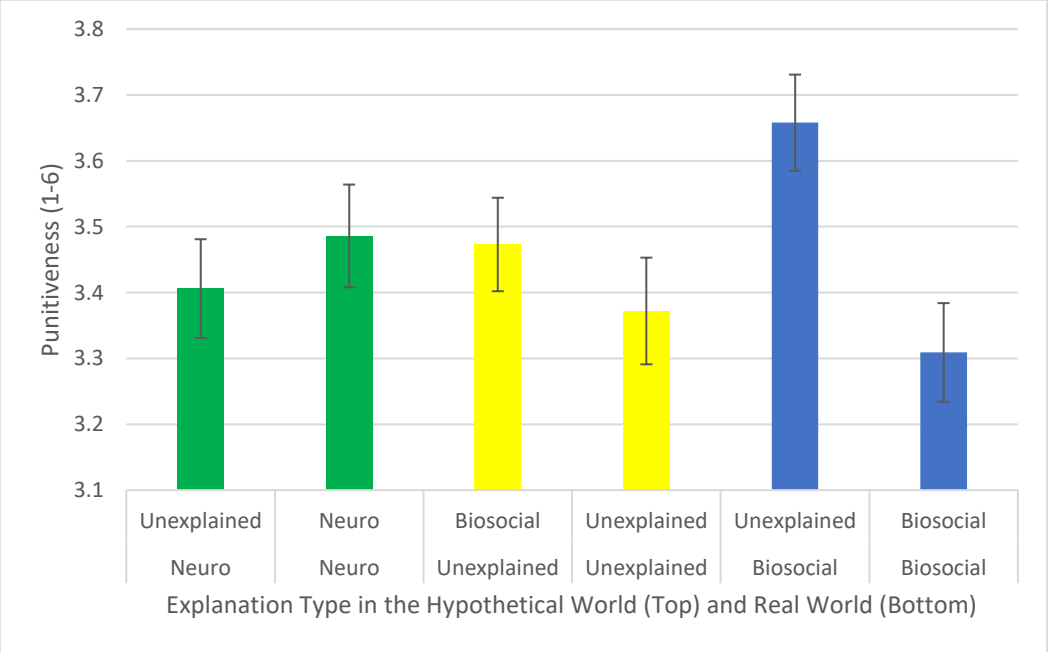


Figure 8. The effect of the Explanation Type (Same in the Hypothetical and Real Worlds vs. Different) on punitiveness three months later. Error bars represent the standard error.

Summary

The results provided some support for both hypotheses. Participants were more punitive towards the real offender than the hypothetical offender on four items (H3), but not the other two. Participants were less punitive towards the real caused offender after first reading the same causal explanation in the hypothetical world (as opposed to a different explanation or no explanation; H4). The real caused offender refers to the offender whose real-life criminal behaviour was described as the causal effect of the brain or genes and the environment.

Discussion

Perceptions of the offender were influenced not only by the type of explanation but also by whether that explanation was a real or hypothetical explanation of real or hypothetical criminal behaviour (H3). When the explanation was applied to the real world, participants were more likely to: attribute fault to the offender, believe the decision to offend reflected the offender as a person, punish the offender and (in the unexplained condition) believe the offender wanted to offend. However, participants also perceived the real offender as less capable of deciding to resist temptation and as less likely to reoffend.

These findings generate an interesting paradox for blame: the real offender was attributed a less moral self, which typically increases blame (Newman et al., 2014, Study 4), yet also less self-control, which typically decreases blame (Pizarro et al., 2003). In fact, the real offender was blamed and punished more; replicating the findings of Roskies and Nichols (2008). Perhaps when the science was applied to the real world, participants shifted their criteria for blame towards those that would minimise the mitigating force of the science; for example, by judging the morality of the Deep Self

rather than self-control. This drive to justify blame may be motivated by the applicability of real science (Shepherd, 2012), the emotions stirred by real crime (e.g., Nichols and Knobe, 2007), and the stated fact that the judgements will be publicised – all unique features of the real-world condition.

Were there facilitatory effects of the hypothetical world?

One unique feature of the museum study was the sequential presentation of the hypothetical and real worlds. The effect of the real explanation on the recommended sentence varied with the nature of the preceding hypothetical explanation (H4). Participants recommended a shorter sentence for the caused offender (than the unexplained offender) only if the same causal explanation had previously been presented in the hypothetical world. The caused offender refers to the offender whose criminal behaviour was described as the causal effect of the brain or genes and the environment. Most impressively, this effect was replicated three months later and in respect to offenders as an entire group: participants reported less punitive attitudes towards offenders in general when, three months earlier, the real biosocial account had been preceded by a biosocial (rather than unexplained) hypothetical account.

The framing of the second survey as a follow-up to the museum study may have reminded participants of the offender presented in the original vignette. Therefore, participants may have been more likely to envisage an opportunistic thief (who was also a university graduate and museum visitor) as the target of their punitiveness rather than the violent stereotype of offenders. Nevertheless, the duration of the effect is surprising. The finding suggests that people are more receptive to the mitigating implications of science that is first presented in a less applicable and less emotive context; in other words, a context that is less threatening to the function of blame in real life.

The facilitatory effect of the hypothetical world may result from participants making less punitive judgements of the hypothetical offender, which then transfer to judgements of the real offender. Such transference might be expected given the drive to reduce cognitive dissonance between the judgements one makes in the hypothetical and real contexts (Festinger, 1957). Note, however, the current study cannot rule out the alternative possibility that the effects simply result from greater exposure to the science. Hence future research is required to compare the effects of prior exposure to the science in an equally threatening and emotive, real-world context.

Nevertheless, a key policy recommendation remains for expert witnesses and other science communicators: emotional resistance to rational science may be somewhat overcome by first presenting the science in a hypothetical context, void of the emotions invoked by its applicability to real crime. Future research might attempt to amplify this facilitatory effect by asking groups of participants to discuss the hypothetically caused world before drawing the parallel with planet Earth, akin to the procedure used in Alien Nation (Hillman and Martin, 2002; Hodson et al., 2009). By basing the intervention on discussion and not simply text, the facilitatory effects of sequential presentation may be enhanced.

Finally, there was an unexpected effect of presenting the unexplained account in the hypothetical world. In the real world, participants were less likely to punish the neurobiological offender than the biosocial offender (H1) when the preceding hypothetical account was unexplained. It was predicted that the neurobiological explanation had greater mitigating potential than the biosocial explanation. It is simply easier to separate the brain from the Deep Self and consequently, believe the neurobiological cause of offending misrepresents a different and more moral Deep Self. Perhaps this effect was only observed after first presenting the unexplained account

because that account highlighted the difference between a world without causal explanations and a world with causal explanations; and by extension, a world with incompatibilist free will and a world without incompatibilist free will.

iii. Conclusion

Greene and Cohen (2004) hypothesised that the communication of neurocriminology would reduce the pursuit of positively retributive punishment by challenging the notion that criminal actions are ultimately free. In order to test this theory, the third empirical study of this thesis assessed whether people might blame and punish offenders less after learning that the brain *causes* criminal behaviour. In this respect, the museum study communicated the philosophical assumptions of neurocriminology rather than correlational findings of the relationship between the brain and criminal behaviour.

In this context, the presentation of neuroscience continued to have a small mitigating effect, including on the recommended sentence; moreover, this effect could be amplified by first presenting the science in a hypothetical context. Interestingly, the mechanisms responsible for this effect differed from the incompatibilist reasoning theorised by Greene and Cohen (2004). The neurobiological offender was punished less because he was perceived to have a more moral disposition and less self-control. Neither of these inferences that the offender has a more moral disposition or less self-control are scientifically valid inferences to make from the neurocriminological idea presented. While moral character has no legal relevance to retributive sentencing (e.g., Fletcher, 1998), self-control does (e.g., Morse, 2004b). Therefore, while the museum study documented a 'gap between what the law officially cares about and what people really care about' (Greene and Cohen, p.1776), the gap was slight and reflected lay concern for moral character rather than incompatibilist free will.

Chapter 9: The impact of neurocriminological ideas on attributions of dangerousness

The retributive justification for punishment is not the only grounds for punishment. There are also consequentialist reasons to punish offenders; those are reasons that concern practical benefits for the future of the offender and society (Carlsmith and Darley, 2008). It is therefore possible that after learning about neurocriminology, people shift from a retributive to a consequentialist rationale for punishment (Greene and Cohen, 2004). More generally, Gill and Getty (2016, p.774) theorise that exposure to explanations of offending may develop people into 'compassionate humanity haters'. Compassionate humanity haters are people who feel compassion for individual offenders, yet disdain for the vicious cycles of criminogenic abuse that create such offenders. In response, the compassionate humanity hater might seek to incapacitate and/or rehabilitate the abusers (i.e., offenders), thereby enacting consequentialist justifications for punishment.

This chapter begins by considering exactly how neurocriminology might lead people to infer offenders are more likely to reoffend and whether this inference would lead people to seek sentences that are more or less punitive than the status quo. Given the absence of research regarding these questions in the criminal context, this chapter begins by considering the effects of presenting neuroscientific explanations of mental illness on the perceived dangerousness of the mentally ill. Subsequently, the focus is shifted to the more criminally relevant context of presenting neuroscience in mock court scenarios, as well as an important theoretical difference between the clinical and criminal contexts. Finally, the chapter discusses relevant findings from the theatre, online and museum studies, which all addressed the question of whether people consider offenders to be more dangerous after learning about neurocriminology.

i. How might neuroscience aggravate sentences?

Even if people did become more consequentialist about punishment after learning about neuroscience, it is unclear whether this consequentialist approach would be less or more punitive than the retributive approach (e.g., Aharoni et al., 2007). Greene and Cohen (2004, p.1784) predict the net effect of communicating neuroscience will be less punitive – 'one day, the law may treat all convicted criminals…humanely'. In contrast, Snead (2007, p.274) predicts the net effect of neurocriminology will be aggravating: 'The criminal regime desired by cognitive neuroscientists would, tragically and ironically, prove far harsher and less humane'. Theorists such as Snead adopt a 'script equating biological theorising to aversive governmental interventions', such as the preventative detention of people merely at risk of offending (Wright and Cullen, 2012, p.238). The script is based on the historical precedent of neuroscience being abused to justify eugenics (Rafter, 2008).

In modern forms of science communication, neurocriminology would never be accompanied by the racist remarks of Lombroso (2006, p.348); who considered offenders to be 'members not of our species but the species of bloodthirsty beasts' (Lombroso, 2006, p.348). Yet still, it is feared that neurocriminology could implicitly increase animalistic perceptions of offenders by withdrawing attention from their individual conscious states at the time of the crime. The novel findings of Capestany and Harris (2014) support this theory: participants recruited brain regions implicated in logical reasoning less when judging the moral responsibility of an impulsive offender described as having a gene, rather than a personality, associated with impulsivity. While this finding was evident only for the less violent of two offences, the study provides some indication that genetic testimony might disable the logical reasoning required to

assess *mens rea*. Perhaps the neurobiological offender is considered less capable of rational choice as a result of dehumanisation.

In turn, the neurobiological offender may be deemed more likely to reoffend, and therefore, in greater need of incapacitation to prevent reoffending. This effect could result from the perception of the neurobiological offender as having less conscious awareness, less conscious control of the offending behaviour, less ultimate control of the offending behaviour or a less moral Deep Self. In other words, the same perceptions that drive the retributively mitigating influence of neurocriminology could also drive its aggravating effects from a consequentialist perspective. As a result, the neurobiological offender may be considered less willing, or less able to enact a willingness, to change. Dar-Nimrod and Heine (2011) review a body of evidence that people perceive genetically caused behaviours to be unchangeable. Similarly, the traits that people attribute to the brain are also the traits that people consider to be the best indicators of moral character (Fernandez-Duque and Schwartz, 2016); suggesting people consider the brain to be a stable influence on behaviour. If the brain is also believed to be unchangeable, rehabilitation may appear futile and incapacitation the only viable option. Hence Rose (2000) expects the aggravating interpretation of the neurobiological offender as unchangeable to overcome any mitigating effects.

Gill and Cerce (2017, p.13) provide evidence that the aggravating interpretation of neurocriminology arises because people believe neurobiological causation undermines conscious control. The participants attributed less capacity for rational choice to the offender with ‘brain damage’ and as a result, reported less desire to inflict ‘strategic punishments’ on this offender. In contrast, descriptions of child abuse exerted no effect on attributions of rational choice and consequently, also no effect on the desire to inflict ‘strategic punishments’. In this study, participants could interpret ‘strategic

punishments' to mean specific deterrence or rehabilitation; in other words, punishments that induce fear of the consequences of offending or address the underlying causes of the offending behaviour (p.15). Therefore, Gill and Cerce conclude that neuroscientific testimony undermines the use of punishment to achieve behaviour change; in turn, the use of punishment to incapacitate offenders may increase.

The perceived inefficacy of rehabilitation may reflect a broader tendency to make fatalistic inferences from evidence of causation; in other words, to infer that offenders are predestined to offend from the evidence that their prior criminal behaviour was caused (Miles, 2013), especially when caused by biology (Dar-Nimrod and Heine, 2011). Causal explanations of behaviour challenge the existence of *uncaused* self-control (i.e., incompatibilist free will), yet may be misperceived to challenge the existence of (even caused) self-control. Without sufficient self-control, the neurobiological offender might appear immune to their own cognitive attempts to change or the environmental interventions of others.

Three approaches have been distinguished in the study of neurocriminology and positive retributivism. The empirical research conducted for this thesis measured perceptions of dangerousness using the same three approaches: the mock court approach, the psychological approach and the approach of experimental philosophy. In previous research, dangerousness has been measured only in mock court experiments. Since there are only a small number of such studies, I first discuss a larger body of research into the effects of communicating neuroscience on attitudes towards mental abnormalities more generally; that is the clinical approach. Subsequently, I consider how the two approaches differ: in the criminal context, attributions of dangerousness may be affected by the concurrent drive to blame (even neurobiological) offenders.

ii. The clinical approach

The stigmatising effects of biology have been documented outside the criminal context, with mixed findings (e.g., Kim and LoSavio, 2009; but see Wright et al., 2012). For example, Walker and Read (2002) found that participants considered schizophrenic patients to be more dangerous and less predictable after reading a genetic explanation of schizophrenia – a finding that has since been replicated with other outcome measures, such as fear of the patient and desire for social distance (Read et al., 2006). In fact, the most consistently reported finding among clinical studies is that genetic explanations increase the desire for social distance from schizophrenics (Angermeyer, Holzinger and Carta, 2011). Biological explanations, which entail invisible and unfamiliar mechanisms to the lay person, resonate with prior 'public images of mental disorder [as]…dominated by…unpredictability and dangerousness' (p.367). Since one either has a genetic variant or not, genetic explanations may also support perceptions of schizophrenics as categorically different from mentally healthy individuals (Dar-Nimrod and Heine, 2011). One might predict similar effects for perceptions of offenders.

However, the relevance of genetics to a thesis on neuroscience might be questioned: while people consider genetic dispositions to be fixed (Dar-Nimrod and Heine, 2011), perhaps the brain is perceived as more changeable. Suggesting otherwise, Deacon and Baird (2009) found a neurobiological explanation of depression reduced the perceived scope for depression to diminish with time or therapy. In fact, Dietrich and Matschinger (2006) found that neurobiological attributions for depression and schizophrenia predict perceptions of dangerousness and unpredictability more strongly than genetic attributions. Therefore, despite the prior emphasis of research on genetics, neuroscience might similarly perpetuate attributions of dangerousness.

iii. The mock court approach

The mock court approach has been used to study lay attributions of dangerousness to offenders suffering deficits in the capacity for rational choice of differing severity. I begin by considering the deficit of principal concern to this thesis – a deficit in impulse control – before discussing the more sensitive measures of sentence aggravation used by other studies. Unlike categorical deficits to the capacity for rational choice (e.g., hallucinations), lay people typically have experience of milder deficits to the capacity for rational choice, such as impulsivity. It is most intuitive to conceive of biological differences as explanations of *categorical* deficits to rational choice (Dar-Nimrod and Heine, 2011). Hence one might question whether people also endorse biological explanations of impulsivity – a trait that more clearly varies on a continuum between mentally healthy and disordered people.

Scurich and Appelbaum (2016) found that presenting genetic evidence of impulsivity increased attributions of dangerousness to the offender in two of three studies – the two that described less serious offences. In the third, most serious case, genetic testimony actually reduced attributions of dangerousness. The latter result is consistent with the findings of Saks et al. (2014): participants perceived a schizophrenic offender to be less dangerous after reading neuroscientific testimony of the diagnosis. Similarly, judges deem a diagnosis of psychopathy to be less aggravating when it is supported by neurogenetic evidence, even after the prosecution has explicitly used that diagnosis to describe the offender as dangerous (Aspinwall et al., 2012). Scurich and Appelbaum speculate this mitigating effect arises when the offender is already considered dangerous or unpredictable by virtue of the serious offence or serious mental disorder. Perhaps in these cases, knowledge of a cause acts to reassure the lay person: when the cause is known, the behaviour is understood, and so rehabilitation can at least

be attempted. Therefore, neurocriminology could even mitigate punitiveness for consequentialist reasons. In real court cases, this could explain the disproportional use of neuroscience by the defence: the prosecution rarely seeks neuroscientific evidence to present the offender as dangerous, despite the potential to collect such evidence (Denno, 2015). Perhaps the prosecutor recognises the consequentialist argument would be unpersuasive.

Indeed, several studies report null effects. Greene and Cahill (2012) found no effect of neuroscientific testimony on attributions of dangerousness to psychotic offenders. Appelbaum, Scurich and Raad (2015) found an effect in only one of three studies, with genetic testimony of impulsivity increasing attributions of dangerousness to the bipolar offender. Since one would expect a bipolar offender to already be conceived as unpredictable, this finding challenges the proposed mitigating effect of neuroscience for such offenders. Appelbaum and Scurich (2014) measured the effects of describing a genetic disposition to act impulsively without any clinical diagnosis. When the disposition was used to explain second-degree (i.e., intentional but not premeditated) murder, participants perceived the offender as more dangerous, with no consequences for the recommended sentence. In sum, there is clearly no consistent evidence of the aggravating interpretation of neuroscience in the criminal context; in fact, the science sometimes mitigates for consequentialist reasons.

It is possible, however, that these studies are not adequately equipped to capture more implicit manifestations of the consequentialist implications of neurocriminology. Although Greene and Cohen (2004) do not specify the exact form of consequentialist punishment that neurocriminology might promote, the most plausible consequentialist alternative to prison would be secure accommodation; secure accommodation would incapacitate offenders to protect the public from future offences and during detention,

would aim to rehabilitate, rather than punish, offenders (Raine, 2013). Without retributivism, however, there would be no objection to detaining offenders for as long as they continue to pose a threat to society. By rejecting retributivism, one rejects the limiting principle of proportionality: if the severity of the sentence must no longer reflect the seriousness of the offence, there is no retributive limit on the length of incapacitation (Morris, 1974).

There is, however, no logical reason to lose the limiting principle of proportionality in response to neurocriminology. The loss would be an unfortunate byproduct of the decision to abandon both components of retributivism simply because one component – the assessment of blame – had been subject to scientific ridicule. The question therefore arises of whether, in response to neurocriminology, lay people relinquish not only the retributive drive for punishment but also the retributive cap on punishment. Thus far, it is clear that lay people do not consistently attribute greater dangerousness or longer prison sentences to offenders with brain abnormalities. However, it remains possible that people still wish to detain offenders in secure accommodation for longer than would be retributively appropriate for their offence type. None of the discussed studies gave participants the opportunity to detain offenders in secure accommodation for so long as the offender is considered dangerous – a more implicit manifestation of consequentialist punitiveness, in the form of incapacitation. Hence Fuss, Dressing and Briken (2015) replicated the seminal study of Aspinwall et al. (2012) in Germany, adding a question about the endorsement of indefinite detention.

Fuss et al. (2015) had reason to expect German judges to be especially sensitive to neurogenetic evidence of psychopathy. Before the 1980s, the only mental disorder that was considered relevant to legal responsibility was organic disorder; in other words, a disorder that changed biological structures and not simply biological

functioning, such as brain activity. In turn, one must evidence diminished legal responsibility (and dangerousness) in order to indefinitely detain an offender under the label of 'involuntary commitment'. With this cultural precedent still in mind, the German judge might assign particular weight to biological evidence. In contrast to the US (Aspinwall et al., 2012), there was no effect of neurogenetic evidence on the sentences sought for psychopaths in Germany (Fuss et al., 2015). However, presentation of neurogenetic evidence by the prosecution quadrupled the number of judges who decided the psychopath should be indefinitely detained in a psychiatric hospital after imprisonment. One might not initially consider this practice of involuntary commitment punitive: perhaps the judges intended to grant the offender access to rehabilitation services that would be unavailable in prison. However, involuntary commitment does tend to lengthen the overall period of detention (Kröber and Lau, 2000). Moreover, judges should expect this outcome, especially in this study where the vignette described the offender as resistant to rehabilitative interventions. Therefore, Fuss et al. do document an aggravating effect of neurocriminology.

Interestingly, attributions of free will and moral responsibility did not differ between conditions (Aspinwall et al., 2012; Fuss et al., 2015). However, unlike in the US (Aspinwall et al.), the neurogenetic evidence did reduce the degree of legal responsibility attributed to the psychopath in Germany (Fuss et al.). Perhaps German judges were more likely to believe the psychopath lacked the (organic) capacity for rational choice and therefore, required involuntary commitment. Neurogenetic evidence might be conceived to be a means of distinguishing a deficit in the capacity from a deficit in the motivation to exercise rational choice. The authors did not measure whether neurogenetic evidence (subsequently) increased attributions of dangerousness. Interestingly, however, neurogenetic evidence had no effect on the use of 'preventative

detention' – a different form of indefinite detention that German judges can administer to dangerous, yet completely rational, offenders. This suggests that neurogenetic evidence *may* prompt people to seek indefinite detention as a result of altering attributions of rational choice, rather than some other basis for judging dangerousness.

The contrasting findings of Aspinwall et al. (2012) and Fuss et al. (2015) suggest the aggravating interpretation of neurocriminology may not be detected in simple sentencing recommendations. Instead, researchers must give participants an explicit opportunity to detain the offender on purely consequentialist grounds. This option was included in the online study. Hence it was possible to test whether neurocriminology also encourages the indefinite detention of the impulsive (and not only the psychopathic) offender. There are reasons to expect it might not be: people may perceive the impulsive offender to be more rational and reformable, and less dangerous and categorically different, than the psychopathic offender. In turn, the impulsive offender may be more immune to aggravating interpretations of the science.

iv. The criminal context

The key difference between the clinical and mock court approaches is that moral concerns are far more likely to arise in the mock court approach. The effects of neuroscience in the courtroom could be distorted by the retributive drive to condemn immoral acts and blame immoral actors. There is some evidence that blame is simply a less relevant concern for lay people judging mentally ill patients who have not committed crime. For example, blame and stigma for mental disorder are uncorrelated (Angermeyer et al., 2011). There may be no baseline desire to blame the moral, yet mentally disordered, patient and so no blame for biological evidence to temper. Indeed, genetic explanations of schizophrenia do not affect the perceived blameworthiness of a

schizophrenic for ‘having developed’ the disorder (Bennett, Thirlaway and Murray, 2008). In this clinical context, it is possible to infer dangerousness from biological causation without fearing that the patient will be excused for offending; because the patient has not offended. Consequently, the aggravating interpretation of neuroscience can proceed without a crime being excused.

This reasoning might explain the clinical findings of Mehta and Farina (1997): even though ‘mentally disturbed’ people were blamed less (than mentally healthy people) for making mistakes on a behavioural task, they were actually punished more for making these same mistakes. One explanation is that the participants believed greater punishment was necessary for the mentally disturbed people to learn from their mistakes – a consequentialist argument that a greater attempt at rehabilitation or deterrence is required to prevent mentally disturbed people from making mistakes again. Perhaps consequentialist reasoning dominates this punishment decision because, in a clinical context, people do not seek to blame mentally disturbed patients:

> Essentialist thinking [promoted by biological explanations] may be more closely tied to [or even function to assert]…the categorical otherness of the affected individual than to [assert]…moral condemnation (Haslam, 2011, p.821).

Consequently, the aggravating interpretation of neuroscience can proceed without moral concerns.

In contrast, in the criminal context, moral concerns may arise when the aggravating and mitigating interpretations depend on each other: if one concludes that the neurobiological offender is more dangerous because he is less rational, it is logical to infer that this offender is also less blameworthy. This dependence derives from the finding that participants attribute less blame and less ‘strategic punishments’ to the offender with ‘brain damage’ for the same reason: because this offender appears to lack the capacity for rational choice (Gill and Cerce, 2017, p.13). Greene and Cahill (2012)

provide evidence that neuroscientific testimony only has a mitigating effect when the offender is described as dangerous: one explanation is that the justification for the dangerous label – a diminished capacity for rational choice – suggests the offender is less deserving of punishment. Lay people have a stubborn drive to blame offenders (Crockett et al., 2014; Fehr and Gächter, 2002; Nadelhoffer et al., 2013b). Therefore, the logical dependence of neuroscientific aggravation on mitigation might deter people from making the aggravating interpretation; in other words, deter people from inferring dangerousness from neurobiological causation.

Of course, people can maintain instinctive attitudes that defy their own logic (Haidt, Bjorklund and Murphy, 2000). Therefore, with sufficient punitive instinct, people might break the logic that ties the aggravating and mitigating interpretations together. In other words, people might expect the neurobiological offender to reoffend without discounting blame for the prior criminal behaviour. There is already evidence that people perceive the brain to be both controllable and stable (Fernandez-Duque and Schwartz, 2016). In turn, the neurobiological cause of offending could be considered both blameworthy and stable, thereby supporting the punitive drive to blame and incapacitate. Hence it is plausible that people seek some means of justifying incapacitation without excusing the act.

Note the logical dependency also operates in the reverse direction. People may interpret their basis for excusing the neurobiological offender (e.g., impaired rationality) as a cue to dangerousness. If lay people do not wish to consider the neurobiological offender dangerous, they may not wish to excuse this offender either. Indeed, Greene and Cahill (2012) found the mitigating effect of neuroscientific testimony disappeared when the offender was no longer described as dangerous. Again, this logical dependency could be broken; for example, by changing the basis for

mitigation from impaired rationality to mere causation. Perhaps therefore, with sufficient drive for leniency, people might excuse the neurobiological offender without also inferring dangerousness. This outcome would require the brain to be perceived as uncontrollable yet changeable on the basis that genetic and environmental influences on the brain change across time. In turn, neurocriminology would reduce the punitive drive both to blame and to incapacitate. This was precisely the prediction made by the theatre study in respect to young offenders.

v. The theatre study

In the theatre study, the communicated neuroscience broke the dependency between the mitigating and aggravating interpretations: participants learnt that the adolescent brain lacks the capacity to exercise rational control over strong impulses, yet also develops that capacity naturally with age. This was the first criminological study to present the brain as changing across time; hence, unlike prior research, the study explicitly challenged the tendency for people to believe that biology is fixed (Dar-Nimrod and Heine, 2011; Miles, 2013). Consequently, it was predicted that the participants would perceive adolescent impulsivity as uncontrollable yet changeable, and subsequently, perceive young offenders as likely to 'grow out' of their offending behaviour.

Hypothesis

The theatre study measured the effects of watching a play about adolescent brain development on attitudes towards a hypothetical young or adult offender who committed a serious assault three times in three years. In chapter six, the findings were discussed that after watching the play, participants recommended a higher age of criminal responsibility (H1) and attributed less moral responsibility to a hypothetical young, but not adult, offender (H2). One further hypothesis was proposed:

H3: Participants would consider the young offender less likely to reoffend (than the adult offender) after watching the play.

Results

Estimated probabilities of reoffending were analysed using an independent-groups ANOVA, with Age of the offender [Young, Adult] and Play [Before, After] as between-subjects factors. Participants perceived the adult offender as more likely to reoffend (M = 75.54, SD = 23.22) than the young offender (M = 69.09, SD = 26.63), F (1, 713) = 10.80, p = .001, ηp^2 = .02. There was also a significant interaction effect of the Age and Play, F (1, 713) = 5.54, p = .019, ηp^2 = .008. Before the play, there was no significant difference between the estimated probabilities of the young (M = 71.10, SD = 25.90) and adult (M = 72.86, SD = 24.12) offenders reoffending, t (341) = .64, p = .523. However, after the play, participants perceived the young offender as significantly less likely to reoffend (M = 67.01, SD = 27.28) than the adult offender (M = 77.65, SD = 22.32), t (364.31) = 4.14, p < .001, d = .43 (equality of variance not assumed, p = .001; Figure 9).

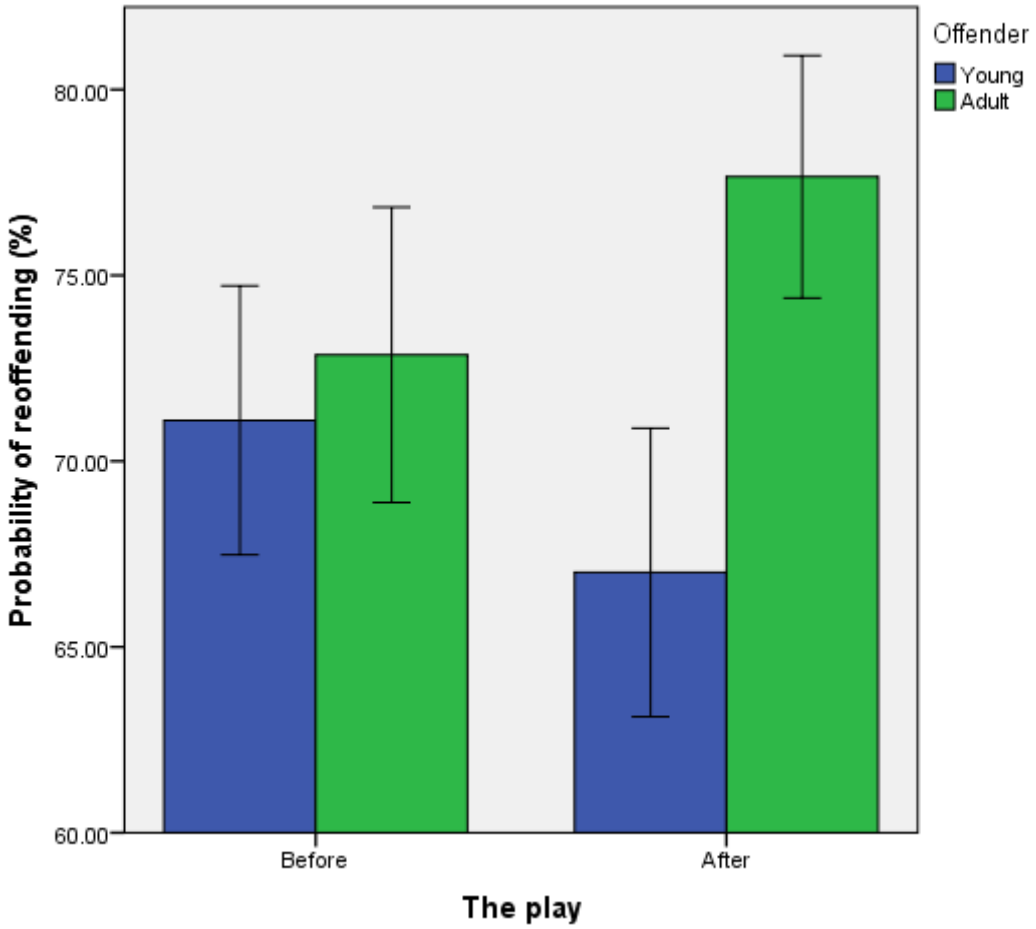


Figure 9. The effect of exposure to neuroscience on the perceived probability of hypothetical young and adult offenders reoffending. Error bars represent the standard error.

Discussion

Participants believed the young offender was less likely to reoffend than the adult offender after, but not before, watching the play (H3). This is a striking result, given the proposal that neuroscience might actually amplify the perceived dangerousness of offenders (e.g., Snead, 2007). This prediction was reversed in the current findings most likely because the play depicted the brain as changing with age, thereby challenging the default perception of biology as unchanging (Dar-Nimrod and Heine, 2011). Therefore, while this study did not document spontaneous 'public excitement about brain plasticity' (Pitts-Taylor, 2010, p.636), it did document receptivity to the idea.

By attributing adolescent behaviour to the developing brain, the play induced belief in the idea that young offenders naturally desist from offending with age, even without punishment or rehabilitative interventions – an idea that is well supported by empirical evidence of the age-crime curve (Hirschi and Gottfredson, 1983; Sampson and Laub, 2003). In fact, desistance is promoted by diverting young offenders away from the formal disciplinary structures of the CJS (McAra and McVie, 2007). For lay people to support diversionary policies in youth justice, it may first be necessary to increase their belief in the existence of the age-crime curve; the current study suggests this play is one viable means of increasing such belief.

Exposure to neuroscience eroded a consequentialist justification for punishment, suggesting that, in the youth justice context, neurocriminology can exert a consequentialist, mitigating influence. As reported in chapter six, the play also reduced the degree of moral responsibility attributed to the young offender. In combination, these findings suggest that lay people are willing to divorce their attributions of stability and controllability: participants were receptive to the message that adolescents are influenced by uncontrollable, yet also unstable, brain mechanisms and therefore, that young offenders are both less blameworthy and less dangerous. In this respect, participants were able to break the hypothesised dependency between the mitigating and aggravating interpretations of neuroscience: attributions of moral responsibility were reduced without any concurrent increase in perceptions of dangerousness. The consequentialist aggravating interpretation of neuroscience is not a *necessary* side-effect of the retributive mitigating interpretation. In science communications that emphasise brain plasticity, such as this play, lay people can reach scientifically valid, mitigating conclusions.

vi. The online study

The theatre study found that participants believed the young offender was less likely to reoffend as his capacity for rational choice developed naturally with age. This suggests that more rational offenders are believed to be less dangerous – a suggestion that is empirically supported by Gill and Getty (2016). Therefore, if neurocriminology does have the potential to aggravate sentences, that potential is most likely to be realised in two contexts: first, when the science is used to explain a deficit in rationality and second, when the irrationality of the offender is unstated, yet inferred from the mere presence of neurobiological causation (Nahmias et al., 2007; the museum study). The hypothesis is clearly that neurocriminology may strengthen the perception of the offender as irrational, and therefore, dangerous. In contrast, one would not expect aggravating effects to arise when the science is used to explain a lapse from an otherwise intact capacity for rational choice. Some people commit impulsive offences following a lapse in their normally adequate capacity for impulse control; in such cases, one can still explain their lapse in neurobiological terms.

The online study aimed to directly test this irrationality requirement for the aggravating interpretation of neurocriminology. In every condition, the offender committed the same violent offence on impulse. The offender was described, in neurobiological or cognitive terms, as having either a controlled disposition (the 'controlled actor') or an impulsive disposition (the 'impulsive actor'); so either an intact or impaired capacity for rational choice.

Hypotheses

In chapter six, the findings were discussed that participants attributed more free choice and fault, but not a longer sentence, to the controlled actor than the impulsive actor

(H1). There was no effect of communicating neuroscience on attributions of free choice, fault or punishment (H3), except neuroscience increased belief that the impulsive actor lacked the capacity (rather than the motivation) to exercise self-control (H2). Two further hypotheses were proposed, in respect to consequentialist reasoning:

H4: Participants would perceive the controlled actor to be less likely to reoffend and more capable of being reformed.

H5: For the controlled actor, neuroscientific testimony would exert no effect on attributions of dangerousness.

Given the very inconsistent findings from previous research, no equivalent hypothesis was made for the impulsive actor.

Results

Perceptions of the offender were analysed using a between-subjects ANOVA, with the Disposition of the actor [Impulsive, Controlled] and its Source [the Self, the Brain] as between-subjects factors. There were main effects of the Disposition on the perceived extent to which the offender: would be likely to reoffend without receiving intervention (F (1, 593) = 28.29, $p < .001$, $\eta_p^2 = .046$) and could be reformed (F (1, 593) = 11.57, $p = .001$, $\eta_p^2 = .019$). Participants were significantly less likely to expect the controlled actor to reoffend (Controlled: $M = 4.00$, $SD = 1.14$; Impulsive: $M = 4.49$, $SD = 1.12$) and significantly more likely to believe the controlled actor could be reformed (Controlled: $M = 4.48$, $SD = 1.02$; Impulsive: $M = 4.19$, $SD = 1.07$).

The control condition

Independent-groups t-tests were conducted to compare the control and experimental conditions. In comparison to the undescribed actor, participants believed the

neurobiologically impulsive actor was more likely to reoffend (t (305) = 3.73, p < .001; Exp: M = 4.59, SD = 1.09; Control: M = 4.10, SD = 1.21) and less likely to have the capacity to resist the impulse to offend (t (305) = 3.24, p = .001; Exp: M = 4.52, SD = 1.07; Control: M = 4.92, SD = 1.11).

Discussion

There was no effect of the neuroscientific testimony on the perceived likelihood of the offender reoffending or the perceived capacity for his behaviour to be changed (H5); this finding corroborates those previous studies that have reported null effects on attributed dangerousness (e.g., Greene and Cahill, 2012). Hence this study challenges the fear that lay people will consider the neurobiological offender unchangeable and consequently, that neurocriminology will aggravate sentences (e.g., Snead, 2007). Fuss et al. (2015) observed a more implicit aggravating effect, whereby neurogenetic evidence motivated the indefinite detention of psychopaths. The current participants received two opportunities to detain (predicted or actual) offenders on purely consequentialist grounds. Yet still, the neuroscientific testimony had no effect on the uptake of those opportunities for impulsive offenders.

Two factors might explain the discrepancy between the current null effects and the positive effects reported by Fuss et al. (2015). First, neuroscience may be more likely to increase the perceived dangerousness of offenders with more severe deficits in rationality (e.g., psychopathic offenders in comparison to impulsive offenders). Second, neurogenetic testimony may be more likely to motivate perceptions of dangerousness than neurobiological testimony. In respect to the second factor, future research is required to test whether people intuitively understand that environmental interventions

can change brain activity; for example, in order to overcome a neurobiological cause of offending and render the neurobiological offender no longer dangerous.

The described disposition of the offender did influence consequentialist judgements: participants tended to perceive the controlled actor as more capable of resisting the impulse to offend, less likely to reoffend and more capable of reform (H4). Participants also perceived the undescribed offender to be less likely to reoffend and more capable of reform than the neurobiologically impulsive actor. This finding suggests that by default, people expect a person who commits a violent offence on impulse to have a normal capacity to exercise self-control. While neuroscience was capable of changing that perception for the impulsive actor, this change did not affect his perceived dangerousness.

vii. The museum study

In the online study, participants made no inference of dangerousness from a neurobiological description of the rational (or the less rational) offender. This finding is consistent with the rationality requirement: neuroscience may motivate perceptions of dangerousness by bolstering a conception of the offender as lacking the capacity for rational choice. However, the study also failed to test the rationality requirement because the participants never made a neurobiological attribution for the offending of the rational actor described in neurobiological terms. In other words, the participants did not infer, from the testimony that his normal capacity for rational choice was caused by the brain, that a lapse in rationality was caused by the brain. Therefore, it remains possible that neuroscience could still amplify attributions of dangerousness to the rational offender.

In response, the museum study provided a better test of the rationality requirement by changing the target of neuroscientific explanation: the brain was described as causing 'every conscious decision to resist or give into temptation' rather than the dispositional tendency to resist or give in to temptation. 'Every conscious decision' was intended to include deviations from the (normally controlled or normally impulsive) disposition of the offender. The disposition of the offender was not stated but he was described as a university graduate returning from vacation with friends and in a relationship – characteristics that might trigger perceptions of a relatively rational actor. Similar to the museum study, the offence was also impulsive; specifically, an act of non-violent opportunistic theft (as opposed to violent assault). The offender in the museum study therefore shares characteristics with the controlled actor in the online study: both might be perceived as offending due to a lapse in self-control, yet, in the museum study, there is a stronger situational explanation of that lapse.

In the museum study, participants read a complete, causal and universally applicable account of self-control in neurobiological, biosocial or unexplained terms; participants then judged an offender in hypothetical and real-world contexts. Since the studies of experimental philosophy do not ask about dangerousness (Nichols, 2011), this was the first study to measure attributions of dangerousness following exposure to a complete causal argument.

Hypothesis

In chapter eight, the findings were discussed that in the hypothetical world, participants recommended a shorter sentence for the neurobiological offender than the biosocial offender, and in the real world, for the biosocial offender than the unexplained offender (H1). These effects were mediated by moral perceptions of the 'person' who offended

(H2). It was also discussed that participants were more punitive towards the real offender than the hypothetical offender on four items (H3), but not the other two. Participants were less punitive towards the real (caused) offender after reading the same causal explanation of his behaviour in the hypothetical world (as opposed to a different explanation or no explanation; H4). One further hypothesis was made:

H5: Participants would not perceive the neurobiological or biosocial offenders to be more dangerous than the unexplained offender. Consequently, there would be no aggravating effect of the science on the recommended sentence.

Results

There was a main effect of the Hypothetical Explanation on the perceived extent to which the hypothetical offender could be changed, F (2, 4180) = 5.68, p = .003, η_p^2 < .01. Participants were significantly more likely to believe the Biosocial offender could be changed (M = 4.83, SD = 1.20) than the Neurobiological offender (M = 4.68, SD = 1.23; p = .001). There were no effects of the Real Explanation, the World Type or any interaction term on the perceived probability of reoffending or the capacity for the offender to be changed (p > .05 in every case).

Discussion

No evidence was documented of the aggravating interpretation of neurocriminology (H3). Neither the neurobiological nor the biosocial explanations affected the perceived probability of reoffending, and just one effect on the perceived scope for reform was observed: in the hypothetical world, participants believed there was greater scope for the biosocial (than the neurobiological) offender to change. Perhaps the biosocial explanation rendered salient the environmental causes of offending, thereby providing a clear basis to behaviour change. Note this finding did not document an aggravating

effect of neurocriminology since the neurobiological offender was judged no differently from the unexplained offender.

This null effect may be particularly striking given the earlier reported finding that the neurobiological offender was attributed less self-control in the same study. In combination, the findings suggest that participants prioritised the opportunity to make a mitigating (over aggravating) inference from the perceived lower of self-control of the neurobiological offender. Therefore, in the pursuit of leniency, participants broke the logical dependency between the two interpretations: the neurobiological offender was attributed a shorter sentence, but not also more dangerousness, for lacking self-control. The science also did not exacerbate perceptions of the Deep Self as a means of justifying greater consequentialist punishment.

The question therefore arises of why there were no aggravating effects of neurocriminology. A likely explanation is that the situational nature of the crime – opportunistic theft – rendered the neuroscience resistant to consequentialist inferences:

> Contextual explanations…might fail to elicit much disdain for humanity [because]…such explanations point to one-off events that might be less emotionally evocative than…enduring maltreatment (Gill and Getty, 2016, p.787).

The current participants may have attributed the neurobiological cause of offending to the unusually tempting situation; as suggested by the improvement in moral perceptions of the offending person in the neurobiological condition. Having made a situational attribution, the participants should expect the cause of offending to expire as soon as offender leaves the criminogenic situation; for example, as soon as he matures, is no longer the last passenger to leave the plane or no longer has a girlfriend who wants to visit a luxury holiday resort that he cannot afford – all features of the vignette. In turn, the neurobiological cause of this particular offence no longer represents a dispositional

trait that could drive future offending. In fact, if participants attributed the criminogenic brain activity to the external situation, these participants should also expect brain activity to change with other environmental inputs; including prosocial environments that may rehabilitate offenders. In turn, neurobiological causation no longer appears to justify a consequentialist need to incapacitate the offender.

More generally, the prediction of neuroscientific aggravation may be based on a misplaced assumption; that being people perceive the neurobiological offender to have less incompatibilist free will, *and therefore*, to be more dangerous. The current null effects suggest that people do not believe offenders need to have some ultimate (i.e., incompatibilist) form of control over their behaviour in order to change. This represents a scientifically accurate point of view; a moral improvement in behaviour may follow an increase in self-control, resulting from a change in the environment to which the offender is exposed. In this respect, behaviour change necessitates only compatibilist (not incompatibilist) free will. It appears that the incompatibilist notion of free will may form the basis of neither perceived dangerousness nor blame (e.g., Nahmias et al., 2005), thereby desensitising the concepts to neuroscientific arguments.

If dangerousness is not inferred from an absence of incompatibilist free will, perhaps it is inferred from perceptions of an immoral Deep Self instead; those being perceptions that a person's most stable core traits are immoral (Sripada, 2010). As discussed in chapter eight, neuroscientific testimony improved moral perceptions of the Deep Self of the offender in the museum study. This improvement might explain why the participants did not conclude the neurobiological offender was more dangerous. By basing judgements on the Deep Self, one also avoids the cognitive dissonance generated by basing judgements on the assessment of incompatibilist free will or rational choice. When neurobiological causation is construed as evidence against incompatibilist free

will or rational choice, retributive and consequentialist concerns conflict: the neurobiological offender deserves less, yet needs more, punishment because he is less blameworthy yet more dangerous (Barth, 2007; Aspinwall et al., 2012). The resultant dissonance between the neuroscientific drive for mitigation and aggravation can be resolved by shifting the basis for attributing blame and dangerousness; from incompatibilist free will and rational choice towards the Deep Self. When the neurobiological explanation is construed as evidence that a more moral Deep Self has been obstructed by a situational cause of offending, retributive and consequentialist concerns no longer conflict: the neurobiological offender deserves, *and needs*, less punishment because he is less blameworthy *and* less dangerous. The retributive interpretation advances, and the consequentialist interpretation permits, a mitigating interpretation of neuroscience; the two punishment rationale align, and the cognitive dissonance is resolved.

In sum, neurocriminology may incentivise lay people to ground their recommended sentences in perceptions of the Deep Self and so to sentence offenders from this compatibilist (rather than an incompatibilist) perspective (Nahmias et al., 2005). When a criminogenic situation also appears to have obstructed a more moral Deep Self, people may not only blame the neurobiological offender less but also deem him safe. In fact, the sensitivity of the criminogenic brain activity to the external situation might evidence the capacity for changes in the external environment to reform the neurobiological offender.

viii. Conclusion

Theorists like Snead (2007) are convinced that neurocriminology will aggravate sentences when the science is used to document a stable *disposition* to offend.

However, this prediction is not supported by the current three studies. In the theatre study, neuroscience was used to describe an impulsive *disposition* that changed naturally with age; consequently, the science actually reduced attributions of dangerousness. In the online study, neuroscience was also used to describe an impulsive *disposition*, with no effects on attributions of dangerousness, the sentence or indefinite detention. In the museum study, participants were not inclined to make a dispositional interpretation of the criminogenic brain activity, or genes and the social environment; in spite of the fact that, arguably, the distal biosocial factors necessitated a dispositional explanation, and the source of the brain activity was ambiguous and so could be explained by disposition.

It could be argued that people would not make a dispositional interpretation of a criminogenic influence that affected an offender with other desirable traits; for example, the museum study described a university graduate as the offender. However, the theatre study described a perpetrator of triple assault and the online study described a violent attack that resulted in brain injury and was vastly disproportionate to the source of provocation; even for these abhorrent acts, participants did not make the aggravating interpretation of neurocriminology. Therefore, the present research questions the prevalence of the aggravating interpretation.

Of course, there may still be contexts that were not tested by the current three studies where the aggravating interpretation does arise. For example, neurocriminology might be used to distance an offender from the moral ingroup when the offender already appears fundamentally different from the ingroup along several dimensions; for example, because the offender has a severe mental illness (e.g., Fuss et al., 2015), serious prior convictions and other cues to outgroup status. In these cases, people may

be sufficiently motivated to interpret a scientific explanation of offending as a confirmation of the immoral Deep Self of the offender.

The risk of this aggravating interpretation may be greater for biosocial (than neurobiological) explanations, given that people are more likely to make dispositional interpretations of biosocial factors (the museum study). In turn, people may dehumanise the outgroup biological offender; in other words, withhold from their perception of this offender the seemingly most human traits, such as the capacity to think and reason (Leyens et al., 2000). Perhaps evidencing this effect, Greene and Cahill (2012) found that neuroscientific testimony reduced attributions of legal responsibility only when the offender had already been described as dangerous; this label may have acted as a cue to his outgroup status. Since neurobiological causation per se is not evidence of dangerousness, the probative value of the science may exceed its prejudicial potential for the most stigmatised offenders.

Chapter 10: The impact of neurocriminological ideas on moral behaviour

Thus far this thesis has assessed the impact of neurocriminology on the attitudes people have towards offenders. In contrast, this chapter considers the impact of the science on the attitudes and behaviour of the offenders themselves. If neurocriminology is integrated into the CJS, this transition will take place within the view of offenders. Consequently, offenders would learn more about the neurobiological predictors of their own antisocial behaviour and mental health conditions not only during the sentencing process and parole hearings but also in prison and through the probation service. In the latter institutions, neurocriminology may be used to tailor rehabilitation services to the neurobiological 'needs' of the particular offender (Glenn and Raine, 2014). Hence, with advancements in science and technology, offenders may develop an understanding of the otherwise hidden contributors to their criminal behaviour. The present chapter addresses how such an understanding might impact their moral behaviour.

The chapter begins by presenting the theoretical backdrop to the Facebook study. The Facebook study considers whether giving people neurofeedback about their degree of psychopathic traits might actually amplify the degree to which they exhibit psychopathic behaviours. The chapter presents four mechanisms by which neurofeedback could bring about psychopathic behaviour: by reducing belief in (compatibilist or incompatibilist) free will, by reducing belief that the world is a just place where people get what they deserve, by increasing belief that one's future is already mapped out, or simply by increasing the believability of the feedback about one's psychopathic traits. The findings of the Facebook study are then presented, before the focus of the chapter shifts slightly: even if the communication of neurocriminology does not increase immoral behaviour in the present, it might increase the salience or recall of memories of having behaved immorally in the past; the museum study

addresses this latter point. Finally, the findings of the Facebook study and the museum study are considered together, with respect to the practical implications of communicating neurocriminology to offenders.

i. The Facebook study

While researchers have begun to capture the qualitative responses of offenders to neurocriminology (Horstkötter, Berghmans and Ruiter, 2012; Horstkötter, Berghmans, Feron and De Wert, 2014), no experimental study has quantified the effects of learning how neuroscience might explain one's own moral behaviour. Hence the Facebook study was designed to measure how people respond to neuroscientific feedback about their own moral disposition. While the study was conducted with lay people for ethical and practical reasons, it provides the basis for predicting how offenders might respond to personalised neuroscience in the CJS; that is the use of neuroscience to explain the traits of a particular offender.

Moral alarm

The Facebook study focused upon a particular moral trait; that was the dispositional tendency to experience moral alarm or not. Moral alarm is the anxiety experienced in the process of causing harm. Psychopathy is the mental disorder characterised by a lack of moral alarm or more generally, a lack of empathy for the suffering of others (Hare, 1991; Aniskiewicz, 1979). Moral alarm reduces the probability of people making decisions for the 'greater good', known as utilitarian decisions; for example, when given the chance to sacrifice one life in order to save five lives, empathy for the one life typically prevents a person making the utilitarian choice to save the five lives (Greene et al., 2008; Greene et al., 2001; Thomson, 1985). In contrast, the utilitarian choice is more likely to be made by people who report feeling less empathy (Bartels and Pizarro,

2011; Conway and Gawronski, 2013; Gleichgerrcht et al., 2013) and people who score higher on the psychopathy scale (Koenigs et al., 2012). While Cima, Tonnaer and Hauser (2010) observed no effect of psychopathy on utilitarian decision making, this may be attributable to the small sample size and lenient categorisation criteria used (Koenigs et al., 2012).

How might neuroscience promote immoral behaviour?

There are four plausible mechanisms by which learning about neuroscience could promote immoral behaviour. First, by providing a causal explanation of behaviour, neuroscience may challenge belief in free will and thereby reduce the burden of taking moral responsibility for the immoral act, potentially including the experience of guilt; hence the expected costs of behaving immorally may be reduced. Second, neuroscience may challenge belief in a just world (BJW), which is the belief that people receive the outcomes that their actions deserve (Lerner, 1980). In turn, lower BJW may undermine the anticipation that moral acts will be rewarded and that immoral acts will be punished; hence the expected benefit of behaving morally may be reduced. Third, neuroscience may increase belief in fatalism and therefore undermine the perception that one could behave morally even if one aspired to behave morally (Miles, 2013; Dar-Nimrod and Heine, 2011). Fourth, neuroscience could simply increase the believability of feedback (Weisberg et al., 2008; Weisberg et al., 2015), thereby strengthening the probability of people conforming their moral behaviour to what the feedback says about their moral behaviour.

These four accounts of the relationship between neuroscience communication and moral behaviour relate to each other. For example, one would expect learning about neuroscience to reduce BJW by challenging belief in free will: if people do not appear

to make free choices, the outcomes of their choices may appear to be unjust. Prior research has considered the behavioural effects of communicating neuroscience as part of some larger generic argument against free will (e.g., Monroe et al., 2017). In contrast, the current study measured the behavioural effects of communicating neuroscience in direct relation to the participant's own traits and without reference to determinism or free will; hence participants could make their own inferences from the science. I will now consider each proposed mechanism of neuroscientific influence in turn.

The behavioural consequences of belief in free will

With reduced belief in free will, neuroscientifically informed people may experience less incentive or obligation to act morally. To test for such effects, participants are typically asked to read a passage of text that challenges their intuitive belief in free will. As a result of reading this passage, participants have been found to be more likely to cheat passively (Vohs and Schooler, 2008; study 1) or proactively (study 2), less likely to help a person in need (Baumeister et al., 2009; studies 1 and 2), more likely to act aggressively (Baumeister et al., 2009; study 3), more likely to conform to the opinion of others (Alquist, Ainsworth and Baumeister, 2013) and less likely to cooperate for the collective good (Protzko, Ouimette and Schooler, 2016). Therefore, challenges to the existence of free will, such as neuroscience, may undermine moral behaviour. It is important to caveat this conclusion with two studies that have reported null effects of challenging belief in free will on cheating (Open Science Collaboration, 2015) and stealing (Monroe, Brady and Malle, 2017). Hence there is clearly a need to address when challenges to free will do promote immoral choices (Baumeister, 2008).

The behavioural consequences of belief in a just world

The behavioural correlates of belief in free will are consistent with the behavioural correlates of the BJW. Specifically, the BJW predicts delayed gratification (Callan et al., 2009), the pursuit of moral long-term goals (Hafer, 2000; Hafer et al., 2005) and the reduction of the perceived uncertainty implicated in pursuing long-term goals (Bal and van den Bos, 2012). Among young prisoners, the BJW predicts their confidence in the pursuit of moral long-term goals (Otto and Dalbert, 2005) and reduced aggression (Otto and Dalbert, 2005; Dalbert and Filke, 2007). In these respects, there appear to be behavioural benefits of the BJW, perhaps especially for offenders (Sutton and Winnard, 2007). Indeed, the relationship between the BJW and long-term goal pursuit is stronger among disadvantaged people, such as ethnic minority group members and poor people (Laurin et al., 2011). Framing their societal disadvantage as a personal failure – an example of the BJW – may motivate disadvantaged people to climb the social hierarchy. If offenders consider the brain beyond their control, neuroscience might undermine the BJW, as an incentive to change.

Note, however, there also exists an alternative possibility: that people are deemed responsible for the 'moral health' of their brains, as discussed in chapter seven. Following this neoliberal reasoning, neuroscience could actually be used to increase the BJW, such that people are deemed responsible for (reforming) their own neurobiological characteristics.

The behavioural consequences of belief in fatalism

Belief in free will and a just world should be considered alongside belief in fatalism; the idea that certain problems (e.g., antisocial traits) cannot be prevented. Extensive empirical research has measured the behavioural consequences of believing that personality traits can or cannot be changed (Dweck, 2006). This work is founded upon

the distinction between the fixed mindset and the growth mindset, also known as entity theory and incremental theory respectively (Hong et al., 1999). The fixed mindset is the belief that one's traits are fixed entities that cannot be changed. The growth mindset is the belief that one's traits can grow with the incremental investment of effort. Students adopting a fixed mindset have less (intrinsic) motivation to continue an intellectual task after performing poorly on that same task; such poor performance is interpreted as evidence of their unchangeable incompetencies (Dweck, 2006). In contrast, those with a growth mindset interpret failure as an opportunity to learn how to improve their future behaviour. Similarly, in the mental health context, biological explanations have been found to increase patient cynicism about their treatment, yet not undermine compliance (Lebowitz, 2014).

While previous research has focused on measuring the behavioural effects of perceiving intelligence as fixed or capable of growth, I predict similar effects to arise from the equivalent perceptions of criminal behaviour: according to the fixed mindset, the causes of criminal behaviour cannot be changed, whereas, according to the growth mindset, those causes can be changed if offenders invest effort in changing their beliefs and their environment. Although neurocriminology actually constitutes evidence neither of the growth mindset nor the fixed mindset, it is possible that lay people interpret neurocriminology as evidence of the fixed mindset. Therefore, the communication of neurocriminology may encourage offenders to adopt a fixed mindset towards their own offending, such that it appears futile to invest effort in rehabilitation services. Since the efficacy of such services requires the compliance and engagement of the offender, a fixed mindset could compromise their efficacy and undermine desistance. The Facebook study aimed to test this reasoning with lay people: after reading a neurobiological explanation of moral alarm, participants may believe there is simply

little point in trying to change their experience of moral alarm; in turn, this could affect their moral behaviour.

The believability of neuroscience

Finally, neuroscientific feedback could simply be more believable than cognitive feedback; even when the objectivity and accuracy of the methods used to collect that feedback are equivalent. In the current study, the neurobiological explanation might more effectively convince people that there exists a moral alarm and/or that their moral alarm is as weak or as strong as the feedback states. I will now address both these forms of believability in turn.

First, the neurobiological explanation may better convince people that there is such a construct as a moral alarm. Indeed, people are more likely to believe cognitive explanations of common behaviours, when those explanations are supplemented with circular neuroscientific jargon that adds nothing in substantial logic to the original explanation (Weisberg et al., 2008; Weisberg et al., 2015). The current study extended this research by replacing the cognitive explanation with neuroscience that made no reference to brain regions and hence was free of jargon. While Weisberg et al. (2008) believed their neuroscience to be superfluous, the authors did label specific brain regions; hence their neuroscience was arguably less superfluous than the neuroscience presented in the current study.

Second, the neurobiological explanation may render the direction of the feedback more believable. Alquist et al. (2013) found that participants were more likely to conform to the judgements of others after their belief in free will had been challenged or their belief in determinism had been increased. Perhaps people respond to deterministic arguments by substituting some of their self-control for social control,

such that their decisions become more subject to the group consensus. The current study extended this research by testing whether neuroscience, as a challenge to free will, could induce greater conformity to the opinion that their moral alarm was stronger or weaker than average. Hence, by reducing belief in free will, neuroscience may increase the receptivity of participants to other opinions, including scientific opinions.

Facebook analysis

Given the practical and ethical issues implicated in measuring the responses of convicted offenders to personalised neuroscientific feedback, in this study, it was unknown whether the participants had ever offended. In the current age of technology, social media has generated major new opportunities to analyse online behaviour by capturing the 'digital footprints' left by millions of people on social networks. Automatic analyses of these 'big data' can now provide people with personalised feedback about their psychological and physical health. For example, Kosinski et al. (2015) analysed the data of millions of Facebook users to create an algorithm capable of predicting users' gender, sexuality, age, personal interests and political views, based simply on their Facebook likes; the analysis can be accessed by anyone free of charge at www.applymagicsauce.com. These algorithms have even been used to predict psychopathic traits among the general population, including the experience of moral alarm (Garcia and Sikström, 2014). Hence, in the current study, participants consented to a fake analysis of their Facebook likes, before reading that they typically experienced above or below average levels of moral alarm. The behavioural effects of providing this feedback in cognitive or neuroscientific terms were subsequently measured.

Hypotheses

This online survey experiment aimed to capture the behavioural consequences of communicating the neuroscience of psychopathy. Participants read that their moral alarm, expressed in neurobiological or cognitive terms, was either 18-22% stronger or 18-22% weaker than average. It was predicted that participants who read their moral alarm was below average would be more likely to display psychopathic tendencies in subsequent behavioural tasks. Specifically, three hypotheses were proposed:

H1: Participants who read that their moral alarm was weak (rather than strong) would be more likely to cheat and display utilitarian reasoning.

H2: The behavioural effects of the false feedback (H1) would be greater when the feedback was expressed in neurobiological (rather than cognitive) terms.

H3: The behavioural effects of the false feedback (H1) would be mediated by self-control, belief in free will, and the guilt experienced in response to moral dilemmas.

Method

Design

This online experiment adopted an independent groups design. Two variables were manipulated between groups: the communicated degree of psychopathic traits (above-average or below-average) and the description of such traits either in neurobiological or cognitive terms (Table 15). Participants responded through the online survey platform, Qualtrics.

Table 15.

Summary of Conditions for the Facebook Study (n = Number of Participants per Condition).

	Participants read that their moral alarm is…	
	Stronger than average	Weaker than average
A Neurobiological Characteristic	*n* = 179	*n* = 198
A Mental Characteristic	*n* = 197	*n* = 176

Participants

Participants were recruited using adverts posted on social media and distributed through the mailing lists of university departments and societies in Austria, Germany, Italy, Slovenia, Finland, Norway and the UK. Anyone aged 18 or older and capable of understanding English was eligible to participate; before the study began, participants were asked to rate their English competence. Upon completing the study, participants were entered into a lottery, with the chance to win £60.

Procedure

From the outset, participants were informed that the aim of the study was to test a new means of measuring personality traits; specifically, a computerised analysis of a person's Facebook likes that can infer the degree of anxiety that person experiences while committing immoral acts – a trait known as moral alarm that psychopaths lack (Blair, 2006). The study was split into three sections; the full procedure is summarised in Figure 10. In the first section, participants completed the psychopathy scale from the Short Dark Triad (SD3) questionnaire (Jones and Paulhus, 2014). For this scale, participants indicated their agreement (1 = *strongly disagree*, 5 = *strongly agree*) with nine statements describing psychopathic traits; the scale exhibited internal consistency

(Cronbach's α = .69). By including this scale, it became possible to compare the associations between moral decision making and actual psychopathy, and therefore, not only the false feedback that participants received.

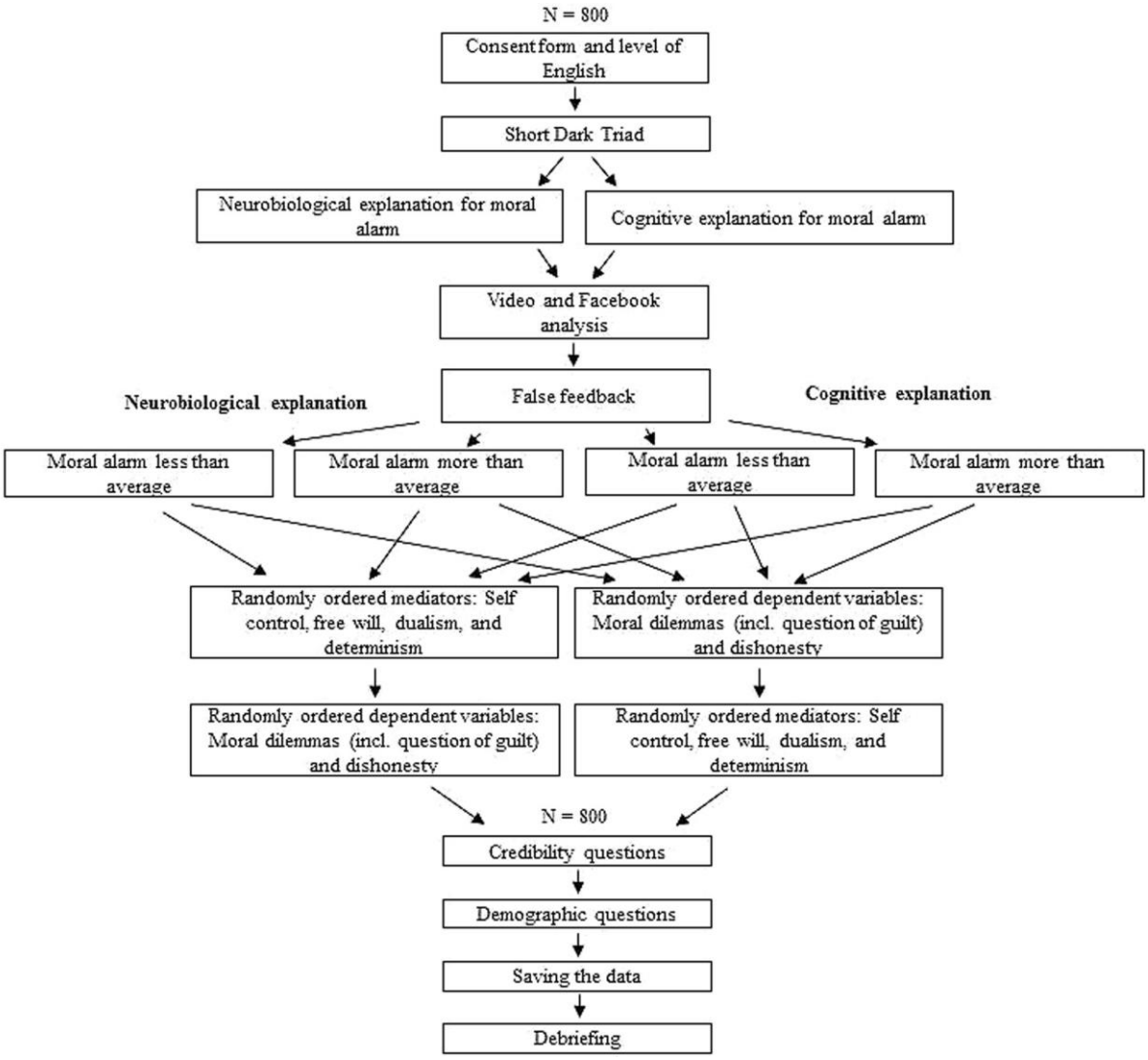


Figure 10. The order of tasks.

In the second part of the study, participants read about several new concepts. To test their comprehension of the concepts, a single multiple-choice question was included at the end of every page of descriptive text. First, participants read either a cognitive or a neurobiological explanation of moral alarm. For each of the various conditions, participants were randomly allocated in equal proportions (e.g., to read the

neurobiological or cognitive explanation). The explanations were adapted from the explanations presented by Aspinwall et al. (2012), whose explanations are based on the neurocognitive model of psychopathy (Blair, 2006). In contrast to the original explanations, references to genetics were removed, such that the two explanations (for the neurobiological and cognitive conditions respectively) were scientifically equivalent and differed by only a few single words.

The explanations read as follows; the words manipulated between the four experimental conditions are italicised:

> The *brain's/mind's* moral alarm // Extensive research shows that human *brains/minds* have a moral alarm. The moral alarm is the *physical/psychological* system that produces feelings of anxiety when you behave badly. // When humans behave badly, their *brain/mind* normally generates particular *electrical signals and chemical reactions/thoughts and emotions* that produce feelings of anxiety. The purpose of this anxiety is to *physically/psychologically* reduce your desire to behave badly.

Subsequently, participants read that it is difficult to infer the degree to which a person experiences moral alarm simply by asking the person directly. People lack insight into their own character and are biased towards presenting themselves in a socially desirable light. In contrast, implicit behaviours, such as liking particular Facebook groups and pages, may provide a more accurate source of information about personality traits, including the experience of moral alarm.

Next, participants watched a video, explaining that researchers have designed a computer algorithm that can analyse a person's Facebook likes to infer even intimate characteristics of the Facebook user (e.g., intelligence, ethnicity and political views). The video combined a variety of national media coverage of the innovative Facebook research conducted by Michal Kosinski (Kosinski, Stillwell and Graepel, 2013). After watching the video, participants entered a shortened and anonymised weblink to their Facebook profile page supposedly so that their own Facebook likes could be analysed

by the computer algorithm. In fact, the entered weblink was not saved, analysed or used in any way. Nevertheless, in the debrief to the study, participants were given a link to the website where one can access the computerised analysis (applymagicsauce.com). After entering the link to their profile, participants received false feedback about their own moral alarm. Moral alarm was deliberately described without referring to psychopathy in order to avoid triggering the popular negative conception of psychopathy. Participants were randomly allocated to read one of four types of feedback, stating that their own moral alarm was either 18-22% stronger or 18-22% weaker than the average moral alarm and again, either in neurobiological or cognitive terms.

The explanation read as follows; the words manipulated between the four experimental conditions are italicised:

> Your *brain/mind* // We would now like to tell you more about people like you, who have an 18-22% *stronger/weaker* moral alarm than the average person. // The moral alarm is the *physical/psychological* system in the *brain/mind* that produces feelings of anxiety when you behave badly. The purpose of this anxiety is to *physically/psychologically* reduce your desire to behave badly. Since your moral alarm is 18–22% *stronger/weaker* than the average moral alarm, you are 18-22% *less/more* likely to behave badly than the average person. This is true of anyone with an 18–22% *stronger/weaker* moral alarm. // People have moral alarms of different strengths because of *physical/psychological* differences in how their *brains/minds* work. When people with a *brain/mind* like yours behave badly, their *brain/mind* generates *more/less* of the *electrical signals and chemical reactions/thoughts and emotions* that produce feelings of anxiety. // Therefore, people with a *brain/mind* like yours feel 18–22% *more/less* anxious when they behave badly. Consequently, people with a *brain/mind* like yours are 18-22% *less/more* likely to behave badly.

The third section of the study required participants to complete measures for the hypothesised mediators and the outcome variables, the order of which were counterbalanced. In other words, half of the participants responded to the mediators, followed by the outcome variables, and the remaining half responded to the outcome variables, followed by the mediators. The items for every scale can be found in

Appendix D. In respect to the mediators, participants completed the Determinism Subscale and the Free Will Subscale of the Free Will Inventory, and a measure of dualistic beliefs. For the Determinism Subscale of the Free Will Inventory (Nadelhoffer et al., 2014), participants indicated their agreement (1 = *strongly disagree*, 7 = *strongly agree*) with five statements in support of determinism (e.g., 'Every event that has ever occurred, including human decisions and actions, was completely determined by prior events'). In the current study, this Determinism Subscale exhibited high internal consistency (Cronbach's $\alpha = .83$).

For the Free Will Subscale of the Free Will Inventory (Nadelhoffer et al., 2014), participants also indicated their agreement (1 = *strongly disagree*, 7 = *strongly agree*) with five statements regarding free will (e.g., 'People ultimately have complete control over their decisions and their actions'). Since it avoids religious terms, the Free Will Inventory was used instead of the alternative Free Will And Determinism Plus (FAD+) scale (Paulhus and Carey, 2011). The Free Will Subscale also exhibited high internal consistency (Cronbach's $\alpha = .82$).

In order to measure dualistic beliefs, participants completed a modified version of the thought experiment designed by Forstmann and Burgmer (2015). The authors asked participants to imagine that scientists had developed a device capable of duplicating any person in a matter of seconds, using highly advanced technology. Participants read that after placing a person in a special chamber, a computer could scan the entire person, every molecule and atom, and store the information digitally. Subsequently, the information could be used to recreate the scanned person from basic chemical elements in a second chamber, resulting in a 100% identical copy of the scanned person, with a 100% success rate.

In contrast to the original task, the participants in this study were asked to imagine that *they* were placed in the chamber and duplicated. Participants then indicated their agreement (1 = *definitely no*, 7 = *definitely yes*) with six statements regarding whether their duplicate shared properties of themselves. Three of the statements concerned psychological properties of relevance to the experimental manipulation (e.g., 'Is the moral alarm in your duplicate the same strength as the moral alarm in you?'). The remaining three statements concerned physical properties (e.g., 'Does your duplicate have the same eye colour as you?'). If people do separate the mind from the body, the participants should ascribe the same physical, yet different psychological, properties to the duplicate of their original selves.

Next, participants completed a test of self-control, using a modified online version of the famous marshmallow test (Mischel and Ebbesen, 1972). Although every participant was entered into a lottery for taking part in the study, participants had to decide about the timing of that prize; specifically, whether they would prefer to know the lottery outcome immediately after completing the study or three months later. The incentive for choosing the latter option – the option that required greater self-control – was a £60 rise in the prize money.

To measure utilitarian reasoning, participants responded to three different types of moral dilemma: a difficult personal dilemma, an impersonal dilemma and an easy personal dilemma. In all of these dilemmas, participants must decide whether to initiate one death in order to save a greater number of other people. For the difficult personal dilemma, participants were given the crying baby dilemma (Greene, Sommerville and Nystrom, 2001). In the crying baby dilemma, participants decide whether to smother their hypothetical child to death to avoid catching the attention of enemy soldiers and thereby save several other lives, including their own. For the impersonal dilemma,

participants were given the standard trolley dilemma. In the standard trolley dilemma (Foot, 1978), participants decide whether to allow a runaway trolley to kill five workmen or to pull a lever that will switch the direction of the trolley, such that only one workman is killed instead. For the easy personal dilemma, participants were given the footbridge dilemma. In the footbridge dilemma (Thomson, 1985), participants decide whether to push a heavy man off a railway bridge to block a runaway trolley from killing five workmen. After making each decision, participants indicated their agreement (1 = *strongly disagree*, 6 = *strongly agree*) with having experienced guilt for making the decision.

To measure dishonesty, participants completed an online version of the die-under-the-cup task (Shalvi et al., 2011). Participants rolled a virtual die three times, under the instruction that the outcome of their first throw would determine the value of the lottery prize for participating in the study: the higher the outcome, the greater the value of the prize (1 = £10, 2 = £20, 3 = £30, 4 = £40, 5 = £50 and 6 = £60). Hence participants were given an incentive to misreport the outcome of their first roll to increase the value of the lottery prize. In place of a physical die, participants simply clicked a button and read that a virtual die had been thrown, with a particular outcome obtained. The outcomes of the virtual die were always a two, then a six, and finally a three in this order; by fixing the outcomes, it was possible to analyse deception at the individual (and not just the group) level. After rolling the die three times, participants were asked to report the outcome of their first roll (from 1 to 6) within a time frame of 30 seconds. Participants were informed that they had to report the outcomes because the computer software was unable to do so since the die was an 'external plug-in'. Participants were warned that if they failed to report the outcome of the roll within 30 seconds, the prize would be fixed to the minimum of £10. The 30 second time limit was

visible from a ticking counter and was included given evidence that time pressure increases dishonesty. It was important to design the task such that potential floor effects (i.e., near zero rates of dishonesty) were avoided (Shalvi, Eldar and Bereby-Meyer, 2012).

The order of presenting each mediator and each outcome variable was randomly determined for every participant. This further instance of counterbalancing was important given the potential for order effects: participants who read that they had a strong moral alarm may have subsequently made more moral decisions, yet that very display of moral decision making may have reduced their subsequent motivation to behave morally. Following moral behaviour, people *become* more likely to behave immorally (Blanken, van de Ven and Zeelenberg, 2015). This effect of moral licensing may result from a change in the capacity or the motivation to act morally (Merritt, Effron and Monin, 2010). The random order of task presentation was designed to overcome such effects.

Before concluding the study, participants indicated the believability (1 = *strongly disbelieved*, 6 = *strongly believed*) of the (below-average or above-average) feedback about their moral alarm and the (neurobiological or cognitive) explanation of their moral alarm. Participants also provided basic demographic information: age, gender, nationality, field of study and which device they used to complete the study. Participants could also enter their email address to receive the lottery prize should they win. The email address was stored separately to their survey responses to protect the anonymity of the data. Finally, participants read an extensive debrief about the false feedback.

Results

The attitudinal and behavioural measures were analysed using a between-subjects MANCOVA; the described Strength of the moral alarm [Weak, Strong] and the Explanation Type [Cognitive, Neurobiological] were entered as between-subjects factors. The participant's actual degree of psychopathy and the believability of both manipulations were entered as control variables. There was only one main effect of the Explanation Type on belief in free will, F (1, 743) = 4.08, p = .044, η_p^2 = .005: participants were significantly more likely to believe in free will after reading the Neurobiological explanation of moral alarm (Neurobiological: M = 4.37, SD = 1.32) than after reading the Cognitive explanation (Cognitive: M = 4.19, SD = 1.22). The results for the other outcome variables were not significant (Table 16). The absence of the predicted effects may be attributable to the fact that 35% of the sample (N = 266) disbelieved or strongly disbelieved one or both manipulations. Therefore, I also ran the same MANCOVA after excluding these disbelievers, which reduced the sample size from 750 to 484; Table 17 indicates the number of participants per condition. Once again, there was a main effect of the Explanation Type on belief in free will, F (1, 477) = 5.76, p = .017, η_p^2 = .012: participants were significantly more likely to believe in free will after reading the Neurobiological explanation of moral alarm (Neurobiological: M = 4.36, SD = 1.25) than after reading the Cognitive explanation (Cognitive: M = 4.15, SD = 1.19).

Table 16.

P Values for the Predominantly Non-Significant Effects of the MANCOVA Run with All Participants (The One Significant Effect Emboldened).

	Explanation Type	Strength of moral alarm	Interaction
Free will	**.04**	.76	.17
Dualism	.55	.88	.48
Determinism	.79	.42	.79
Self-control	.67	.87	.88
Guilt	.66	.06	.66
Utilitarianism	.73	.82	.12
Cheating	.68	.56	.68

Table 17.

Number of Participants in Each Condition After Excluding Disbelievers of the Manipulation.

Explanation Type	Moral Alarm	
	Weak	Strong
Neuroscientific	80	159
Cognitive	108	137

Summary

The results did *not* support the hypotheses. Participants who read that their moral alarm was weak (rather than strong) were *not* more likely to cheat and display utilitarian

reasoning (H1), regardless of whether the moral alarm was described in neurobiological or cognitive terms (H2).

Discussion

It was predicted that lay people would be more likely to cheat and exhibit utilitarian reasoning after reading that their moral alarm was weak (H1), especially when this feedback was stated in neurobiological terms (H2). In fact, no such effects of the neurobiological or cognitive feedback were observed. The feedback also exerted no effects on almost all the remaining outcome measures, including self-control, belief in dualism, belief in determinism, and the guilt experienced in response to the utilitarian dilemmas (H3). Although it may be a false positive (p = .044), one effect was observed: participants were more likely to believe in free will after reading the neurobiological (as compared with the cognitive) explanation of moral alarm.

It is possible that the null effects would not have arisen if participants had faced real moral dilemmas; given evidence that hypothetical moral choices differ from real moral choices (FeldmanHall et al., 2012). It could also be argued that the null effects resulted from the fact that 35% of the participants disbelieved one or both the manipulations. It may therefore appear appropriate to test the hypotheses after excluding those disbelievers. When such an approach was taken, the one effect of neuroscientific feedback on belief in free will remained. However, after excluding 35% of the sample, participants were no longer randomly allocated to experimental conditions. Therefore, although I could control for psychopathy and believability, other third variables (rather than the manipulated feedback) might explain why belief in free will was higher in the neurobiological condition. Nevertheless, the fact that the

neuroscientific feedback exerted this one effect both before and after excluding the disbelievers suggests the effect is, in fact, robust.

The question therefore stands of how the neuroscientific feedback could have *increased* belief in free will. This result is not only in the opposite direction of the predicted effect (H3) but also conflicts with previous findings that neurobiological explanations reduce attributions of free will to specific actors (Nahmias et al., 2007; Shepherd, 2012) and reduce generic belief in free will (Monroe et al., 2017; Shariff et al., 2014). This possible discrepancy between the current result and previous findings may concern how the neurobiological basis of moral alarm was described; notably, in the current study, the science was related to the self and without mention of determinism. In contrast, previous studies described the brain as having a deterministic influence on the behaviour of other people.

People may wish to reject neuroscience that appears to promote fatalistic inferences; in this case, the inference that nothing, not even a change in environments or a change in beliefs, could modify the strength of *their* moral alarm. Although neurobiological causation is not evidence of fatalism, Miles (2013) suggests that lay people spontaneously infer fatalism from determinism; and perhaps also neuroscience. In turn, people may respond to neuroscience by reasserting their seemingly challenged capacity for behaviour change – and in this study, make that reassertion by reporting greater belief in free will.

This account of the unpredicted effect could also explain the surprising findings of Caspar et al. (2017): in their study, female participants were *less* likely to administer electric shocks to a fellow participant (for profit) after reading deterministic messages. Counter to previous research (e.g., Baumeister et al., 2009), therefore, antisocial

behaviour was reduced by belief in determinism. Perhaps their finding also represents a resistance to the feared fatalistic implication of neuroscience: by delivering fewer shocks, the participants actively resisted the opportunity to exploit determinism to justify antisocial behaviour. Since people believe themselves to have more free will than others (Pronin and Kugler, 2010), people may be especially unreceptive to the seemingly fatalistic implications of neuroscience *for oneself*. Therefore, the unexpected findings of the current study and that by Caspar et al. may result from the (unjustified) fear that neuroscience implies fatalism.

Still, greater belief in free will among the neuroscientifically informed participants is especially surprising for the participants who read they had a weak moral alarm. I hypothesised that such participants would have a particularly strong incentive to endorse a fatalistic interpretation of neuroscience; this interpretation would justify immoral behaviours that could be conceived as resulting from their weak, yet unchangeable, moral alarm. In the mental health domain, sufferers of mental disorder 'appear generally receptive' to biological explanations of their disorder, perhaps because the science mitigates the experience of blame and shame for their disorder (Lebowitz, 2014, p.81). In comparison, the current participants did not show that receptivity to the science, even when the science could be used to their advantage.

The current findings are, however, consistent with previously reported null effects of challenging belief in free will on cheating (Open Science Collaboration, 2015) and stealing (Monroe et al., 2017). The null effects are also consistent with two qualitative studies that directly asked offenders for their opinion on neuroscience: in interviews, serious young offenders tended to reject neuroscience and genetics as explanations for their wrongdoing; instead claiming the importance of social influences and asserting their continued capacity for choice, responsibility and blame (Horstkötter,

Berghmans and Ruiter, 2012; Horstkötter, et al., 2014). It is therefore presumptive to expect offenders to want to excuse their actions.

Horstkötter et al. (2014) speculate two interesting incentives for offenders to retain a sense of choice over their decision to offend: their dignity and their identity. In respect to dignity, 'even though [offending] may be a wrong choice, at least, it was their own choice' (p.8). In contrast, biology appears to offer an animalistic explanation that threatens their dignity. In respect to identity, proud offenders may seek to protect 'their own, rebellious, voice', rather than perceive their behaviour to be the predictable outcome of unfortunate biological circumstances (Horstkötter et al., 2014, p.8). One might expect this incentive for retaining belief in free will to be particularly salient among psychopathic offenders. Psychopaths may construe moral responsibility as the opportunity to take credit, rather than receive blame, for their deviant behaviour; in turn, their grandiose depiction of the self can be maintained (Hare, 1993).

Contrary to neutralisation theory (Sykes and Matza, 1957), therefore, people may wish to retain responsibility for their deviant traits. For example, the current participants may have wished to retain 'credit' for their moral alarm, since even a weak moral alarm reaps certain advantages; for example, the ability to stay calm in emotional situations (Dutton, 2013) and the 'ability' to intimidate others (Horstkötter et al., 2014). For psychopaths, 'being good is bad' (Topalli, 2005, p.797). Therefore, belief in free will may function as an important mechanism for retaining the sense of *having developed* a particular identity.

ii. The museum study

In spite of the absence of behavioural effects in the Facebook study, perhaps learning about neurocriminology still affects the ability to recall or willingness to report acts of

minor deviance from the past. Therefore, after reading the biosocial, neurobiological or unexplained account of self-control, participants in the museum study reported how frequently they had committed a series of deviant acts. The following hypothesis was proposed:

H6: Participants would be more likely to report deviant acts from the past after reading the neurobiological (rather than the unexplained) account of self-control.

Results

There was a main effect of reading the Real Explanation of self-control on self-reported deviance, F (2, 4100) = 4.36, p = .013, η_p^2 < .01. In the real world, participants were more likely to report everyday deviance after reading a Neurobiological Explanation of self-control (M = 1.79, SD = .57) than after reading an Unexplained Account of self-control (M = 1.73, SD = .55; p = .003). This effect was mediated by attributions of fault to the offender described in the vignette: the Neurobiological Explanation of self-control (β = .04, p = .031) no longer significantly predicted self-reported deviance (β = .02, p = .172) after attributions of fault were entered into the regression model, whereas attributions of fault did significantly predict deviance (β = -.08, p < .001).

Discussion

Participants were less likely to attribute fault to the neurobiological (than the unexplained) offender, and consequently, were more likely to report their own minor acts of deviance after reading the neurobiological (than the unexplained) account of self-control (H5). Since the participants were randomly allocated to conditions, the finding does not represent a difference in the actual levels of deviance between the two groups. The mediating role of fault suggests that participants may have been more

willing to disclose their own deviance because the neuroscience appeared capable of partially excusing that deviance.

It is unknown whether participants were more able to recall, or more willing to report, minor acts of deviance after reading the neuroscience; the science could amplify the salience or the social acceptability of deviance. If scientific explanation normalised more deviant aspects of the self, one might even expect the science to blur the moral divide between the participant and offenders. In turn, this could constitute an additional mechanism by which neurocriminology might improve perceptions of offenders.

The museum study originally featured a confession task in which participants were asked to write about a time they had resisted or given into temptation. On the third day of the study, the museum required the task to be removed since the disclosed temptations were unexpectedly severe (e.g., infidelity, theft and drug dealing). The museum staff had no concerns about the confession task until they read the responses given by visitors. Therefore, the decision to close down the confession task could partially be interpreted as an attempt to maintain the moral image of visitors to the museum. Ironically, the confession task was included specifically to test the prediction that participants would be less concerned about maintaining their moral image after reading the neurobiological explanation; the deviance finding suggests this was the case.

Practical implications

The deviance finding also poses certain practical implications for the rehabilitation of offenders. An important first step towards effective rehabilitation is acknowledging the problem; the findings of the museum study suggest people are more willing to acknowledge their deviant selves after reading a neurobiological explanation for

deviance. Therefore, perhaps the science could facilitate the recognition (and subsequent correction) of wrongdoing. The Facebook study found that neurocriminology also increased belief in free will, which lay people typically define as self-control (Monroe and Malle, 2010). In turn, stronger belief in self-control might motivate greater engagement in rehabilitation if belief in self-control is a prerequisite of the growth mindset; the prosocial attitude that one's traits can change (Dweck, 2006).

There is reason to believe that a scientifically accurate interpretation of neurocriminology could promote behaviour change. Specifically, the science could increase the likelihood of offenders believing that their prior offending behaviour was caused (by particular factors) and therefore, that by addressing those causes, future offending could be prevented. In other words, neurocriminology could render the causal basis of prior offending a salient target for change. Even if offenders are reluctant to take responsibility for past behaviour (Sykes and Matza, 1957), one can still assert the offender's responsibility for the better informed future self. In fact, the behavioural benefits of believing in free will (e.g., Baumeister et al., 2009) may be the benefits of believing in the capacity to *consciously* control *future* behaviour.

In contrast, belief in free will for past behaviour may be an antisocial force: 'Free will becomes the legitimating excuse that is used to ignore the plight of the most unfortunate' (Miles, 2013, p.213). Indeed, people who believe the world is a just place for other people exhibit more prejudice towards the elderly (Bègue and Bastounis, 2003), the poor (Furnham and Gunter, 1984; Sutton and Douglas, 2005), refugees (Montada, 1998), people with AIDS (Connors and Heaven, 1990) and the unemployed (Reichle et al., 1998). Such prejudice may detract from attempts to understand the causes of disadvantage, including criminogenic disadvantage. As Waller (2011, p.150) claims, 'Moral responsibility practices block inquiry and militate against understanding

(the first step toward changing)'. Therefore, rehabilitation might be a higher priority in a world that better recognised the (neurobiological) causes of offending.

Belief in the efficacy of rehabilitative efforts does not require belief in an incompatibilist free will. Rehabilitation is a process *caused* by changes in the social environment, resultant brain activity and associated beliefs. Therefore, belief in (neurobiological) causation could actually empower belief in the feasibility of behaviour change. The antisocial effects of challenging belief in free will (e.g., Baumeister et al., 2009) could be avoided by clarifying that (neurobiological) causation does not refute self-control or the influence of the environment. The most effective approach to rehabilitation may be to assert the self-control of the offender, yet also recognise that self-control is influenced by environmental factors and hence, seek to manoeuvre those factors to promote self-control. This approach would require lay people, including offenders, to distinguish between incompatibilist free will and self-control. Future research is therefore required to test means of correcting people from intuitively conflating the two constructs (Miles, 2013; Nahmias et al., 2005).

iii. Conclusion

This chapter began by suggesting the communication of neurocriminology might increase immoral behaviour by challenging the belief in (compatibilist or incompatibilist) free will, by challenging the belief that the world is a just place where people get what they deserve, by increasing the fatalistic belief that the future is already mapped out, or simply by increasing the believability of feedback about one's moral traits. The findings of the Facebook study and the museum study suggested a similar, but slightly different, possibility to these prior hypotheses: people may construe neurocriminology as a challenge to their identity, their dignity and their capacity for

change. The response to such an inference was different to the response predicted at the start of this chapter: participants in the Facebook study reported greater belief in free will, perhaps as a means of rejecting the feared misuse of neuroscience to support fatalism. The Facebook study also challenged prior evidence that neuroscientific communications increase immoral behaviour by giving people the chance to relinquish moral responsibility for their actions; no such effects were observed.

The museum study explored the possibility that even if neurocriminology does not increase immoral behaviour in the present, it may increase the recall of immoral behaviour from the past. Indeed, the museum study found that people were more likely to report their own acts of minor deviance after exposure to neurocriminology. In turn, this finding generated a novel practical implication for the rehabilitation of offenders: neurocriminology could actually promote behaviour change by encouraging people to acknowledge their moral weaknesses. In turn, those weaknesses might be better remedied by shifting the focus of rehabilitation programmes away from implicitly blaming offenders for the behavioural outcomes of their moral weaknesses to understanding the causal origins of those moral weaknesses. In this respect, neuroscientific communications could actually reduce immoral behaviour by encouraging a stronger recognition of what causes offending behaviour and so what could prevent future offending behaviour.

Chapter 11: Limits on the impact of neurocriminological ideas

This thesis began by outlining discrepancies between the lay and neuroscientific perspectives on what causes crime and what justifies punishment. Neurocriminologists tend to attribute criminal behaviour to the brain and, in turn, to the genetic and environmental factors that influence brain development and brain activity (Greene and Cohen, 2004). According to this neurobiological explanation, offending is entirely caused by prior uncontrollable factors. Therefore, the offender is never considered *deserving* of punishment; yet punishment might still be appropriate for consequentialist reasons, namely crime prevention. In contrast, lay people prefer environmental explanations of offending (e.g., Gabbidon and Boisvert, 2012) and retributive, rather than consequentialist, justifications for punishment (Carlsmith and Darley, 2008). Lay people implicitly justify retributive punishment on the basis that even when criminal behaviour is entirely the result of prior uncontrollable factors, the typical offender still has the capacity for rational choice (e.g., Nahmias et al., 2005). Therefore, most lay people and incompatibilist neurocriminologists have different views about crime causation and the appropriate justifications for punishment.

I aimed to anticipate the outcome of the discrepancies between these views; specifically, whether lay people might change their perceptions of offenders after considering the neurocriminological perspective. The major discrepancy between the lay and scientific perspectives on punishment, discussed in chapter four, was the first indicator that lay people would be unreceptive to the scientific perspective. Chapters six, eight and nine empirically documented this lacking receptivity in two respects. First, participants did not make the grand mitigating interpretation of neuroscience hypothesised by Greene and Cohen (2004). In other words, although neuroscience had a slight mitigating influence, lay people did not relinquish the retributive justification for

punishment. Second, participants did not make the aggravating interpretation of neuroscience hypothesised by Snead (2007). In other words, lay people did not harness the consequentialist justification for punishment. Both findings document the relative stability of the default lay perspective on punishment in the face of neuroscientific arguments about crime causation.

This chapter aims to understand how these two findings relate to each other by discussing the relationship between retributive and consequentialist justifications for punishment. It is unsurprising that neurocriminology fails to exert major mitigating influence when one considers the tendency for lay attributions of blame to persist independently of causal explanation. The question therefore arises of why people continue to gauge blame by the perceived capacity of the offender to make rational choices rather than free choices; understanding why lay people are compatibilist in this respect should unearth the underlying reason for the resistance of neuroscientific mitigation.

One answer is that the compatibilist basis for positive retributivism is *conveniently* difficult to attack: the typical offender does wish and intend to offend, and no science will disconfirm this fact. Hence, by adopting compatibilist criteria for the blameworthiness dimension of the proportionality principle, retributivist theory provides a robust and *seemingly* principled outlet for a positively retributive instinct. Positive retributivism is the unrestrained pursuit of punishment to blame offenders for their moral wrongdoing rather than the deliberated calculation of their blameworthiness for the act, as retributive theory would require. By only assessing the capacity for rational choice, the compatibilist stops the assessment of blame short of the underlying cause of offending; the free nature of criminal intent is simply assumed or ignored (Kolber, 2016). Such a grand assumption renders this reasoning only superficially

principled, arousing the suspicion that a positively retributive instinct stopped the assessment of blame short of the most logical basis for blame: a free decision to offend that has no prior cause (Harris, 2012).

This chapter begins by discussing whether the influence of neurocriminology is limited by the instinctive pursuit of positive retributivism. Subsequently, the instinct behind that pursuit is examined, with respect to the consequentialist function that such an instinct might fulfil in evolutionary terms. The chapter raises the question of whether the wrongfulness of the criminal act motivates censure, administered via blame, and if so, whether it is possible to censure offenders without blame. The potential consequentialist functions of blame are then discussed in more detail; particular attention is given to the notion that people seek to blame offenders in order to instil guilt, shame and other emotions that discourage offending and to stigmatise offenders to reduce their capacity to survive, flourish and reproduce in this society. In other words, people may attribute blame to facilitate the rehabilitation and/or incapacitation of offenders. Finally, the chapter reports the empirical findings of the museum study and the postcard study, both of which explored whether people might wish to censor neurocriminology (or its mitigating implications) from offenders – the research question at the core of this chapter.

i. Does positive retributivism limit the influence of neurocriminology?

Retributive punishment could plausibly be motivated by positive retributivism and restrained only to the minimum extent necessary to maintain the legitimacy of the CJS:

> Reasons usually given to justify punishment do not explain why it exists. They serve only to conceal the truth, that the scheme of punishment is a barbaric system by which society tries to ‘get even’ with the criminal (Menninger, 1946, p.448).

This reasoning would be consistent with the more general theory that people reach even their 'rational' judgements on emotional grounds, yet then seek to rationalise those judgements with logical argument (Haidt, 2001). In this context, people might reach punitive judgements as a result of positive retributivism, yet then, seek to rationalise their punitiveness in terms of retributive theory that encompasses proportionality. If this theory of punishment is correct, people should seek more severe punishments than would be permitted by the principle of proportionality when their severity can be disguised; in other words, people should be covertly retributive.

Covert retributivism is:

> [The desire] to impose retributive penalties on norm violators that exceed the limits that society places on punishment, and…[hence to] look for alternative covert ways of inflicting pain on norm violators (Fincher and Tetlock, 2015, p.629).

Fincher and Tetlock empirically documented covert retributivism: participants were more likely to send a rapist to a violent (than a non-violent) prison when the rationale for this punitive decision could be disguised by appealing to other described benefits of the violent prison (e.g., cost, security, recidivism rate and education completion rate). Therefore, people appear to be positively retributive rather than (also) negatively retributive when consequentialist concerns can be used to disguise that positive retributivism. The covert desire for the violent prison might be compared to the covert desire to punish the neurobiological offender – both might be examples of positive retributivism that are justified on seemingly legitimate grounds; in the neuroscientific example, the offender's capacity for rational choice. In the theatre, online and museum studies conducted for this thesis, participants clearly sought 'retributive penalties…that exceed the limits that [neurocriminology] places on punishment' (Fincher and Tetlock, 2015, p.629). The distinction between overt and covert motives can be found in other

domains; for example, people report overt disapproval (Scurich and Monahan, 2016), yet covert desire (Hetey and Eberhardt, 2014), for racialized sentencing.

Although Fincher and Tetlock (2015, study 4) have not analysed the lay response to neuroscience, they have examined the lay response to another form of science that advocates for leniency: DNA evidence of wrongful convictions. In this study, half the participants received the opportunity to subject a rapist to state corporal punishment, while the other half received the opportunity to send the rapist to a more violent prison. In other words, the violent nature of the supplementary punishment remained the same, but positive retributivism could only be pursued overtly (corporal punishment) or covertly (the violent prison). After deciding whether to administer the additional punishment, participants learnt that DNA evidence found the convict to be innocent and were then asked to make the same punishment decision for a different rapist. The previous evidence of wrongful conviction reduced overt punitiveness, but not covert punitiveness, towards the second rapist. In other words, the participants were positively retributive, acting with concern for negative retributivism (i.e., responding to the evidence of wrongful conviction) only when their decision could be scrutinised. In contrast, both overt and covert punitiveness towards the second rapist increased when participants read about false acquittals; in other words, when the DNA evidence called for a more punitive (rather than lenient) response. Therefore, people are overtly punitive when that punitiveness can be justified through retributive theory that reaches the same punitive conclusion as positive retributivism.

In sum, the engrained drive for positive retributivism (i.e., covert retributivism) might explain why people adopt the compatibilist criterion for blame that renders neurocriminology incapable of overthrowing positive retributivism. The work of Fincher and Tetlock (2015) raises the interesting possibility that even if

neurocriminology reduces overt punitiveness, covert punitiveness might remain intact. Since compatibilism adopts a narrow (rather than thoroughly encompassing) definition of choice to maintain punitiveness, it might more accurately be described as an emotional, *positively* retributive (rather than deliberated, retributive) basis for sentencing. Therefore, positive retributivism might provide the ultimate limit on neuroscientific mitigation.

ii. Do consequentialist concerns limit the influence of neurocriminology?

If positive retributivism explains the limited influence of neurocriminology, this raises the further question of why people might be so positively retributive. One answer is that a positively retributive instinct evolved to implement the consequentialist goal of upholding group values (Tyler and Boeckmann, 1997). Consequentialist goals have been attached to retributive punishment; for example, von Hirsch (1976) advocates punishment that communicates censure for the seriousness of the offence and the blameworthiness of the offender. Duff (2001) argues that the function of censure extends further into the future, granting the offender a reparative opportunity to repent; that is to endure sorrow as a means of restoring the status of the moral values disrespected by the offending act. Censure is communicated in front of the victim, other potential offenders and society at large, with the aim of reasserting the status of the damaged moral values. Perhaps, therefore, consequentialist concerns motivate positive retributivism, which again would seem to limit the influence of neurocriminology.

Does the wrongfulness of crime motivate blame?

Retributive punishment may function to communicate censure regarding both the seriousness of the offence and the blameworthiness of the offender. Neurocriminology presents a challenge to only one of these two foundations for retributive punishment:

the blameworthiness of the offender, not the seriousness of the offence. In lay sentencing practice, however, the two appear to be inextricably linked: judgements of offence seriousness drive attributions of blame, even though it is unclear why an offender should be more morally responsible for a more severe offence (Knobe, 2003). Hence it is plausible that no matter how strong the neuroscientific argument against blame, the concept will persist simply because a moral wrong has been committed; a fact that neuroscience will never change.

An uncontested source of blame is the perceived intentionality of the offender (Malle et al., 2014). There is a large body of evidence to suggest that people attribute greater intentionality to the offender whose actions *accidentally* cause greater harm, compared to the offender whose same actions accidentally cause less harm (Knobe, 2003; Sripada, 2012a). For example, people attribute greater intentionality to a street robber who causes brain damage (rather than minor bruises) as a result of the victim hitting her head on the edge of the pavement (Tetlock et al., 2010). This 'side-effect effect' suggests that attributions of intentionality increase with the harm caused by the offending act. Clark et al. (2014) extending this research by investigating a different, more generic basis for blame: belief in free will, as observed in experimental and census data. Participants were found to report greater belief in free will after reading about an immoral – as opposed to carefully matched, amoral – event. Actually experiencing an immoral event had the same effect: students reported greater belief in free will after learning that a class member had cheated on the class exam. Even national belief in free will correlates positively with the crime and homicide rates of a country. While the latter finding is correlational, the preceding experiments suggest a causal relationship between the two foundations of retributive punishment:

experiencing a moral violation increased belief in free will. Hence mere judgements of offence seriousness might motivate blame.

Clark et al. (2014) also found that a desire for punishment mediated the relationship: the immoral event initiated punitiveness, which predicted belief in free will. Therefore, people may blame offenders in order to legitimise their independent desire for punishment. Moreover, an unpublished study found that conservatives tend to report greater belief in free will than liberals as a result of judging the same immoral act to be more wrongful (Everett et al., submitted). This finding provides a direct example of people inferring the existence of free will, as a justification for blame, from the wrongfulness of an act. Perhaps conservatives are less tolerant of wrongfulness; since such an intolerance also increases with belief in free will (Martin, Rigoni and Vohs, 2017). Greene and Cohen (2004) expect neuroscience to undermine some grounds for blame independent from the wrongfulness of the act. However, the mere perception of wrongfulness is sufficient to motivate blame and presumably, therefore, to rebut the neuroscientific attack on blame. No matter the neuroscientific disincentive to blame, the incentive of offence seriousness remains.

Can censure be maintained without blame?

If blame functions to communicate censure, one might ask whether censure can be communicated without blame. There may be a substitute for blame that could fulfil its consequentialist function in a neuroscientific world; for example, perhaps the offender could be separated from the offence, such that the offence is censured but not the person who offended (Braithwaite, 1989). However, one might doubt the efficacy of such an approach for effectively upholding moral values. The approach of morality without blame would pose contradictory implications for censure: the act would be censured,

yet the offender would also be considered undeserving of censure. This would be the equivalent of upholding moral values by censuring the offence without ever finding the guilty offender. Without a blameworthy target, censure loses its influence. This might explain why people who report less belief in free will (and so presumably, are also less likely to blame offenders for their actions) are less inclined to believe in the legitimacy of morality as a concept (Bergner and Ramon, 2013). The distinction between right and wrong – and so the appropriateness of censuring wrongs – simply lose meaning in a world without blame, such as the world envisaged by Greene and Cohen (2004). In this context, it appears slightly ludicrous to expect neuroscience to overthrow blame, as the chief means of upholding group values (Tyler and Boeckmann, 1997); this key point will be explored in the subsequent sections.

What is the consequentialist function of blame?

There certainly appears to be a strong relationship between blame and the maintenance of group values. Reminding people of their collective (rather than individual) identity increases victim blaming (Van Prooijen and van den Bos, 2009). In the offender context, people who define themselves in a more collective manner (i.e., in terms of their social relations) report greater support for both 'constructive' (relative to punitive) penalties and retributive (relative to instrumental) penalties (Gollwitzer and Bücklein, 2007). This supports the notion that constructive and retributive penalties could be interrelated rather than conflicting: people with a more social definition of themselves may support retributive penalties more and do so more in the pursuit of 'constructive' goals; such as the restoration of damaged group values. Therefore, retributive punishment may be a means to consequentialist ends (Figure 11).

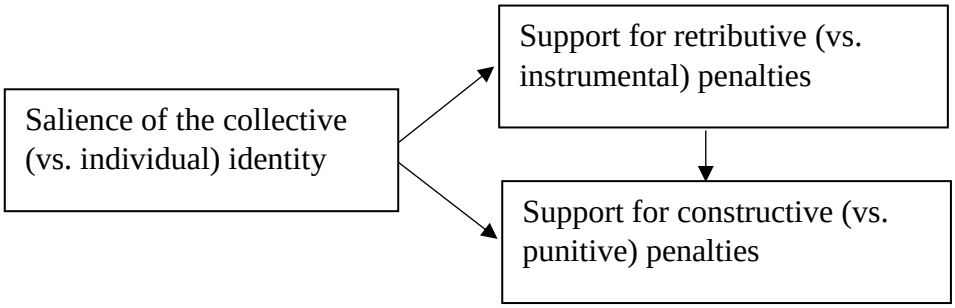


Figure 11. Interpreting the findings of Gollwitzer and Bücklein (2007).

People have a fundamental need to maintain group values. By making moral decisions, people gain the acceptance of ingroup members and thereby increase their sense of belonging to the ingroup (Baumeister and Leary, 1995). Hence social exclusion causes people to act more aggressively (Twenge et al., 2001) and more impulsively (Baumeister et al., 2005); without the desire to belong, moral decision making loses utility. By signalling a commitment to group values, blame restores a damaged sense of belonging to the ingroup. Hence reading about a repeat ingroup offender motivates punitiveness by reducing the sense of group cohesion and increasing moral outrage (Gollwitzer and Keller, 2010). If people abandoned blame in response to neuroscience, this would equate to abandoning a major supply of group belonging. Indeed, people who report less belief in free will do report a lesser sense of belonging and consequently, are less likely to believe that life is meaningful (Moynihan et al., 2017). Therefore, people may resist the neuroscientific attack on blame given the social benefits of the concept, such as the way it seems to bind groups together.

By reserving the supply of group belonging to moral decision makers, blame might discourage moral violations. Indeed, Rucker et al. (2004, p.673) found that participants more strongly supported positively retributive grounds for punishing an ATM thief, after reading about a ‘threat to social order’; for example, a low conviction

rate or the increasing prevalence of ATM theft. The subsequent rise in punitiveness towards the offender was explained by this greater pursuit of positive retributivism rather than the desire to deter this offender or other likely offenders. In other words, consequentialist fears triggered punitiveness by increasing the drive for positive retributivism rather than the drive for deterrence. Perhaps, therefore, the drive for positive retributivism evolved to address consequentialist concerns, such as rising or undetected crime. In this respect, neuroscientific challenges to blame may be resisted in the pursuit of crime prevention. In turn, this raises the question of how blame might discourage offending and therefore, in which contexts neuroscientific challenges to blame might be most resisted.

How does blame fulfil this consequentialist function?

In the evolutionary model of retributivism, there are two mechanisms by which retributive punishment could maintain societal cooperation with group values; those are rehabilitation and incapacitation (Crockett et al., 2014). Although this claim was challenged in the previous chapter (Waller, 2011), blame could facilitate rehabilitation: offenders could be dissuaded from offending by receiving censure for their immoral actions (Fehr and Gächter, 2002). It remains contended whether lay people do seek to punish in order to realise this 'altruistic' goal (Guala, 2012). Blame could also facilitate incapacitation: the consequences of censure could disadvantage offenders in their societal pursuit to survive, reproduce and gain status (Boyd et al., 2003; Rand et al., 2010). The disadvantages of censure (e.g., stigma and incapacitation) would impair their evolutionary fitness; in other words, their capacity to transmit their genes to future generations, where their genes might include genetic precursors to offending (Ferguson, 2010).

There exists an important difference between these two mechanisms of blame: whereas rehabilitation requires the successful communication of censure to the offender, incapacitation does not. This difference raises important predictions for when people are likely to be more or less receptive to neurocriminology. If people punish altruistically – that is with the goal of changing the views of the offender –, participants should be particularly reluctant to neuroscientific mitigation within view of the offender. In contrast, if people are retributively punitive only in order to justify incapacitation, people should be equally unreceptive to neuroscientific mitigation within and beyond the view of the offender. Therefore, to identify moderators of neuroscientific influence, it is important to consider whether people specifically seek to communicate censure to the offender.

Evidence suggests people do seek to communicate blame to the offender. For example, Roberts et al. (2007, p.87) found that the most explicitly supported purpose of punishment was 'making offenders acknowledge and take responsibility for the harm they have done' – a goal that attracted more support than specific deterrence, rehabilitation, incapacitation or victim satisfaction. There is also experimental evidence that victims derive satisfaction from punishment only when censure has been communicated to the offender. In such behavioural experiments, participants are assigned to the position of victim; those victims who subsequently choose to punish an offender report greater satisfaction (Gollwitzer, Meder and Schmitt, 2011) and display fewer implicit signs of anger (Gollwitzer and Denzler, 2009) on one condition: that they also receive a message from the offender, saying he understood the retributive reason for punishment (compared to no message). One might therefore conclude that victims seek retributive punishment for the purpose of communicating blame to the offender.

Since neurocriminology renders blame unjustified, victims may be especially reluctant to neuroscientific mitigation.

Victims may be happier when censure is communicated to the offender, yet this finding does not indicate whether people actually adapt their sentencing decisions to whether that communicative goal can be realised. In other words, people may pursue the communication of censure even when the rehabilitative purpose of retribution cannot be fulfilled. Indeed, third-party judges and victims punish offenders more for unfair (than fair) transactions, even knowing that the offender will never learn that he has been punished (Crockett et al., 2014). This suggests that people still tailor their sentences to retributive concerns even when punishment cannot teach the offender a lesson. Yet still, these concerns are retributive, not consequentialist, because the participants are anonymous and only interact with each player once in the game. Hence participants are most likely responding to retributive concerns with the aim of incapacitating, rather than rehabilitating, the worst offenders. One would therefore expect people to resist neuroscientific mitigation even beyond the view of the offender; where incapacitation is the only relevant function of blame.

One study even suggests that, while it may be satisfying to communicate censure to the offender, sentencing decisions operate independently of that goal (Nadelhoffer et al., 2013b). Specifically, third-party judges punished offenders just as severely in contexts where the offender would never learn that he had been punished. This suggests that people do not tailor sentences to their communicative potential. Note, however, Crockett et al. (2014) failed to replicate this effect: victims and third-party judges punished offenders to a greater extent when the offender would know. Therefore, people may sometimes tailor retributive punishments to the chance of

fulfilling communicative goals. While the evidence is inconsistent, this would suggest people are more likely to resist neuroscientific mitigation within view of the offender.

iii. The museum study

In the museum study, participants made judgements about a seemingly real offender, either believing that the results of the study would be published in a national newspaper or in a national prison newspaper; so either within the view of everyone or explicitly within the view of offenders. In this respect, the study tested the effects of the expected audience of the findings on the inferences participants made from neurocriminology. The neuroscientific mitigation of blame would impede the rehabilitative function of communicating censure to offenders. Therefore, people may be more likely to resist neuroscientific mitigation when the expected spectators of that resistance are offenders. This reasoning generated the following hypothesis:

H7: Participants would recommend a longer sentence for the neurobiological offender when the expected audience of that sentencing decision was the prison population (rather than the general population).

Results

There was no effect of the Target Audience on the recommended sentence ($p > .05$) but there was one effect of the Target Audience on the believability of the Neurobiological Explanation; in other words, these two variables interacted (F (2, 4100) = 5.60, p = .004, $\eta_p^2 < .01$): participants were significantly more likely to believe the Neurobiological Explanation when the Target Audience of their responses was the prison population (M = 4.25, SD = 1.02) rather than the general population (M = 4.10, SD = 1.14; p = .006; Figure 12). There was no equivalent effect of the Target Audience for the Unexplained or Biosocial levels of the Real Explanation ($p > .05$).

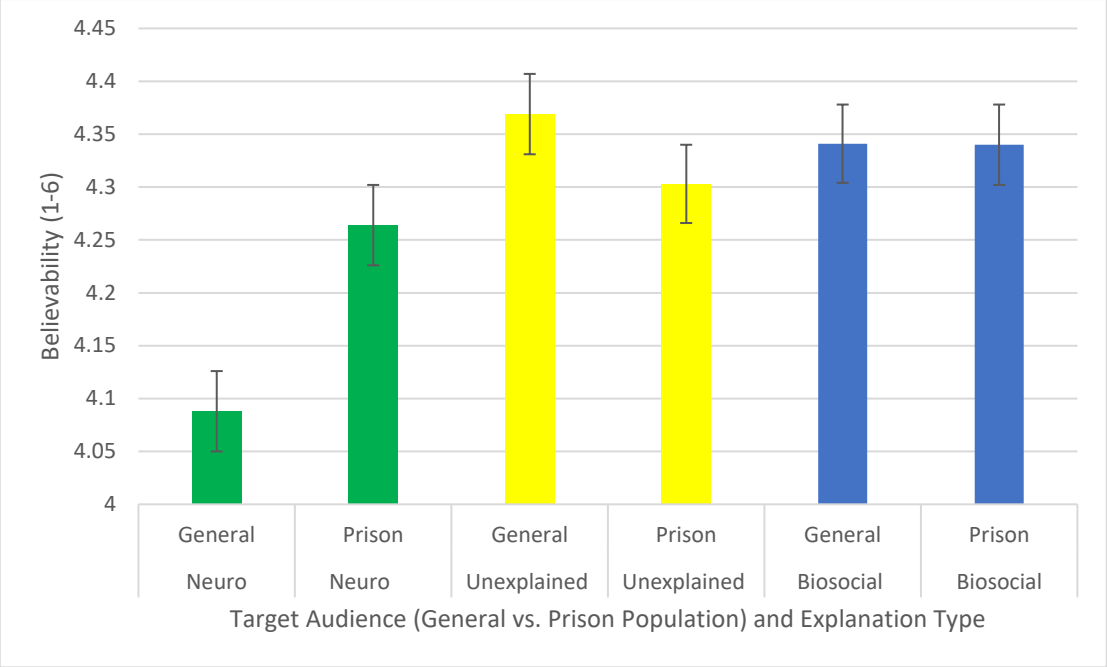


Figure 12. The effect of the Target Audience (General vs. Prison Population) and Explanation Type (Neuro vs. Unexplained vs. Biosocial). Error bars represent the standard error.

Discussion

There was no punitive effect of informing participants that their responses would be communicated to the prison population; hence the study provided no support for H7. In other words, participants were no more resistant of neuroscientific mitigation within the view of offenders. One might consider this finding a ceiling effect: perhaps participants could not resist neuroscientific mitigation any further since that would entail rejecting the relevance of neuroscience entirely. The null effect is consistent with the finding that people punish offenders just as severely when the offender will never learn he has been punished (Nadelhoffer et al., 2013b). These findings suggest that even if the satisfaction of punishment is derived from communicating censure to the offender (Gollwitzer and Denzler, 2009; Gollwitzer and Bushman, 2011), people do not tailor their sentences to the plausibility of gaining that satisfaction.

Participants were more likely to believe in the neurobiological explanation after reading that their views would be communicated to the prison population rather than the general population. Since this increase in believability was not accompanied by any improvement in perceptions of the offender, this raises the question of why believability increased. A punitive perspective on lay intuitions might reason that by finding the science more believable, participants were able to distance their own social identity from that of prisoners (Turner and Tajfel, 1982). Neuroscience might be used to bolster the perceived distance between the participant and prisoners, without also requiring the mitigation of blame. Such a stance would represent a 'dark side' of compatibilism; whereby offenders are considered both blameworthy and fundamentally different from the participant.

iv. The postcard study

The postcard study addressed the same question as the manipulation of the target audience in the museum study; that is the question of whether people fear neurocriminology because of its capacity to impede the communication of censure to offenders. If people perceive neuroscience as a threat to free will (e.g., Nahmias et al., 2007), the decision to communicate neuroscience is a decision to communicate a threat to belief in free will – and by questioning free will, one is questioning the belief that life outcomes, including censure, are deserved (Lerner, 1980). Hence neurocriminology might appear to undermine the legitimacy of expressing censure for criminal life outcomes. The postcard study measured the lay response to the prospect of communicating neurocriminology to offenders using the lost letter technique.

The lost letter technique

The lost letter technique (LLT) is a means of measuring how lay people respond to the address on an envelope – or in this study, the message on a postcard – that is stamped and sealed, yet clearly undelivered (Merritt and Fowler, 1948). The willingness of people to post the apparently lost letter indicates the willingness of people to facilitate delivery of the letter to its intended audience. Participants are unaware of their participation in LLT experiments; hence the method should capture more implicit attitudes towards controversial topics, free from socially desirable responding (Cahill and Sherrets, 1979; Milgram et al., 1965).

In this study, the LLT was adapted, such that participants encountered lost postcards, promoting either a 'brain'-based or 'person'-based explanation of offending, targeted at either prisoners or students. If people consider neuroscience to challenge the existence of free will, people may be reluctant to communicate a brain-based explanation of offending for two reasons. First, people may fear the implications of communicating neuroscience to offenders. Second, people may experience less incentive or obligation to make the 'moral choice' of returning the postcard after reading a brain-based explanation of moral choice. I will now address each of these reasons in turn.

Returning the postcard: Communicating neurocriminology to offenders

It is unclear whether people punish offenders irrespective of the opportunity for the offender to receive the censure conveyed by punishment. While two studies suggest punitiveness is not sensitive to whether the offender will learn he has been punished (Nadelhoffer et al., 2013b; the museum study - see p.267), one study suggests it is (Crockett et al., 2014). Perhaps, therefore, other features of the context moderate the

sensitivity of punitiveness to its communicative function; one such feature might be the group membership of the offender.

When the offender is an ingroup member, people may be more motivated to use retributive punishment to rehabilitate (rather than incapacitate) the offender. This could explain why people pursue positively retributive grounds for punishment, more for ingroup (than outgroup) offenders, when the guilt of the offender is certain (van Prooijen, 2006) and the status of the ingroup is high (van Prooijen and Lam, 2007). Under these conditions, perhaps people are sufficiently motivated to rehabilitate the ingroup offender by communicating censure. In contrast, when the offender is an outgroup member, people may be more motivated to use retributive punishment to incapacitate (rather than rehabilitate) the offender. This could explain why people are more punitive towards ethnic outgroup members, for example (e.g., Steffensmeier and Demuth, 2001), and in fact, covertly desire racialised sentencing (Hetey and Eberhardt, 2014). People may consider the outgroup offender less worthy or capable of rehabilitation and hence, be more motivated to incapacitate the outgroup offender by communicating censure. Therefore, the group membership of the offender may moderate whether the consequentialist function of retributivism becomes rehabilitation or incapacitation.

It is important to consider the consequentialist function of punishment when predicting whether people will be more likely to resist neurocriminology *within the view of the offender*. When the function of blame is rehabilitation, the presence of the offender should exacerbate the resistance of neuroscientific mitigation; because this retributive form of rehabilitation requires the communication of blame to the offender, which neuroscience obstructs. In contrast, when the function of blame is incapacitation, the presence of the offender should not affect the resistance of neuroscientific

mitigation; because incapacitation does not require the communication of blame to the offender. Hence the willingness to communicate a neuroscientific challenge to blame may depend on the function of blame.

In the postcard study, the person-based explanation of offending implied that offenders 'choose' to offend. Therefore, the person-based explanation could also incentivise the *choice to refrain* from offending, since the explanation depicts offending *to be a choice*; in other words, this postcard might be seen to promote the scope for reform and rehabilitation. Consequently, if people wish to communicate censure with the aim of rehabilitating offenders, then people should be especially willing to facilitate delivery of the choice-based postcards *to offenders*. In contrast, the brain-based explanation appears to attribute offending to a different, uncontrollable cause (Nahmias et al., 2007). Therefore, the brain-based explanation could incentivise offenders to resign their lives to a predestined criminal path; in other words, this postcard might be seen to undermine the scope for reform and rehabilitation. Consequently, if people wish to communicate censure, with the aim of rehabilitating offenders, then people should be especially reluctant to facilitate delivery of the brain-based postcards *to offenders*. In sum, if the aim of censure is rehabilitation, the target audience should moderate the return rate of the brain-based postcard.

In contrast, if people express censure to justify incapacitation (rather than rehabilitation), then the brain-based postcards should be preferred, irrespective of whether offenders or students are the intended recipients. By suggesting that criminal behaviour is uncontrollable, the brain-based postcard might be seen to promote the appropriateness of incapacitation (irrespective of the target audience). Therefore, the predicted effect of the target audience on the return rate of the brain-based postcards may depend on the consequentialist function of censure. Given evidence that censure is,

in part, expressed to achieve rehabilitation (Crockett et al., 2014), an important hypothesis follows: people should be less willing to communicate a brain-based explanation of offending to prisoners than to students (H2).

There may also be certain types of people who particularly fear the potential for neuroscience to challenge the legitimacy of censure. In particular, the SES of participants may influence their willingness to communicate neuroscience to prisoners, if the science is seen to challenge their experience of censure. For example, relative to people of moderate status, high-status people are more likely to believe that people earn their positions in societal hierarchies (Calhoun and Cann, 2010; Hunt, 2000; Smith and Green, 1984). Perhaps in the current study, therefore, the richest participants would be the most motivated to assert the legitimacy of their high status, in part, by asserting the legitimacy of the low status of offenders. If neuroscience is believed to conflict with this motivation, then the richest participants may be the least willing to communicate the science. Given the exploratory nature of this hypothesis, its direction was unspecified. Instead, it was simply hypothesised that the reluctance to communicate a brain-based explanation of offending (H1/H2) might interact with the socio-economic status (SES) of the borough where the postcard was dropped (H3).

Returning the postcard: A moral act

It might be considered the morally right choice to return a lost postcard rather than ignore or dismiss it. Indeed, the return rate for lost letters has previously been considered a measure of altruism because, in order to post the envelope, participants must sacrifice their own time for no personal reward (Holland et al., 2012). No empirical study has tested the effects of communicating neuroscience on moral decision making such as altruism. However, as discussed in the preceding chapter, some

evidence suggests that people are less likely to make moral choices after reading an explicit challenge to free will (e.g., Baumeister et al., 2009). Therefore, in the current study, participants may have felt less incentive or obligation to 'make' the moral choice of returning a postcard that attributes moral choice to the brain.

It may appear implausible that people would consciously reason about their own moral responsibility in deciding whether or not to return a postcard. However, neither this study nor prior evidence suggests such effects are mediated by conscious reasoning. When deterministic messages have been found to reduce the probability of people 'deciding' to exercise self-control, the mechanism for this effect is typically unconscious (Rigoni, Pourtois and Brass, 2015) and the decision is made in far less time than conscious reasoning would permit; for example, in 1.28 seconds (Rigoni et al., 2012) or at least less than two seconds (Lynn et al., 2013). Although people accurately recognise that their self-control has been reduced after learning about determinism (Rigoni et al., 2012), exposure to determinism has also been found to delay *preconscious* motor signals in the brain (Rigoni et al., 2011) and not the conscious equivalent of those signals (Rigoni et al., 2015). Similarly, Lynn et al. (2014) found that deterministic messages reduced a person's implicit, but not explicit, sense of agency. Consequently, disbelief in free will may particularly affect moral choices made on intuition, as opposed to reason. For example, Protzko et al. (2016) found that learning about determinism reduced co-operation only when participants were required to make their moral choices quickly rather than slowly, and so, intuitively rather than deliberatively. The decision to return a lost postcard is likely to be intuitive and implicit, not deliberative and explicit. Hence neuroscience could reduce the unconsidered moral motivation to return a postcard.

Hypotheses

This field experiment aimed to capture the willingness of lay people to communicate a neurocriminological claim. The LLT was used to compare the return rates of postcards promoting either a brain-based or person-based explanation of offending for prisoners or students. Three hypotheses were proposed:

H1: Participants would be less likely to facilitate the communication of a brain-based (than a person-based) explanation of offending.

H2: This reluctance to communicate a brain-based explanation of offending (H1) would be greater when prisoners (rather than students) were the intended recipients of the postcard.

H3: The reluctance to communicate a brain-based explanation of offending (H1/H2) might interact with the socio-economic status (SES) of the borough where the postcard was dropped.

Method

Design

This field experiment adopted an independent groups design. Three variables were manipulated between groups: the explanation of offending (the brain or the person), the intended recipients of this explanation (students or prisoners) and the SES of the borough where the postcard was dropped (Table 18). An even distribution of boroughs was selected to represent each of the four categories of SES: very poor, poor, rich and very rich. The stamped and addressed postcards were dropped on pavements on rain-free weekdays in London between 7th March and 14th April 2017. Participants responded simply by deciding whether to place the postcard in a post box; so the

dependent variable was the proportion of returned postcards. It was not possible to gain consent from the participants, who provided no data beyond the decision to return the postcard.

Table 18.

Summary of Conditions for the Postcard Study (Numbers Indicate the Number of Participants per Condition).

	The postcards attributed criminal behaviour to:							
	The person				The brain			
	The postcards were dropped in boroughs defined as:							
	Very poor	Poor	Rich	Very rich	Very poor	Poor	Rich	Very rich
The postcards were addressed to:								
Prisoners	52	52	52	52	52	52	52	52
Students	52	52	52	52	52	52	52	52

Participants

Data were collected from a convenience sample of participants, with no recruitment or selection criteria. Since no questions could be asked of the participants, the demographics of the sample are unknown. The number of postcards dropped (N = 832) was selected to account for the moderate response rates recorded by prior research; from 37% in poor neighbourhoods to 87% in rich neighbourhoods (Holland et al., 2012). In the current study, the return rate was 35.34% (N = 294).

Postcard content

Participants were led to believe that the stamped postcards belonged to a scientific campaign promoting either a brain-based or a person-based explanation of criminal behaviour. According to this cover story, the writer of the postcard had picked up the

postcard, with the intention of sending it to a friend who taught prisoners or students about crime.

The four types of postcard differed only by a few single words. The text on the front cover read as follows; the words manipulated between the four experimental conditions are italicised: 'Most criminal behaviour is completely caused by the failure of the *person*/*brain* to resist impulses' (Figure A5 in Appendix E). On the back cover, the message read as follows:

> [Hand-written:] Hi Alex, I saw this and thought the *prisoners*/*students* you teach might be interested. // [Typed:] Extensive research has discovered that most criminal behaviour is entirely the result of the failure of the *person*/*person's brain* to resist impulses. Please take one of these postcards to help us spread the message *to other people*/*to people who have committed crime*. // [Hand-written:] Sam // [Typed in small print:] This postcard is part of a scientific campaign (Figure A6 in Appendix E).

The postcards were addressed to: '[Hand-written:] Organisation for Educating *Prisoners*/*Students* in Crime, Centre for Criminology, University of Oxford, St Cross Building, St Cross Rd, Oxford, OX1 3UL'. A dot was also added to a corner of the back cover to indicate the SES of the borough where the postcard would be dropped.

London boroughs

The SES of the borough was calculated from the combined average degree of inequality, homelessness, housing quality, unemployment, income, benefits and education (Trust for London and New Policy Institute, 2016). Postcards were dropped in four boroughs from each of the four categories of SES: the poorest (Barking and Dagenham, Newham, Brent, Ealing), the moderately poor (Enfield, Haringey, Waltham Forest, Lewisham), the moderately rich (Hackney, Southwark, Tower Hamlets, Croydon) and the richest (Islington, Lambeth, Merton, Kensington and Chelsea).

Procedure

The postcards were dropped face-up so that pedestrians did not mistake the postcard for a flyer. I and three research assistants wrote and dropped 52 postcards in each of the four categories of London borough. In total, therefore, 208 postcards were dropped in the poorest boroughs, 208 in the moderately poor, 208 in the moderately rich and 208 in the richest (52 brain-based postcards to prisoners, 52 brain-based to students, 52 person-based to prisoners and 52 person-based to students). Each of these 208 postcards were dropped across four time slots: 9-11am, 11am-1pm, 1pm-3pm and 3pm-5pm. While dropping the postcards, each dropper rotated through the four types of postcard. The aim of this distribution process was to randomly allocate participants to conditions, maintaining sufficient distance between postcards to minimise the probability of one participant encountering two postcards.

Results

In the analyses involving SES, one postcard was excluded because the SES of the borough was ambiguous. A chi-square test of goodness of fit was conducted to measure differences between the observed and expected return rates for the conditions of each manipulation: the described Cause of offending [Brain, Person], the Recipients of the postcard [Prisoners, Students] and the SES of the borough [Poorest, Poor, Rich, Richest]. There was no effect of the described Cause of offending ($p > .05$) or the Recipients of the postcard ($p > .05$). There was, however, an effect of SES, χ^2 (3, N = 294) = 14.75, p = .002. The postcards dropped in the very rich boroughs were 74.12% more likely to be returned than those dropped in the moderately rich boroughs, χ^2 (1, N = 159) = 11.63, p = .001, were 50.75% more likely to be returned than those dropped in the moderately poor boroughs, χ^2 (1, N = 168) = 6.88, p = .009, and were 50.75% more

likely to be returned than those dropped in the very poor boroughs, χ^2 (1, N = 168) = 6.88, p = .009 (Figure 13).

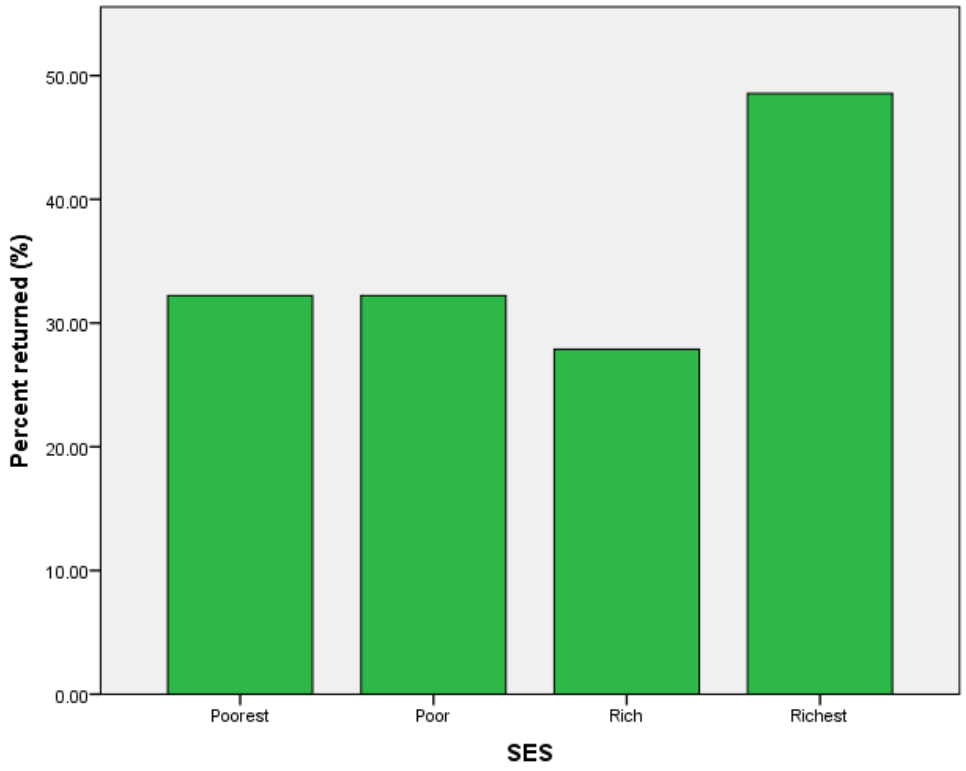


Figure 13. The effect of SES on the return rate.

A chi-square test of association was conducted to measure the relative independence of the Cause [Brain, Person], the Recipients [Prisoners, Students] and the SES [Extreme, Moderate]. There was a significant association between the Cause and Recipients at the Extreme level of SES, χ^2 (1, N = 293) = 4.57, p = .033, Phi = .17, but not at the Moderate level of SES (p > .05). In boroughs of extreme SES, the brain-based postcards were 35.85% less likely to be returned when the recipients were prisoners rather than students (Figure 14; odds = .64), χ^2 (1, N = 87) = 4.15, p = .042. In contrast, the person-based postcards were 25.02% more likely to be returned when the recipients were prisoners rather than students (odds = 1.25), though this difference was not significant (p > .05). Therefore, the odds of the participants returning the prisoner-

intended postcards were 48.69% lower when the postcard promoted a brain-based (rather than a person-based) explanation of offending (odds ratio = .51); this effect was not observed in boroughs of moderate SES (Figure 15). Table 19 shows the return rate for every condition.

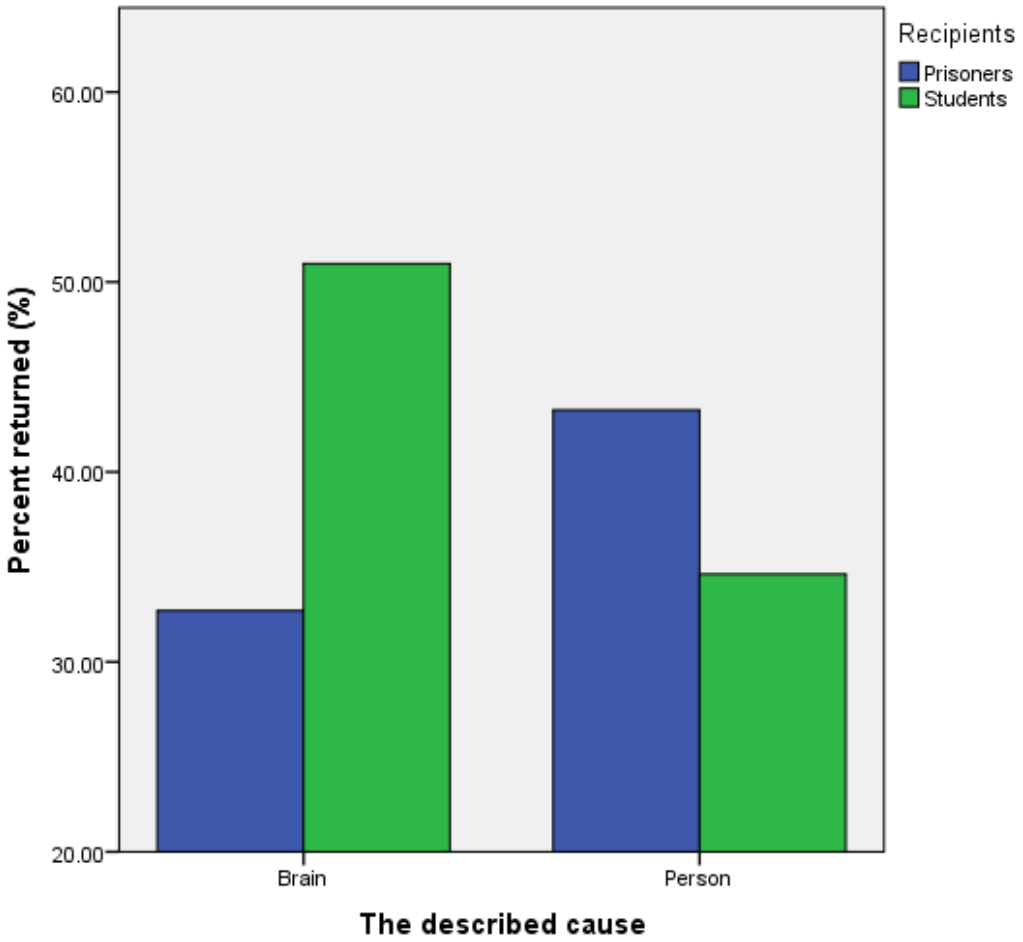


Figure 14. The effects of the described Cause and the intended Recipients in boroughs of extreme SES.

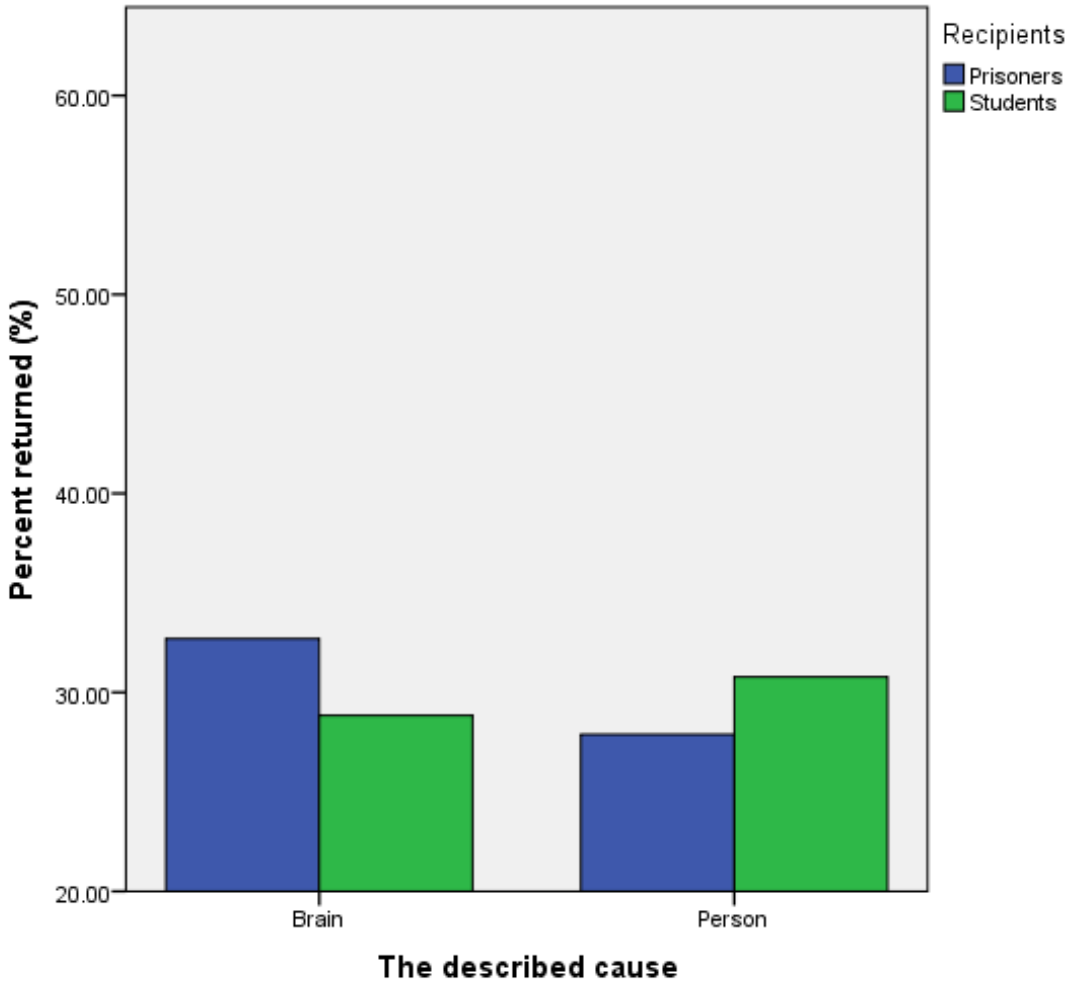


Figure 15. The effects of the described Cause and the intended Recipients in boroughs of moderate SES.

Table 19.

The Proportion (%) of Postcards Returned in Each Condition.

		The described Cause	
Recipients	SES	Person	Brain
Students	Extreme	34.62	50.96
	Moderate	30.77	28.85
Prisoners	Extreme	43.27	32.69
	Moderate	27.88	32.69

Summary

The results did *not* support the first two hypotheses. Participants were equally willing to communicate the brain-based and person-based explanations of offending (H1) for both offenders and students (H2). However, the third hypothesis was supported: participants were less willing to communicate the brain-based explanation of offending to prisoners than to students, only in boroughs of extreme SES (H3).

Discussion

It was hypothesised that lay people would seek to censor the communication of neurocriminology (H1), especially when prisoners were the intended recipients (H2). In contrast to these hypotheses, people were equally willing to communicate the brain-based and person-based explanations of offending for both offenders and students. Nevertheless, people were more reluctant to communicate the brain-based explanation of offending to prisoners than to students, in boroughs of extreme SES (very poor: Barking and Dagenham, Newham, Brent, Ealing; very rich: Islington, Lambeth, Merton, Kensington and Chelsea). Supporting H3, this effect was not found for the person-based explanation or in boroughs of moderate SES (moderately poor: Enfield, Haringey, Waltham Forest, Lewisham; moderately rich: Hackney, Southwark, Tower Hamlets, Croydon). Although it may be a fluke result, I will discuss possible explanations of the finding that are consistent with theory and previous research.

It could be suggested that the brain-based explanation reduced the perceived incentive or obligation to make the 'moral choice' of returning the postcard (e.g., Baumeister et al., 2009). However, if this was true, the effect should not be conditional on the intended recipients of the postcard. Since the effect was conditional on SES and the intended recipients, it is more reasonable to argue that people of a certain SES are

genuinely reluctant to communicate neurocriminology to offenders in particular. I interpret this finding from the perspective of neuroscience as a threat to the BJW (belief in a just world): people consider the brain to be at least somewhat beyond conscious control (Nahmias et al., 2007; Shepherd, 2012). Hence the communication of neuroscience may challenge belief in free will (Shariff et al., 2014) and consequently, the BJW. There exist different incentives for protecting the BJW among the richest and poorest people: the incentives of incapacitating and rehabilitating offenders respectively.

First, the richest people may wish to censor scientific reasons to question the deserved nature of their high status and so also, the low status of offenders, for example. Indeed, high-status people tend to make attributions (Kraus et al., 2009) and support policies (Messick and Sentis, 1979) that legitimise the status hierarchy, tend to report greater BJW (Calhoun and Cann, 2010; Hunt, 2000; Smith and Green, 1984) and be happier as a result (Napier and Jost, 2008; O'Brien and Major, 2005). For this reason, the richest participants in the current study may have been reluctant to communicate a challenge to the BJW, perhaps to maintain the appearance of the hierarchy as legitimate and so, punishment as deserved.

Second, the poorest participants may have also sought to censor a neuroscientific challenge to the BJW, yet for a different reason: to maintain the legitimate appearance of the hierarchy, such that it seems possible to earn status (Jost and Banaji, 1994; Jost et al., 2004; Jost et al., 2003). Hence some studies find that lower status people report greater BJW (Hunt, 2000; Smith and Green, 1984) and are less likely to measure status according to biological criteria (Mahalingam, 2007; Kraus and Keltner, 2013). Perhaps these individuals wish to believe in the legitimacy of the hierarchy in order to believe in social mobility and the value of *trying* to gain status

(Laurin et al., 2011); including the value of offenders *trying* to reform themselves. Therefore, the poorest participants may have been reluctant to communicate a challenge to the BJW, perhaps to maintain the perception that, simply through choice, offenders can change.

In sum, the richest participants had an incapacitatory incentive to maintain the BJW, whereas the poorest participants had a rehabilitative incentive to maintain the BJW (Lerner, 1980; Jost and Banaji, 1994). This is certainly a speculative explanation of why these particular participants were reluctant to communicate neuroscience to offenders; however, there is evidence to support the reasoning. Another explanation would be that participants believed the neuroscience would be more relevant and interesting to students than prisoners. Yet it is unclear why only the poorest and richest participants would believe so and why the other (person-based) explanation would not also be deemed more relevant to students.

One conclusion is clear: participants tended to respond differently to the brain-based explanation compared to the person-based explanation. Therefore, some people did not equate the brain to the person, supporting the idea that lay people are intuitively dualistic (Bloom, 2004; Forstmann and Burgmer, 2015). It is also clear that participants were far more willing to return the postcards in the richest boroughs – an effect that replicates the key finding of Holland et al. (2012). In the richest boroughs, people may have more time or incentive to help others, given a greater sense of affiliation to the community (Sampson, 1997; Shaw and McKay, 1942).

The effect of the postcard content was only found in boroughs of extreme SES; hence it was not a robust effect and there was no overall effect. Since this study used a single dependent measure, there were no further indicators to explain why there was no

overall drive to censor the neuroscience. One explanation is that people are not so dualistic that the brain-based explanation would be treated that differently from the person-based explanation. Another possibility is that the brain-based explanation was not interpreted as a major threat to free will (and so the BJW). Indeed, people continue to attribute free will to caused actors (e.g., Nahmias et al., 2005), including actors whose brain causes them to act in a particular way (Mele, 2012; Nahmias et al., 2014). There is even evidence that people consider brain-based traits to be the most controllable (Fernandez-Duque and Schwartz, 2016). Therefore, people may seek to censor explicit challenges to free will, yet not see neuroscience as such a challenge. A third explanation is that the brain-based postcard did challenge belief in free will, yet the implications for the general and prison populations are weighted equally or the motivation for moral behaviour is unaffected. Indeed, three previous studies failed to document effects of disbelief in free will on cheating (Open Science Collaboration, 2015), stealing (Monroe et al., 2017) and moral decision making (the Facebook study). Therefore, moral choices may be somewhat immune to disbelief in free will.

It is also worth considering the simplest explanation of the null effects; that people did not read the postcard. Although the LLT was clearly sensitive to the manipulations used in previous research (Milgram et al., 1965; Cherulnik, 1975), the current study was the first to apply the manipulation text to a postcard rather than an envelope. While this allowed the manipulation text to be longer, people may have attended less to the text as a result. Indeed, the response rate (35%) was lower than that reported by previous researchers (e.g., 61%; Holland et al., 2012). Perhaps, therefore, the lost postcards were less sensitive to manipulations of the text, explaining the overall null effects.

In sum, this exploratory study used the LLT to measure public approval for the communication of neurocriminology. In boroughs of extreme SES, participants were more likely to censor a brain-based explanation of offending for prisoners than for students (yet not a person-based explanation). Therefore, people of extreme SES may be reluctant to educate offenders in the neurobiological causes of their criminal actions. The Facebook study, however, suggests this possible fear of disseminating neuroscience is empirically unjustified: neuroscience did not increase immoral behaviour. Future research is therefore required to explain why certain people fear communicating neuroscience, especially as it relates to the BJW.

v. Conclusion

> Denunciation is typically regarded as a retributivist rationale, although if its purpose is in part educative, then it is forward-looking (Spiranovic et al., 2011, p.290).
>
> Ironically, the goals served by retribution may ultimately be instrumental (Aharoni et al., 2007, p.882).

As these two quotes suggest, blame is a useful concept because it legitimises censure. Consequently, positive retributivism may motivate blame whenever a moral wrong is committed, no matter how strong the neuroscientific argument against blame. In particular, positive retributivism may motivate the incapacitation and rehabilitation of offenders (Crockett et al., 2014). Specifically, people may blame offenders in order to generate guilt, shame and other emotions in the offender, thereby discouraging possible future offending, and in order to stigmatise the offender, thereby undermining the capacity for that offender to survive, flourish and reproduce in this society. This drive to rehabilitate and incapacitate offenders, and thereby prevent further wrongdoing, could account for the resistance of neurocriminological ideas:

> [People] become warier of situational justifications and excuses for violating society's rules to the degree one has been induced to believe, or is already

> predisposed to believe, that…growing numbers of citizens…are tempted to behave badly (Tetlock, 2002, p.462).

Similarly, people may 'become warier of [neurobiological] justifications' with increases in their awareness of crime. People might even fear that 'growing numbers of citizens…are tempted to behave badly' by the availability of neuroscientific explanations for criminal acts.

This theory might explain the limits in the mitigating inferences that people make from neurocriminology, as documented in chapters five to eight. However, the current chapter found little evidence that people actively censure neurocriminology from offenders. In the museum study, participants were no less likely to make mitigating inferences from neurocriminology when they knew those inferences would be communicated to offenders. In the postcard study, participants were no less willing to communicate neurocriminology to offenders than to students. These findings may simply suggest that people do not consider any consequentialist functions of blame to depend on preserving the legitimacy of blame in the eyes of offenders more than in the public eye. One caveat to these conclusions was found in the postcard study: in areas of extreme SES, participants were less willing to communicate neurocriminology to offenders than to students. Therefore, people of extreme SES may be more motivated to censor neurocriminology from offenders; though future research is required to replicate this finding and explore what those motivations might be.

Chapter 12: Conclusion

The key conclusion of the current research is that lay people seem reluctant to accept neuroscientific grounds for mitigating blame and punishment; those grounds being offenders do not deserve punishment, since their criminal behaviour arises from the brain, which, in turn, is the product of uncontrollable factors. In this thesis, I have argued that the drive for positive retributivism, invoked by moral violations, renders lay people reluctant to relinquish the retributive justification for punishment (Fehr and Gächter, 2002). The drive for positive retributivism may have evolved to maintain effective censure, which requires a blameworthy target (Waller, 2014) and seems to have the function of rehabilitating and incapacitating offenders (Crockett et al., 2014), thereby generating a sense of belonging to the moral ingroup (Tyler and Boeckmann, 1997). These consequentialist functions of blame may motivate the resistance of any science that challenges the legitimacy of blame (Fincher and Tetlock, 2015), such as neurocriminology. Therefore, attempts by neurocriminologists to change attitudes towards crime and punishment must always compete with the engrained function of positive retributivism. In turn, neurocriminology may have little effect on attitudes towards offenders.

This thesis empirically documented the limited influence of neurocriminology in a variety of contexts; specifically, contexts in which the vignette linked the neuroscientific explanation to the judged offending to differing extents. Across three studies, neuroscience was presented as a correlational explanation of teenage impulsivity in general (the theatre study), the impulsive character of the judged offender in particular (the online study), and a causal, complete and universally applicable explanation of self-control in general (the museum study). Each of these studies represented a broader approach to the study of neuroscientific influence: the

psychological approach, the mock court approach and the philosophical approach respectively.

An original contribution of this thesis was the focus upon a realistic use of neuroscience that would suggest widespread implications for the CJS. Specifically, in every vignette, neuroscience was used to explain the actions of an impulsive offender; in other words, a common type of offender who knows the difference between right and wrong, and offends following some conscious intent and desire to offend based on an accurate sense of reality. By presenting neuroscience in this realistic context, it became possible to consider mechanisms of neuroscientific influence beyond those that constitute basic misinterpretations of the science; in other words, misinterpretations of the neurobiological offender as unconscious or influenced in a way that bypasses the conscious mind.

This thesis began by posing four research questions:

1. Do neurocriminological ideas reduce attributions of blame?
2. Do neurocriminological ideas increase attributions of dangerousness?
3. Do neurocriminological ideas increase immoral behaviour?
4. Do people seek to censor neurocriminological ideas from offenders?

Chapters five to eight addressed the first of these questions, examining whether lay people might attribute less blame to offenders for their criminal actions, after learning about neurobiological causes of those actions. Chapter nine addressed the second of the four research questions, discussing whether people might consider offenders more likely to reoffend, and less amenable to rehabilitation, after learning about neurocriminology. All of chapters five to nine considered whether the hypothesised changes in attitudes towards blame and dangerousness, respectively, pose implications

for punitiveness. Chapter ten addressed the third of the four research questions, examining whether neurocriminology might promote immoral behaviour if people use the science to somewhat excuse immoral behaviour. Finally, chapter eleven addressed the fourth of the four research questions, discussing why lay attitudes might be resistant to change in response to neurocriminology; in particular, people may make attributions of blame *in order* to discourage immoral behaviour, and given this entrenched function of blame, people may seek to challenge scientific communications that threaten the concept. Table 20 provides a summary of the answers to the four research questions of the thesis, as suggested by each of the five studies conducted. This concluding chapter explores each of the four research questions in turn, relating the present findings to relevant theory and previous research.

Table 20.

A Summary of the Answers to the Four Research Questions Suggested by the Empirical Studies Conducted for this Thesis.

	The psychological approach	The mock court approach			The philosophical approach
	The theatre study	The online study	The postcard study	The Facebook study	The museum study
Participants judged:	A young offender	A rational or an irrational offender	The act of communicating neurocriminology	The psychopathic brain	A neurobiological or biosocial offender
Suggested answers to the four research questions					
Do neurocriminological ideas reduce attributions of blame?	Slightly	Slightly	-	-	Slightly
Do neurocriminological ideas increase attributions of dangerousness?	No	No	-	-	No
Do neurocriminological ideas increase immoral behaviour?	-	-	-	No	-
Do people seek to censor neurocriminological ideas from offenders?	-	-	No	-	No

i. Do neurocriminological ideas reduce attributions of blame?

In this ecologically valid context, the influence of neurocriminology was limited, regardless of how the science was presented. Such converging findings suggest that Greene and Cohen (2004) may be wrong to predict revolutionary effects of neuroscience. In turn, the prediction of revolutionary effects may result from an

unfounded assumption about how people blame offenders; that assumption being that people blame offenders by retracing the causal chain of their behaviour to the very beginning and calculating the degree of incompatibilist free will that lies at this beginning (Greene and Cohen, 2004). Together with previous research (e.g., Nahmias et al., 2005), the museum study suggests this is not the case: people continue to blame offenders whose behaviour is entirely caused by their brain.

When the neurobiological argument did mitigate the sentence in the museum study, this resulted from the offender appearing to have a less immoral character, not a less blameworthy character (the online study), and less self-control, not less incompatibilist free will (the museum and online studies). Therefore, neurocriminology may gain attitudinal influence by changing perceptions of the capacity of the offender to exercise rational choice at the time of the crime; rather than by prompting people to trace the causal chain of events back to the first cause of their behaviour (which would also be difficult for people to envisage, since the first cause would have taken place before birth, according to determinism). By extension, the museum study suggests that people may perceive the brain to obstruct moral character. Consequently, the neurobiological offender may appear to lack not just the capacity to act upon reason, but the capacity to act upon reasons informed by his stable character. In turn, the neurobiological offender may be considered less deserving of punishment for both legally relevant and irrelevant reasons.

A blameworthy alternative to the free self in a neuroscientific world

An important conclusion of this thesis is that neuroscience may influence attitudes through a very different mechanism to the one proposed by Greene and Cohen (2004). Whereas these authors claim that neuroscience will end belief in an incompatibilist

notion of free will, I suggest that in fact, neuroscience may simply improve moral perceptions of the Deep Self of the offender. The very nature of this mechanism may explain why the effects of neuroscience were limited. I argue that the Deep Self Model poses less severe, and so less threatening, consequences for the concept of blame.

Although the Deep Self Model would predict that people blame the neurobiological offender less (than the unexplained or biosocial offender), the Model would still predict sizeable attributions of blame to be made to the neurobiological offender. In this Model, the mitigating influence of neurocriminology derives from the perception that, at some point in time before the crime, the neurobiological offender lacked the intent to offend. Yet this offender is still conceived to have made a *conscious* choice to offend at the time of the crime. Indeed, the presence of conscious choice is sufficient for people to blame predetermined offenders in the real world (Nahmias et al., 2006; Nahmias et al., 2007; Nichols and Knobe, 2007).

It is possible that people prefer to premise blame on attributions of incompatibilist free will yet rely on the compatibilist concern of conscious choice when it becomes the only remaining justification for blame. Indeed, there is evidence that people prefer to believe in the existence of an incompatibilist free will (Nichols and Knobe, 2007; Sarkissian et al., 2010); perhaps they also prefer that belief to justify the attribution of moral responsibility to offenders. After all, this incompatibilist criterion for blame is adopted in judgements of hypothetical offenders who do not appear to represent real offenders (Nahmias et al., 2007; Roskies and Nichols, 2008). Yet, when determinism is applied to the real world, participants appear to shift towards a compatibilist criterion for free will, as the only remaining grounds for attributing blame (e.g., Nahmias et al., 2005).

Similarly, when neurocriminology challenges the existence of an incompatibilist free will, people may shift to the only alternative justification for blame: the continued existence of conscious choice. In this respect, neurocriminology may motivate a covert shift in the grounds on which people attribute moral responsibility:

> People [may] *want* incompatibilist free will, but, if they find themselves in a [neuroscientific] situation…without it, they will settle for a less satisfying kind of compatibilist free will instead (Nahmias et al., 2005, p.575).

The communication of neuroscience may render increasing numbers of people 'without it' – lacking belief in incompatibilist free will. In response, people may simply render incompatibilist free will irrelevant to moral responsibility.

Lay conceptions of the brain may be similarly motivated. By default, people may conceive the brain as unconscious and this conception may arise in judgements that pose no consequences for neurobiological actors in real life. Indeed, people do not attribute choice to neurobiological actors who appear to be unconscious or consciously helpless – perhaps because these actors are not conceived to represent real humans (Nahmias et al., 2007). Similarly, people do not consider the brain sufficient or necessary to attribute agency to non-human actors; studies have found that participants attribute agency to certain actors that clearly lack a brain, such as organisations (Knobe and Prinz, 2008), God and robots, and also do not attribute agency to certain actors that clearly have brains, such as frogs and newborn babies (Gray, Gray and Wegner, 2007). Collectively, this research suggests that by default, the brain is perceived as unconscious.

The perception of the brain as unconscious would typically excuse neurobiological offenders since not even the compatibilist criterion for blame is met in these cases (Nahmias et al., 2007). Yet people also perceive traits they believe to be brain-based as stable across the lifespan (Fernandez-Duque and Schwartz, 2016). If

stable brain-based traits are conceived to form the Deep Self and people hold offenders responsible for their Deep Self (Sripada, 2010), one would expect people to blame the neurobiological offender. Therefore, the lay person might experience a certain cognitive dissonance between blaming the offender – since he has an immoral Deep Self – and excusing the offender – since he is unconscious or consciously helpless (Festinger, 1957). In turn, the adopted solution to this dissonance may vary between hypothetical and real-world contexts.

The dissonance could be resolved in one of two ways: either by excusing the offender or by shifting the conception of the brain as an unconscious entity to a conscious entity, such that the compatibilist criterion for blame is met. In seemingly hypothetical scenarios, people may be content with excusing the neurobiological offender (Nahmias et al., 2007). However, in real-life scenarios, where blame is motivated by the violation of moral values (Clark et al., 2014), people may shift their conception of the brain to a conscious entity and their criterion for blame to conscious choice. It then becomes possible to blame the real neurobiological offender; the compatibilist perspective legitimises the positively retributive punishment of an immoral Deep Self.

This does not render neuroscientific arguments for mitigation powerless; it simply highlights a limit on the mitigating perception of situationally induced brain activity as an obstruction to the Deep Self. That limit is the competing perception of brain-based traits as dispositional and so as reflecting an immoral Deep Self, thereby rendering the neurobiological offender blameworthy (Fernandez-Duque and Schwartz, 2016). The dispositional or situational source of brain activity is ambiguous; hence the Deep Self model describes a mechanism for neuroscientific mitigation that is certainly

more limited, and perhaps for this reason, more readily enacted, than the mechanism proposed by Greene and Cohen (2004).

ii. Do neurocriminological ideas increase attributions of dangerousness?

Some theorists fear that neurocriminology will promote the fatalistic belief that rehabilitation is futile, thereby increasing support for the incapacitation of offenders and potential offenders to prevent offending (Rafter, 2008; Snead, 2007). However, the current research documented no evidence that people make this aggravating interpretation of neurocriminology. Most noticeably, in the online study, participants did not consider even a *violent* impulsive offender more dangerous or more in need of preventative detention after reading a neurobiological explanation of the impulsive disposition. In fact, in the theatre study, participants believed the adolescent offender was less likely to reoffend after learning about the neuroscience of adolescence; therefore, people are receptive to the message that their neurobiological tendencies change over time. Together, these findings provide some empirical rest for the fear that neurocriminology might propel attributions of dangerousness.

Fear that the science will be abused may result from the unfounded assumption that people calculate dangerousness from their assessment of the very first cause of criminal behaviour. Yet perhaps the fear of abuse is unjustified if people actually infer dangerousness from the moral character of the offender and his capacity to exercise rational choice. For example, in the museum study, the offender was perceived to have less self-control as a result of the neuroscientific testimony. Although one might expect the offender to subsequently appear more dangerous, neuroscience also improved perceptions of his moral character. Hence the potential for reduced attributions of self-control to increase perceived dangerousness might be countered by improved

perceptions of moral character. In this respect, there is a good theoretical reason to question the hypothesis that dangerousness will be inferred from neurocriminology.

iii. Do neurocriminological ideas increase immoral behaviour?

This thesis also assessed the potential for neurocriminology to incentivise immoral behaviour by appearing to excuse, and thereby legitimise, deviance. Although exposure to the science increased the willingness of participants to recall past deviance, it did not increase deviance in the present. In fact, when participants learnt how the science explained their own moral behaviour, this only increased their belief in free will. It is therefore possible that the communication of neurocriminology actually promotes behaviour change: people may be more likely to recognise the causes of their prior deviance and to believe that those causes can be addressed. By challenging an incompatibilist notion of free will, yet asserting self-control, neurocriminology may have prosocial implications.

iv. Do people seek to censor neurocriminological ideas from offenders?

Regardless of whether blame actually 'works' to reduce immoral behaviour, it is plausible that people blame offenders with the unconscious goal of reducing further offending. Indeed, through the attribution of blame, people may be seeking to instil in the offender guilt, shame and other emotions that discourage offending, and to stigmatise offenders to undermine their capacity to survive, succeed and reproduce in this society. In other words, people may have evolved to protect the legitimacy of blame, as a means of facilitating the rehabilitation and incapacitation of offenders (Crockett et al., 2014). In turn, this provides a theoretical framework for explaining the limits on the impact of neurocriminology on attributions of blame (as observed in the

empirical research conducted for this thesis). Blame may simply be too useful a concept to give up.

This theory that people pursue retributivism for consequentialist reasons may explain their resistance to the mitigating implications of neurocriminology. However, it did not lead the current participants to take the opportunity to censor neurocriminology from offenders (any more than from the general public). In the museum study, for example, participants still made the same (small) mitigating inferences from neurocriminology when they knew their sentencing decisions would be specifically communicated to offenders. In the postcard study, people were as willing to communicate neurocriminology to offenders as to students. Therefore, to the extent that people do pursue positive retributivism for consequentialist reasons, this pursuit did not lead the current participants to protect the legitimacy of blame in the eye of offenders any more than in the public eye. The only exception to this finding was observed in the postcard study: in areas of extreme SES, participants were more reluctant to communicate neurocriminology to offenders than to students – a finding that warrants further research.

v. Future research

This thesis examined the implications of communicating neurocriminology to lay people in general. Future research should consider how those implications vary for different types of people along dimensions that are relevant to CJ, such as people who have recently offended or recently been victimised. As the direct recipients of wrongdoing, victims may be more likely to blame the offender in proportion to their moral outrage rather than any scientific argument. There is also reason to believe that psychopathy, as a trait that varies between individuals, may moderate the effect of

neuroscience on public attitudes towards offenders. People who score higher on a measure of psychopathy may be less likely to seek punishment in response to moral outrage, given their reduced tendency to experience distress as a result of learning about wrongdoings (Blair, 2006); in turn, people who score higher on psychopathy may be more amenable to scientific arguments against blame. It has already been found that more psychopathic individuals do not tailor the sentences they recommend (in mock court cases) to the degree of intentionality displayed by the hypothetical offender (Aharoni et al., 2007). Perhaps, therefore, the sentencing recommendations of people who score higher on psychopathy would be more sensitive to scientific arguments against the existence of an incompatibilist free will. In other words, psychopathy may amplify the effect of learning about neurocriminology on one's attitudes towards offenders.

vi. Final note

More generally, the findings of this thesis suggest that people face a certain difficulty or reluctance in simulating the subjective experiences of other people (Twenge and Foster, 2008). It would be difficult and uncomfortable for the average person to adopt the neurobiological desires of a particular offender in the moments leading up to the crime. An ex-prisoner who participated in the museum study noted this astutely:

> *[The quote originally presented here cannot be made freely available via ORA for reasons regarding the remit of the museum study's ethical approval. The quote was sourced from personal communication over email with a participant in the museum study]*.

It is possible to teach people about the neurobiologically driven subjective experience outside the law-abiding life; indeed, the current research invited participants

to simulate that neurobiological experience. Yet their '*[part of the omitted quote]*' was negligible in comparison to their drive to censor those who have violated group values and thereby threatened their sense of belonging (Tyler and Boeckmann, 1997). Hence this thesis strongly questions Greene and Cohen's (2004, p.1784) claim that as a result of neuroscience, 'retributivist notions of criminal responsibility [may] give way to consequentialist ones, thus radically transforming our approach to criminal justice'. I conclude that the implications of communicating neurocriminology will be minimal, not 'radical', giving rise to a slight improvement in attitudes towards offenders.

While this thesis has documented the resistance of lay people to neuroscientific grounds for overthrowing retributive justice, future research may wish to consider *how* people will continue to reject the mitigating implications of neurocriminology. The future will bring major developments in brain scanning technology, as well as new ways of communicating philosophical ideas and neuroscientific findings to the public. These developments are likely to make people increasingly aware of the current (compatibilist) basis for their attributions of moral responsibility, such that it becomes more difficult for people to misunderstand what it means to consider blame compatible with determinism. Consequently, the way in which people resist neurocriminology may shift, away from asserting compatibilist justifications for blame and towards questioning the validity, ethics and communication of neurocriminology.

If greater scepticism is cast on compatibilist criteria for moral responsibility, people may seek alternative means of resisting the mitigating implications of neurocriminology. Conducting research in the field may become more difficult, with increasing concern about the ethics of neurocriminology, especially regarding its use in prediction and early intervention (Raine, 2013), perhaps leading to restrictions on the use of brain scanning with forensic populations. People may also resist attempts to

communicate the science on a mass scale, for example, through mainstream educational curricula. This resistance is clear in criminology as a field, with claims that neuroscience ‘has been systematically excluded from the discipline’ (Wright et al., 2008, p.326). Indeed, between 1996 and 2006, references to biology featured in less than 4% of *Criminology* papers and in the abstracts of only two criminology PhDs (Wright et al.). Hence a key contribution of this thesis has been to contest this trend, highlighting the need to consider neuroscience in the criminological study of public opinion regarding crime and punishment.

Acknowledgements

This theatre study was made possible by the National Theatre and Company Three. I would like to thank the Executive Director of Company Three, Adam Coleman, and Stephen Hayes, Head of House Management at the National Theatre, for recognising the value of this research. The museum study was made possible by the Live Science programme at the Science Museum in London. I would like to thank Lottie Dodwell for making the practical arrangements necessary to conduct the study at the museum. Finally, thank you to my supervisor, Professor Ben Bradford, for embracing this new and controversial subfield of criminology.

References

Aharoni, E. and Fridlund, A., 2012. Punishment without reason: Isolating retribution in lay punishment of criminal offenders. *Psychology, Public Policy, and Law*, 18(4), pp.599–625.

Aharoni, E., Funk, C., Sinnott-Armstrong, W. and Gazzaniga, M., 2008. Can neurological evidence help courts assess criminal responsibility? Lessons from law and neuroscience. *Annals of the New York Academy of Sciences*, 1124(1), pp.145–160.

Aharoni, E., Weintraub, L.L. and Fridlund, A.J., 2007. No skin off my back: Retribution deficits in psychopathic motives for punishment. *Behavioral Sciences & the Law*, 25(6), pp.869–889.

Alicke, M., 1990. Incapacitating conditions and alteration of blame. *Journal of Social Behavior and Personality*, 5(6), pp.651–664.

Alicke, M.D., 1992. Culpable causation. *Journal of Personality and Social Psychology*, 63(3), pp.368–378.

Alicke, M.D. and Davis, T.L., 1990. Capacity responsibility in social evaluation. *Personality and Social Psychology Bulletin*, 16(3), pp.465–474.

Allen, T.T., Trzcinski, E. and Kubiak, S.P., 2012. Public Attitudes Toward Juveniles Who Commit Crimes. *Crime & Delinquency*, 58(1), pp.78–102.

Alquist, J.L., Ainsworth, S.E. and Baumeister, R.F., 2013. Determined to conform: Disbelief in free will increases conformity. *Journal of Experimental Social Psychology*, 49(1), pp.80–86.

Angermeyer, M., Holzinger, A. and Carta, M., 2011. Biogenetic explanations and public acceptance of mental illness: Systematic review of population studies. *The British Journal of Psychiatry*, 199(5), pp.367–372.

Aniskiewicz, A.S., 1979. Autonomic components of vicarious conditioning and psychopathy. *Journal of Clinical Psychology*, 35(1), pp.60–67.

Appelbaum, P.S. and Scurich, N., 2014. Impact of behavioral genetic evidence on the adjudication of criminal behavior. *Journal of the American Academy of Psychiatry and Law*, 42(1), pp.91–100.

Appelbaum, P.S., Scurich, N. and Raad, R., 2015. Effects of behavioral genetic evidence on perceptions of criminal responsibility and appropriate punishment. *Psychology, Public Policy, and Law*, 21(2), pp.134–144.

Aspinwall, L.G., Brown, T.R. and Tabery, J., 2012. The double-edged sword: Does biomechanism increase or decrease judges' sentencing of psychopaths? *Science*, 337(6096), pp.846–849.

Bal, M. and van den Bos, K., 2012. Blaming for a better future: Future orientation and associated intolerance of personal uncertainty lead to harsher reactions toward innocent victims. *Personality & Social Psychology Bulletin*, 38(7), pp.835–844.

Barnes, J.C., TenEyck, M., Boutwell, B.B., Beaver, K.M., Bandura, A.J., Benson, M.L., Greer, L.F., DeMaris, A., Wyk, J.V., Berger, S.L., Kouzarides, T., Shiekhattar, R., Shilatifard, A., Black, M.C., Basile, K.C., Breiding, M.J., Smith, S.G., Walters, M.L.,

Merrick, M.T., al., et, Bouchard, T.J., Lykken, D.T., McGue, M., Segal, N.L., Tellegen, A., Campbell, J.C., Charney, E., English, W., Chio, J., Jeong, B., Polcari, A., Rohan, M.L., Teicher, M.H., Cleveland, H.H., Beekman, C., Zheng, Y., Coker, A.L., Davis, K.E., Arias, I., Desai, S., Sanderson, M., Brandt, H.M., al., et, Edleson, J.L., Egeland, B., Franklin, T.B., Russig, H., Weiss, I.C., Gräff, J., Linder, N., Michalon, A., al., et, Harris, J.R., Harris, K.M., Harris, K.M., Halpern, C.T., Whitsel, E., Hussey, J., Tabor, J., Entzel, P., al., et, Hines, D.A., Saudino, K.J., Hines, D.A., Saudino, K.J., Hines, D.A., Saudino, K.J., Kalmuss, D., Lussier, P., Farrington, D.P., Moffitt, T.E., Neale, M.C., Maes, H.H.M., Pinker, S., Pinto, L.A., Sullivan, E.L., Rosenbaum, A., Wyngarden, N., Umhau, J.C., Miller, M.W., al., et, Plomin, R., DeFries, J.C., McClearn, G.E., McGuffin, P., Purcell, S., Raine, A., Roberts, A.L., Gilman, S.E., Fitzmaurice, G., Decker, M.R., Koenen, K.C., Silverman, J.G., Raj, A., Mucci, L.A., Hathaway, J.E., Stith, S.M., Rosen, K.H., Middleton, K.A., Busch, A.L., Lundeberg, K., Carlton, R.P., Straus, M.A., Straus, M.A., Hamby, S.L., Boney-McCoy, S., Sugarman, D.B., Consortium, T.E.P., Truman, J.L., Turkheimer, E. and Widom, C., 2013. Indicators of domestic/intimate partner violence are structured by genetic and nonshared environmental influences. *Journal of Psychiatric Research*, 47(3), pp.371–6.

Baron, J. and Ritov, I., 1994. Reference points and omission bias. *Organizational Behavior and Human Decision Processes*, 59(3), pp.475–498.

Bartels, D.M. and Pizarro, D.A., 2011. The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition*, 121(1), pp.154–161.

Barth, A., 2007. A double-edged sword: The role of neuroimaging in federal capital sentencing. *American Journal of Law & Medicine*, 33(2–3), pp.501–522.

Baumeister, R., DeWall, C. and Ciarocco, N., 2005. Social exclusion impairs self-regulation. *Journal of Personality and Social Psychology*, 88(4), pp.589–604.

Baumeister, R. and Leary, M., 1995. The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, 117(3), pp.497–529.

Baumeister, R.F., 2008. Free will in scientific psychology. *Perspectives on Psychological Science*, 3(1), pp.14–19.

Baumeister, R.F., Masicampo, E.J. and Dewall, C.N., 2009. Prosocial benefits of feeling free: Disbelief in free will increases aggression and reduces helpfulness. *Personality & Social Psychology Bulletin*, 35(2), pp.260–268.

Beall, E., Meindl, P., Iyer, R., Monterosso, J., Graham, J. and Schwarz, B., 2013. Biology or bad character? Brain scan images increase determinism and decrease moral attributions for criminal acts. In: *Paper presented at the Annual Conference of the Society for Personality and Social Psychology*. New Orleans, L.A.

Beaver, K.M., Barnes, J.C. and Boutwell, B., 2015. Introduction. In: *The Nurture vs. Biosocial Debate in Criminology*. Thousand Oaks, CA: SAGE Publications, pp.1–8.

Beaver, K.M., Wright, J.P., DeLisi, M. and Vaughn, M.G., 2008. Genetic influences on the stability of low self-control: Results from a longitudinal sample of twins. *Journal of Criminal Justice*, 36(6), pp.478–485.

Bègue, L. and Bastounis, M., 2003. Two spheres of belief in justice: Extensive support

for the bidimensional model of belief in a just world. *Journal of Personality*, 71(3), pp.435–463.

Bennett, L., Thirlaway, K. and Murray, A.J., 2008. The stigmatising implications of presenting schizophrenia as a genetic disease. *Journal of Genetic Counseling*, 17(6), pp.550–559.

Bergner, R.M. and Ramon, A., 2013. Some implications of beliefs in altruism, free will, and nonreductionism. *The Journal of Social Psychology*, 153(5), pp.598–618.

Björnsson, G., 2014. Incompatibilism and "Bypassed" Agency. In: A.R. Mele, ed., *Surrounding Free Will: Philosophy, Psychology, Neuroscience*. New York, NY: Oxford University Press, pp.95–118.

Björnsson, G., 2016. Outsourcing the deep self: Deep self discordance does not explain away intuitions in manipulation arguments. *Philosophical Psychology*, 29(5), pp.637–653.

Björnsson, G. and Persson, K., 2013. A unified empirical account of responsibility judgments. *Philosophy and Phenomenological*, 87(3), pp.611–639.

Blair, R., 2006. The emergence of psychopathy: Implications for the neuropsychological approach to developmental disorders. *Cognition*, 101(2), pp.414–442.

Blakemore, S.-J. and Choudhury, S., 2006. Development of the Adolescent Brain: Implications for Executive Function and Social Cognition. *Journal of Child Psychology and Psychiatry*, 47(3–4), pp.296–312.

Blanken, I., van de Ven, N. and Zeelenberg, M., 2015. A meta-analytic review of moral licensing. *Personality and Social Psychology Bulletin*, 41(4), pp.540–558.

Bloom, P., 2004. *Descartes' Baby: How The Science of Child Development Explains What Makes Us Human*. New York, NY: Basic Books.

Bode, S., He, A., Soon, C., Trampel, R. and Turner, R., 2011. Tracking the unconscious generation of free decisions using uitra-high field fMRI. *PLoS ONE*, 6(6), p.e21612.

Boutwell, B.B. and Beaver, K.M., 2010. The role of broken homes in the development of self-control: A propensity score matching approach. *Journal of Criminal Justice*, 38(4), pp.489–495.

Boutwell, B.B. and Lewis, R., 2015. Some kind of madness: The biosocial origins of intimate partner violence. In: K.M. Beaver, J.C. Barnes and B.B. Boutwell, eds., *The Nurture vs. Biosocial Debate in Criminology*. Thousand Oaks, CA: SAGE Publications.

Boyd, R., Gintis, H., Bowles, S. and Richerson, P.J., 2003. The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences of the United States of America*, 100(6), pp.3531–5.

Braithwaite, J., 1989. *Crime, Shame and Reintegration*. Cambridge: Cambridge University Press.

Brass, M. and Haggard, P., 2007. To do or not to do: The neural signature of self-control. *Journal of Neuroscience*, 27(34), pp.9141–9145.

De Brigard, F., Mandelbaum, E. and Ripley, D., 2009. Responsibility and the Brain

Sciences. *Ethical Theory and Moral Practice*, 12(5), pp.511–524.

Buchen, L., 2012. Science in court: Arrested development. *Nature*, 484(7394), pp.304–6.

Burns, J.M. and Swerdlow, R.H., 2003. Right orbitofrontal tumor with pedophilia symptom and constructional apraxia sign. *Archives of Neurology*, 60(3), p.437.

Cahill, N. and Sherrets, S., 1979. The lost-letter technique: A measure of social and political attitudes in three socio-economic groups. *Psychological Reports*, 45(1), pp.144–146.

Calhoun, L.G. and Cann, A., 2010. Differences in assumptions about a just world: Ethnicity and point of view. *The Journal of Social Psychology*, 134, pp.765–770.

Callahan, L.A., Steadman, H.J., McGreevy, M.A. and Robbins, P.C., 1991. The volume and characteristics of insanity defense pleas: An eight-state study. *Journal of the American Academy of Psychiatry and the Law*, 19(4), pp.331–338.

Callan, M.J., Shead, N.W. and Olson, J.M., 2009. Foregoing the labor for the fruits: The effect of just world threat on the desire for immediate monetary rewards. *Journal of Experimental Social Psychology*, 45(1), pp.246–249.

Campbell, A. and Muncer, S., 1990. Causes of crime. *Criminal Justice and Behavior*, 17, pp.410–419.

Capestany, B. and Harris, L., 2014. Disgust and biological descriptions bias logical reasoning during legal decision-making. *Social Neuroscience*, 9(3), pp.265–277.

Carlsmith, K., Darley, J.M. and Robinson, P.H., 2002. Why do we punish?: Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology*, 83(2), pp.284–299.

Carlsmith, K.M., 2006. The roles of retribution and utility in determining punishment. *Journal of Experimental Social Psychology*, 42(4), pp.437–451.

Carlsmith, K.M., 2008. On justifying punishment: The discrepancy between words and actions. *Social Justice Research*, 21(2), pp.119–137.

Carlsmith, K.M. and Darley, J.M., 2008. Psychological aspects of retributive justice. In: *Advances in Experimental Social Psychology*. Academic Press Inc, pp.193–236.

Carlsmith, K.M., Monahan, J. and Evans, A., 2007. The function of punishment in the 'civil' commitment of sexually violent predators. *Behavioral Sciences & the Law*, 25(4), pp.437–448.

Carlsmith, K.M. and Sood, A.M., 2009. The fine line between interrogation and retribution. *Journal of Experimental Social Psychology*, 45(1), pp.191–196.

Carlsmith, K.M., Wilson, T.D. and Gilbert, D.T., 2008. The paradoxical consequences of revenge. *Journal of Personality and Social Psychology*, 95(6), pp.1316–1324.

Carroll, J., Perkowitz, W., Lurigio, A.J. and Weaver, F.M., 1987. Sentencing goals, causal attributions, ideology, and personality. *Journal of Personality and Social Psychology*, 52(1), pp.107–118.

Caspar, E.A., Vuillaume, L., Magalhães De Saldanha da Gama, P.A. and Cleeremans,

A., 2017. The influence of (dis)belief in free will on immoral behavior. *Frontiers in Psychology*, 8, p.20.

Caspi, A., McClay, J., Moffitt, T.E., Mill, J., Martin, J., Craig, I.W., Taylor, A. and Poulton, R., 2002. Role of genotype in the cycle of violence in maltreated children. *Science*, 297(5582), pp.851–854.

Chen, G. and Einat, T., 2015. To punish or not to punish - that is the question: Attitudes of criminology and criminal justice students in Israel toward punishment. *International Journal of Offender Therapy and Comparative Criminology*, 61(3), pp.347–367.

Cherulnik, P.D., 1975. An independent validation of the lost-letter technique. *The Journal of Social Psychology*, 96(2), pp.299–300.

Cheung, B.Y. and Heine, S.J., 2015. The double-edged sword of genetic accounts of criminality. *Personality and Social Psychology Bulletin*, 41(12), pp.1723–1738.

Cima, M., Tonnaer, F. and Hauser, M.D., 2010. Psychopaths know right from wrong but don't care. *Social Cognitive and Affective Neuroscience*, 5(1), pp.59–67.

Clark, C.J., Luguri, J.B., Ditto, P.H., Knobe, J., Shariff, A.F. and Baumeister, R.F., 2014. Free to punish: A motivated account of free will belief. *Journal of Personality and Social Psychology*, 106(4), pp.501–513.

Cochrane, R.E., Grisso, T. and Frederick, R.I., 2001. The relationship between criminal charges, diagnoses, and psycholegal opinions among federal pretrial defendants. *Behavioral Sciences & the Law*, 19(4), pp.565–582.

Cohen, L.E. and Felson, M., 1979. Social change and crime rate trends: A routine activity approach. *American Sociological Review*, 44(4), pp.588–608.

Connors, J. and Heaven, P.C.L., 1990. Belief in a just world and attitudes toward AIDS sufferers. *The Journal of Social Psychology*, 130(4), pp.559–560.

Conway, P. and Gawronski, B., 2013. Deontological and utilitarian inclinations in moral decision making: A process dissociation approach. *Journal of Personality and Social Psychology*, 104(2), pp.216–235.

Crick, F., 1994. *The Astonishing Hypothesis*. New York, NY: Scribner's.

Critcher, C.R., Inbar, Y. and Pizarro, D.A., 2013. How quick decisions illuminate moral character. *Social Psychological and Personality Science*, 4(3), pp.308–315.

Crockett, M., Özdemir, Y. and Fehr, E., 2014. The value of vengeance and the demand for deterrence. *Journal of Experimental Psychology: General*, 143(6), pp.2279–2286.

Dalbert, C. and Filke, E., 2007. Belief in a personal just world, justice judgments, and their functions for prisoners. *Criminal Justice and Behavior*, 34(11), pp.1516–1527.

Dar-Nimrod, I. and Heine, S.J., 2011. Genetic essentialism: On the deceptive determinism of DNA. *Psychological Bulletin*, 137(5), pp.800–818.

Dar-Nimrod, I., Heine, S.J., Cheung, B.Y. and Schaller, M., 2011. Do scientific theories affect men's evaluations of sex crimes? *Aggressive Behavior*, 37(5), pp.440–449.

Darley, J., Carlsmith, K. and Robinson, P., 2000. Incapacitation and just deserts as motives for punishment. *Law and Human Behavior*, 24(6), pp.659–683.

Deacon, B.J. and Baird, G.L., 2009. The chemical imbalance explanation of depression: Reducing blame at what cost? *Journal of Social and Clinical Psychology*, 28(4), pp.415–435.

Delmage, E., 2013. The minimum age of criminal responsibility: A medico-legal perspective. *Youth Justice*, 13(2), pp.102–110.

Denno, D.W., 2015. The myth of the double-edged sword: An empirical study of neuroscience evidence in criminal cases. *Boston College Law Review*, 56(2), pp.493–551.

Dietrich, S. and Matschinger, H., 2006. The relationship between biogenetic causal explanations and social distance toward people with mental disorders: Results from a population survey in Germany. *International Journal of Social Psychiatry*, 52(2), pp.166–174.

Duff, R.A., 2001. *Punishment, Communication, and Community*. New York, NY: Oxford University Press.

Dutton, K., 2013. *The Wisdom of Psychopaths*. London, UK: Arrow.

Dweck, C.S., 2006. *Mindset: The New Psychology of Success*. New York, NY: Random House.

Erskine, H., 1974. The polls: Causes of crime. *The Public Opinion Quarterly*, 38(2), pp.288–298.

Everett, J.A.C., Clark, C.J., Luguri, J.B., Earp, B.D., Ditto, P.H. and Shariff, A., n.d. Political differences in free will belief are driven by differences in moralization. *Unpublished manuscript*.

Falco, D.L. and Martin, J.S., 2012. Examining punitiveness: Assessing views toward the punishment of offenders among criminology and non-criminology students. *Journal of Criminal Justice Education*, 23(2), pp.205–232.

Farah, M.J. and Hook, C.J., 2013. The seductive allure of 'seductive allure'. *Perspectives on Psychological Science*, 8(1), pp.88–90.

Farmer, E., 2011. The age of criminal responsibility: Developmental science and human rights perspectives. *Journal of Children's Services*, 6(2), pp.86–95.

Feather, N. and Souter, J., 2002. Reactions to mandatory sentences in relation to the ethnic identity and criminal history of the offender. *Law and Human Behavior*, 26(4), pp.417–438.

Fehr, E. and Gächter, S., 2002. Altruistic punishment in humans. *Nature*, 415, pp.137–140.

Feld, B., 2013. Adolescent criminal responsibility, proportionality, and sentencing policy: Roper, Graham, Miller/Jackson, and the youth discount. *Law & Inequality*, 31, pp.263–330.

FeldmanHall, O., Mobbs, D., Evans, D., Hiscox, L., Navrady, L. and Dalgleish, T., 2012. What we say and what we do: The relationship between real and hypothetical moral choices. *Cognition*, 123(3), pp.434–441.

Ferguson, C.J., 2010. Genetic contributions to antisocial personality and behavior: A

meta-analytic review from an evolutionary perspective. *The Journal of Social Psychology*, 150(2), pp.160–180.

Fernandez-Duque, D., Evans, J. and Christian, C., 2015. Superfluous neuroscience information makes explanations of psychological phenomena more appealing. *Journal of Cognitive Neuroscience*, 27(5), pp.926–944.

Fernandez-Duque, D. and Schwartz, B., 2016. Common sense beliefs about the central self, moral character, and the brain. *Frontiers in Psychology*, 6, p.2007.

Festinger, L., 1957. *A Theory of Cognitive Dissonance*. Stanford, CA: Stanford University Press.

Fincher, K.M. and Tetlock, P.E., 2015. Brutality under cover of ambiguity: Activating, perpetuating, and deactivating covert retributivism. *Personality & Social Psychology Bulletin*, 41(5), pp.629–642.

Flanagan, T.J., 1987. Change and influence in popular criminology: Public attributions of crime causation. *Journal of Criminal Justice*, 15(3), pp.231–243.

Fletcher, G.P., 1998. *Basic Concepts of Criminal Law*. New York, NY: Oxford University Press.

Foot, P., 1978. The Problem of Abortion and the Doctrine of the Double Effect. In: *Virtues and Vices*. Oxford, UK: Basil Blackwell.

Forstmann, M. and Burgmer, P., 2015. Adults are intuitive mind-body dualists. *Journal of Experimental Psychology:*, 144(1), pp.222–235.

Frankfurt, H.G., 1969. Alternate possibilities and moral responsibility. *The Journal of Philosophy*, 66(23), p.829.

Frankfurt, H.G., 1971. Freedom of the will and the concept of a person. *The Journal of Philosophy*, 68(1), p.5.

Freiman, C. and Nichols, S., 2011. Is desert in the details? *Philosophy and Phenomenological Research*, 82(1), pp.121–133.

Furnham, A. and Gunter, B., 1984. Just world beliefs and attitudes towards the poor. *British Journal of Social Psychology*, 23(3), pp.265–269.

Furnham, A. and Henderson, M., 1983. Lay theories of delinquency. *European Journal of Social Psychology*, 13(2), pp.107–120.

Fuss, J., Dressing, H. and Briken, P., 2015. Neurogenetic evidence in the courtroom: A randomised controlled trial with German judges. *Journal of Medical Genetics*, 52(11), pp.730–7.

Gabbidon, S.L. and Boisvert, D., 2012. Public opinion on crime causation: An exploratory study of Philadelphia area residents. *Journal of Criminal Justice*, 40(1), pp.50–59.

Gajos, J.M., Beaver, K.M., Gertz, M. and Bratton, J., 2014. Public opinion of genetic and neuropsychological contributors to criminal involvement. *Journal of Criminal Justice Education*, 25(3), pp.368–385.

Galvan, A., Hare, T.A., Parra, C.E., Penn, J., Voss, H., Glover, G. and Casey, B.J.,

2006. Earlier development of the accumbens relative to orbitofrontal cortex might underlie risk-taking behavior in adolescents. *The Journal of Neuroscience*, 26(25), pp.6885–92.

Garcia, D. and Sikström, S., 2014. The dark side of Facebook: Semantic representations of status updates predict the Dark Triad of personality. *Personality and Individual Differences*, 67, pp.92–96.

Garland, D., 2001. *The Culture of Control: Crime and Social Order in Contemporary Society*. Chicago: University of Chicago Press.

Genschow, O., Rigoni, D. and Brass, M., 2017. Belief in free will affects causal attributions when judging others' behavior. *Proceedings of the National Academy of Sciences of the United States of America*, 114(38), pp.10071–10076.

Giacopassi, D.J. and Blankenship, M.B., 1991. The effects of criminal justice pedagogy on student attitudes. *American Journal of Criminal Justice*, 16(1), pp.97–103.

Gill, M. and Cerce, S., 2017. He never willed to have the will he has: Historicist narratives,"civilized" blame, and the need to distinguish two notions of free will. *Journal of Personality and Social Psychology*, 112(3), pp.361–382.

Gill, M.J. and Getty, P.D., 2016. On shifting the blame to humanity: Historicist narratives regarding transgressors evoke compassion for the transgressor but disdain for humanity. *British Journal of Social Psychology*, 55(4), pp.773–791.

Gleichgerrcht, E., Young, L., Carpenter, J., Galinsky, A. and Schweizer, S., 2013. Low levels of empathic concern predict utilitarian moral judgment. *PLoS ONE*, 8(4), p.e60418.

Glenn, A.L. and Raine, A., 2014. Neurocriminology: Implications for the punishment, prediction and prevention of criminal behaviour. *Nature Reviews Neuroscience*, 15(1), pp.54–63.

Goldson, B., 2013. 'Unsafe, unjust and harmful to wider society': Grounds for raising the minimum age of criminal responsibility in England and Wales. *Youth Justice*, 13(2), pp.111–130.

Goldstein, A.M., Morse, S.J. and Shapiro, D.L., 2003. Evaluation of criminal responsibility. In: A.M. Goldstein, ed., *Forensic Psychology*. New York, NY: Wiley, pp.381–406.

Gollwitzer, M. and Bücklein, K., 2007. Are "we" more punitive than "me"? Self-construal styles, justice-related attitudes, and punitive judgments. *Social Justice Research*, 20(4), pp.457–478.

Gollwitzer, M. and Bushman, B.J., 2011. Do victims of injustice punish to improve their mood? *Social Psychological and Personality Science*, 3(5), pp.572–580.

Gollwitzer, M. and Denzler, M., 2009. What makes revenge sweet: Seeing the offender suffer or delivering a message? *Journal of Experimental Social Psychology*, 45(4), pp.840–44.

Gollwitzer, M. and Keller, L., 2010. What you did only matters if you are one of us: Offenders' group membership moderates the effect of criminal history on punishment severity. *Social Psychology*, 41, pp.20–26.

Gollwitzer, M., Meder, M. and Schmitt, M., 2011. What gives victims satisfaction when they seek revenge? *European Journal of Social Psychology*, 41(3), pp.364–74.

Gorin, M., 2013. What makes an intuition a compatibilist intuition? A response to Sripada. *Philosophia*, 41(4), pp.1205–1215.

Gottfredson, M. and Hirschi, T., 1990. *A General Theory of Crime.* Palo Alto, CA: Stanford University Press.

Graham v. Florida, 560 U.S. 48 [2010].

Gray, H., Gray, K. and Wegner, D., 2007. Dimensions of mind perception. *Science*, 315(5812), p.619.

Greene, E. and Cahill, B.S., 2012. Effects of neuroimaging evidence on mock juror decision making. *Behavioral Sciences & the Law*, 30(3), pp.280–296.

Greene, E. and Dodge, M., 1995. The influence of prior record evidence on juror decision making. *Law and Human Behavior*, 19(1), pp.67–78.

Greene, J. and Cohen, J., 2004. For the law, neuroscience changes nothing and everything. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 359(1451), pp.1775–1785.

Greene, J., Sommerville, R. and Nystrom, L., 2001. An fMRI investigation of emotional engagement in moral judgment. *Science*, 293(5537), pp.2105–2108.

Greene, J.D., Morelli, S.A., Lowenberg, K., Nystrom, L.E. and Cohen, J.D., 2008. Cognitive load selectively interferes with utilitarian moral judgment. *Cognition*, 107(3), pp.1144–1154.

Gromet, D.M., Goodwin, G.P., Tang, S., Nadelhoffer, T. and Sinnott-Armstrong, W.P., n.d. Mind, brain, and character: How neuroscience affects people's views of wrongdoers. *Unpublished manuscript*.

Gruber, D. and Dickerson, J.A., 2012. Persuasive images in popular science: Testing judgments of scientific reasoning and credibility. *Public Understanding of Science*, 21(8), pp.938–948.

Guala, F., 2012. Reciprocity: Weak or strong? What punishment experiments do (and do not) demonstrate. *Behavioral and Brain Sciences*, 35(1), pp.1–15.

Gurley, J.R. and Marcus, D.K., 2008. The effects of neuroimaging and brain injury on insanity defenses. *Behavioral Sciences & the Law*, 26(1), pp.85–97.

Hafer, C.L., 2000. Investment in long-term goals and commitment to just means drive the need to believe in a just world. *Personality and Social Psychology Bulletin*, 26(9), pp.1059–1073.

Hafer, C.L., Bègue, L., Choma, B.L. and Dempsey, J.L., 2005. Belief in a just world and commitment to long-term deserved outcomes. *Social Justice Research*, 18(4), pp.429–444.

Haidt, J., 2001. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108(4), pp.814–834.

Haidt, J., Bjorklund, F. and Murphy, S., 2000. *Moral dumbfounding: When intuition*

finds no reason. Unpublished manuscript.

Haji, I. and Cuypers, S.E., 2004. Moral responsibility and the problem of manipulation reconsidered. *International Journal of Philosophical Studies*, 12(4), pp.439–464.

Harden, K. and Tucker-Drob, E., 2011. Individual differences in the development of sensation seeking and impulsivity during adolescence: Further evidence for a dual systems model. *Developmental Psychology*, 47(3), pp.739–46.

Hare, R.D., 1991. *The Hare Psychopathy Checklist-Revised: Manual*. Toronto, ON: Multi-Health Systems.

Hare, R.D., 1993. *Without Conscience: The Disturbing World of the Psychopaths Among Us*. New York, NY: Guildford Publications.

Harris, S., 2012. *Free Will*. New York, NY: Free Press.

Haslam, N., 2011. Genetic essentialism, neuroessentialism, and stigma: commentary on Dar-Nimrod and Heine (2011). *Psychological Bulletin*, 137(5), pp.819–824.

Hayes, A.F., 2013. *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach*. Guildford, UK: Guildford Press.

Heider, F., 1958. *The Psychology of Interpersonal Relations*. New York, NY: Wiley.

Hetey, R.C. and Eberhardt, J.L., 2014. Racial disparities in incarceration increase acceptance of punitive policies. *Psychological Science*, 25(10), pp.1949–1954.

Hillman, J. and Martin, R.A., 2002. Lessons about gay and lesbian lives: A spaceship exercise. *Teaching of Psychology*, 29(4), pp.308–311.

Hines, D. and Saudino, K., 2004. Genetic and environmental influences on intimate partner aggression: A preliminary study. *Violence and Victims*, 19(6), pp.701–718.

von Hirsch, A., 1976. *Doing Justice: The Choice of Punishments*. New York, NY: Hill and Wang.

von Hirsch, A., 2001. Proportionate sentences for juveniles: How different than for adults? *Punishment & Society*, 3(2), pp.221–36.

von Hirsch, A., 2009. The recidivist premium: For and against. In: A. von Hirsch and A. Ashworth, eds., *Principled sentencing: Readings on Theory and Practice*, 3rd ed. Oxford, UK: Hart Publishing, pp.148–162.

Hirschi, T., 1969. A control theory of delinquency. In: *Causes of Delinquency*. Oakland, California: University of California Press, pp.16–34.

Hirschi, T. and Gottfredson, M., 1983. Age and the Explanation of Crime. *American Journal of Sociology*, 89(3), pp.552–84.

Hodson, G., Choma, B.L. and Costello, K., 2009. *Experiencing Alien-Nation: Effects of a simulation intervention on attitudes toward homosexuals*. *Journal of Experimental Social Psychology*, .

Holland, J., Silva, A.S., Mace, R., Wikström, P. and Loeber, R., 2012. Lost letter measure of variation in altruistic behaviour in 20 neighbourhoods. *PLoS ONE*, 7(8), p.e43294.

Hong, Y., Chiu, C., Dweck, C. and Lin, D., 1999. Implicit theories, attributions, and coping: A meaning system approach. *Journal of Personality and Social Psychology*, 77(3), pp.588–599.

Hook, C. and Farah, M., 2013. Look again: Effects of brain images and mind–brain dualism on lay evaluations of research. *Journal of Cognitive Neuroscience*, 25(9), pp.1397–1405.

Hopkins, E.J., Weisberg, D.S. and Taylor, J.C.V., 2016. The seductive allure is a reductive allure: People prefer scientific explanations that contain logically irrelevant reductive information. *Cognition*, 155(3), pp.67–76.

Horstkötter, D., Berghmans, R., Feron, F. and De Wert, G., 2014. 'One can always say no.' Enriching the bioethical debate on antisocial behaviour, neurobiology and prevention: Views of juvenile delinquents. *Bioethics*, 28(5), pp.225–234.

Horstkötter, D., Berghmans, R. and Ruiter, C. De, 2012. "We are also normal humans, you know?" Views and attitudes of juvenile delinquents on antisocial behavior, neurobiology and prevention. *International Journal of Law and Psychiatry*, 34(5), pp.289–297.

Hough, M. and Roberts, J. V., 2004. *Youth Crime and Youth Justice: Public Opinion in England and Wales*. Bristol, UK: The Policy Press.

Hunt, M., 2000. Status, religion, and the 'belief in a just world': Comparing African Americans, Latinos, and Whites. *Social Science Quarterly*, 81(1), pp.325–343.

Inbar, Y., Pizarro, D.A. and Cushman, F., 2012. Benefiting from misfortune: When harmless actions are judged to be morally blameworthy. *Personality and Social Psychology Bulletin*, 38(1), pp.52–62.

Jensen, F.E. and Nutt, A.E., 2016. *The Teenage Brain: A Neuroscientist's Survival Guide to Raising Adolescents and Young Adults*. New York, NY: Harper Collins.

Johnson, M.P., 2005. Domestic violence: It's not about gender - or is it? *Journal of Marriage and Family*, 67(5), pp.1126–1130.

Jones, D.N. and Paulhus, D.L., 2014. Introducing the Short Dark Triad (SD3). *Assessment*, 21(1), pp.28–41.

Jones, O.D., Wagner, A.D., Faigman, D.L. and Raichle, M.E., 2013. Neuroscientists in court. *Nature Reviews Neuroscience*, 14(10), pp.730–736.

Jost, J.T. and Banaji, M.R., 1994. The role of stereotyping in system-justification and the production of false consciousness. *British Journal of Social Psychology*, 33(1), pp.1–27.

Jost, J.T., Banaji, M.R. and Nosek, B.A., 2004. A decade of system justification theory: Accumulated evidence of conscious and unconscious bolstering of the status quo. *Political Psychology*, 25(6), pp.881–919.

Jost, J.T., Pelham, B.W., Sheldon, O. and Ni Sullivan, B., 2003. Social inequality and the reduction of ideological dissonance on behalf of the system: Evidence of enhanced system justification among the disadvantaged. *European Journal of Social Psychology*, 33(1), pp.13–36.

Kane, R., 1999. Responsibility, luck, and chance: Reflections on free will and indeterminism. *The Journal of Philosophy*, 96(5), pp.217–240.

Kansas Statute, ., 1994. *Sexually Violent Predator Act*.

Kant, I., 1797. *The Metaphysics of Morals*. New York, NY: Cambridge University Press.

Karstedt, S. and Farrall, S., 2006. The moral economy of everyday crime: Markets, consumers and citizens. *British Journal of Criminology*, 46(6), pp.1011–1036.

Keller, L.B., Oswald, M.E., Stucki, I. and Gollwitzer, M., 2010. A closer look at an eye for an eye: Laypersons' punishment decisions are primarily driven by retributive motives. *Social Justice Research*, 23(2–3), pp.99–116.

Kelley, H.H., 1973. The processes of causal attribution. *American Psychologist*, 28(2), pp.107–128.

Kiehl, K.A., Smith, A.M., Hare, R.D., Mendrek, A., Forster, B.B., Brink, J. and Liddle, P.F., 2001. Limbic abnormalities in affective processing by criminal psychopaths as revealed by functional magnetic resonance imaging. *Biological Psychiatry*, 50(9), pp.677–684.

Kim, N. and LoSavio, S., 2009. Causal explanations affect judgments of the need for psychological treatment. *Judgment and Decision Making*, 4(1), pp.82–91.

Kliemann, D., Young, L., Scholz, J. and Saxe, R., 2008. The influence of prior record on moral judgment. *Neuropsychologia*, 46(12), pp.2949–2957.

Knobe, J., 2003. Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology*, 16(2), pp.309–324.

Knobe, J., 2014. Free Will and the Scientific Vision. In: E. Machery and E. O'Neill, eds., *Current Controversies in Experimental Philosophy*. New York, NY: Routledge, pp.69–85.

Knobe, J. and Prinz, J., 2008. Intuitions about consciousness: Experimental studies. *Phenomenology and the Cognitive Sciences*, 7(1), pp.67–83.

Koenigs, M., Kruepke, M., Zeier, J. and Newman, J.P., 2012. Utilitarian moral judgment in psychopathy. *Social Cognitive and Affective Neuroscience*, 7(6), pp.708–714.

Kolber, A.J., 2016. Free Will as a Matter of Law. In: D. Patterson and M.S. Pardo, eds., *Philosophical Foundations of Law and Neuroscience*. Oxford, UK, pp.10–28.

Kosinski, M., Matz, S., Gosling, S. and Popov, V., 2015. Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist*, 70(6), pp.543–556.

Kosinski, M., Stillwell, D. and Graepel, T., 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America*, 110(15), pp.5802–5.

Kraus, M. and Keltner, D., 2013. Social class rank, essentialism, and punitive judgment. *Journal of Personality and Social Psychology*, 105(2), pp.247–261.

Kraus, M., Piff, P. and Keltner, D., 2009. Social class, sense of control, and social explanation. *Journal of Personality and Social Psychology*, 97(6), pp.992–1004.

Kröber, H.L. and Lau, S., 2000. Bad or Mad? Personality disorders and legal responsibility? The German situation. *Behavioral Sciences & the Law*, 18(5), pp.679–690.

Krueger, F., Hoffman, M., Walter, H. and Grafman, J., 2014. An fMRI investigation of the effects of belief in free will on third-party punishment. *Social Cognitive and Affective Neuroscience*, 9(8), pp.1143–1149.

Lambert, E.G., 2004. Assessing the crime and punishment views of criminal justice majors: How different are they from other majors? *Criminal Justice Studies*, 17(3), pp.245–257.

Laurin, K., Fitzsimons, G. and Kay, A., 2011. Social disadvantage and the self-regulatory function of justice beliefs. *Journal of Personality and Social Psychology*, 100(1), pp.149–171.

Lebowitz, M.S., 2014. Biological conceptualizations of mental disorders among affected individuals: A review of correlates and consequences. *Clinical Psychology: Science and Practice*, 21(1), pp.67–83.

Lerner, M.J., 1980. *The Belief in a Just World: A Fundamental Delusion*. New York, NY: Plenum Press.

Leyens, J.P., Paladino, P.M., Rodriguez-Torres, R., Vaes, J., Demoulin, S., Rodriguez-Perez, A. and Gaunt, R., 2000. The emotional side of prejudice: The attribution of secondary emotions to ingroups and outgroups. *Personality and Social Psychology Review*, 4(2), pp.186–197.

Libet, B., 1985. Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences*, 8(4), pp.529–539.

Locke, D. and Frankfurt, H., 1975. Three concepts of free action. *Supplementary Volumes of the Aristotelian Society*, 49, pp.95–125.

Lombroso, C., 2006. *Criminal Man*. Durham: Duke University Press.

Lynn, M.T., Van Dessel, P. and Brass, M., 2013. The influence of high-level beliefs on self-regulatory engagement: Evidence from thermal pain stimulation. *Frontiers in Psychology*, 4, p.614.

Lynn, M.T., Muhle-Karbe, P.S., Aarts, H. and Brass, M., 2014. Priming determinist beliefs diminishes implicit (but not explicit) components of self-agency. *Frontiers in Psychology*, 5, p.1483.

Mackey, D.A. and Courtright, K.E., 2000. Assessing punitiveness among college students: A comparison of criminal justice majors with other majors∗. *The Justice Professional*, 12(4), pp.423–441.

Mackie, J.L., 1982. Morality and the Retributive Emotions. *Criminal Justice Ethics*, 1, pp.3–10.

Maguire, E.A., Gadian, D.G., Johnsrude, I.S., Good, C.D., Ashburner, J., Frackowiak, R.S. and Frith, C.D., 2000. Navigation-related structural change in the hippocampi of

taxi drivers. *Proceedings of the National Academy of Sciences of the United States of America*, 97(8), pp.4398–4403.

Mahalingam, R., 2007. Essentialism, power, and the representation of social categories: A folk sociology perspective. *Human Development*, 50, pp.300–319.

Malle, B., Guglielmo, S. and Monroe, A., 2014. A theory of blame. *Psychological Inquiry*, 25(2), pp.147–186.

Martin, N.D., Rigoni, D. and Vohs, K.D., 2017. Free will beliefs predict attitudes toward unethical behavior and criminal punishment. *Proceedings of the National Academy of Sciences of the United States of America*, 114(28), pp.7325–7330.

Maruna, S. and King, A., 2009. Once a criminal, always a criminal?: 'Redeemability' and the psychology of punitive public attitudes. *European Journal on Criminal Policy and Research*, 15(1–2), pp.7–24.

Matravers, M., 2016. De-moralising retributivism: Agency, blame and humanity in criminal law theory and practice. In: J. Jacobs and J. Jackson, eds., *The Routledge Handbook of Criminal Justice Ethics*. Routledge, pp.144–163.

McAra, L. and McVie, S., 2007. Youth Justice?: The Impact of System Contact on Patterns of Desistance from Offending. *European Journal of Criminology*, 4(3), pp.315–45.

McCabe, D. and Castel, A., 2008. Seeing is believing: The effect of brain images on judgments of scientific reasoning. *Cognition*, 107(1), pp.343–352.

McGlynn, R.P. and Dreilinger, E.A., 1981. Mock juror judgment and the insanity plea: Effects of incrimination and sanity information. *Journal of Applied Social Psychology*, 11(2), pp.166–180.

Mehta, S. and Farina, A., 1997. Is being "sick" really better? Effect of the disease view of mental disorder on stigma. *Journal of Social and Clinical Psychology*, 16(4), pp.405–419.

Mele, A., 2012. Another scientific threat to free will? *The Monist*, 95(3), pp.422–440.

Menninger, K.A., 1946. *The Human Mind*. Whitefish, Montana: Kessinger Publishing Co.

Merritt, A.C., Effron, D.A. and Monin, B., 2010. Moral self-licensing: When being good frees us to be bad. *Social and Personality Psychology Compass*, 4(5), pp.344–357.

Merritt, C. and Fowler, R., 1948. The pecuniary honesty of the public at large. *The Journal of Abnormal and Social Psychology*, 43(1), pp.90–93.

Messick, D.M. and Sentis, K.P., 1979. Fairness and preference. *Journal of Experimental Social Psychology*, 15(4), pp.418–434.

Metcalfe, C., Pickett, J.T. and Mancini, C., 2015. Using path analysis to explain racialized support for punitive delinquency policies. *Journal of Quantitative Criminology*, 31(4), pp.699–725.

Michael, R.B., Newman, E.J., Vuorre, M., Cumming, G. and Garry, M., 2013. On the (non) persuasive power of a brain image. *Psychonomic Bulletin & Review*, 20(4), pp.720–725.

Miles, D. and Carey, G., 1997. Genetic and environmental architecture on human aggression. *Journal of Personality and Social Psychology*, 72(1), pp.207–217.

Miles, J.B., 2013. 'Irresponsible and a Disservice': The integrity of social psychology turns on the free will dilemma. *British Journal of Social Psychology*, 52(2), pp.205–218.

Milgram, S., Mann, L. and Harter, S., 1965. The lost-letter technique: A tool of social research. *Public Opinion Quarterly*, 29(3), pp.437–438.

Miller v. Alabama, 567 U.S. __ [2012].

Mischel, W. and Ebbesen, E., 1972. Cognitive and attentional mechanisms in delay of gratification. *Journal of Personality and Social Psychology*, 21(2), pp.204–218.

Moffitt, T.E., Caspi, A. and Rutter, M., 2005. Strategy for investigating interactions between measured genes and measured environments. *Archives of General Psychiatry*, 62(5), pp.473–81.

Monroe, A.E., Brady, G.L. and Malle, B.F., 2017. This isn't the free will worth looking for: General free will beliefs do not influence moral judgments, agent-specific choice ascriptions do. *Social Psychological and Personality Science*, 8(2), pp.191–199.

Monroe, A.E. and Malle, B.F., 2010. From uncaused will to conscious choice: The need to study, not speculate about people's folk concept of free will. *Review of Philosophy and Psychology*, 1(2), pp.211–224.

Montada, L., 1998. Belief in a just world: A hybrid of justice motive and self-interest? In: L. Montada and M.J. Lerner, eds., *Responses to Victimization and Belief in a Just World*. New York, NY: Plenum Press, pp.217–246.

Monterosso, J., Royzman, E.B. and Schwartz, B., 2005. Explaining away responsibility: Effects of scientific explanation on perceived culpability. *Ethics & Behavior*, 15(2), pp.139–158.

Monterosso, J. and Schwartz, B., 2012. Did your brain make you do it? *The New York Times*. [online] 27 Jul. Available at: <http://www.nytimes.com/2012/07/29/opinion/sunday/neuroscience-and-moral-responsibility.html?_r=0>.

Moore, M.S., 1997. *Placing Blame: A Theory of Criminal Law*. Oxford, UK: Oxford University Press.

Morewedge, C., Gray, K. and Wegner, D., 2010. Perish the forethought: Premeditation engenders misperceptions of personal control. In: R. Hassin, K. Oschsner and Y. Trope, eds., *Self Control in Society, Mind, and Brain*. Oxford, UK: Oxford University Press, pp.260–278.

Morgan, M., 2013. *Blame My Brain: The Amazing Teenage Brain Revealed*. London, UK: Walker.

Morris, H., 1968. Persons and Punishment. *The Monist*, 52, pp.475–501.

Morris, N., 1974. The future of imprisonment: Toward a punitive philosophy. *Michigan Law Review*, 72(6), p.1161.

Morris, N., 1982. *Madness and the Criminal Law*. Chicago, IL: University of Chicago

Press.

Morse, S., 2004a. New neuroscience, old problems. In: B. Garland, ed., *Neuroscience and the Law: Brain, Mind, and the Scales of Justice*. New York, NY: Dana Press, pp.157–198.

Morse, S., 2004b. Reason, results, and criminal responsibility. *University of Illinois Law Review*, 1(1), pp.363–444.

Morse, S., 2006. Moral and legal responsibility and the new neuroscience. In: J. Illes, ed., *Neuroethics: Defining the issues in theory, practice, and policy*. Oxford, UK: Oxford University Press, pp.33–50.

Morse, S.J., 2008a. Determinism and the death of folk psychology: Two challenges to responsibility from neuroscience. *Minnesota Journal of Law, Science & Technology*, 9(1), pp.1–36.

Morse, S.J., 2008b. Psychopathy and criminal responsibility. *Neuroethics*, 1(3), pp.205–212.

Moynihan, A.B., Igou, E.R. and van Tilburg, W.A.P., 2017. Free, connected, and meaningful: Free will beliefs promote meaningfulness through belongingness. *Personality and Individual Differences*, 107, pp.54–65.

Murphy, J.G., 2007. Legal Moralism and Retribution Revisited. *Criminal Law and Philosophy*, 1, pp.5–20.

Murray, D. and Nahmias, E., 2014. Explaining away incompatibilist intuitions. *Philosophy and Phenomenological Research*, 88(2), pp.434–467.

Nadelhoffer, T., Gromet, D., Goodwin, G., Nahmias, E., Sripada, C. and Sinnott-Armstrong, W., 2013. The mind, the brain, and the law. In: T.A. Nadelhoffer, ed., *The Future of Punishment*. New York, NY: Oxford University Press, pp.193–212.

Nadelhoffer, T., Heshmati, S. and Kaplan, D., 2013. Folk retributivism and the communication confound. *Economics and Philosophy*, 29(2), pp.235–261.

Nadelhoffer, T., Shepard, J., Nahmias, E., Sripada, C. and Ross, L.T., 2014. The free will inventory: Measuring beliefs about agency and responsibility. *Consciousness and Cognition*, 25, pp.27–41.

Nadelhoffer, T., Sripada, C., Gromet, D. and Sinnott-Armstrong, W., n.d. Neuroscience, Future Dangerousness, and the Criminal Law. *Unpublished manuscript*.

Nahmias, E., 2006. Folk fears about freedom and responsibility: Determinism vs. reductionism. *Journal of Cognition and Culture*, 6(1–2), pp.215–237.

Nahmias, E., Coates, D.J. and Kvaran, T., 2007. Free will, moral responsibility, and mechanism: Experiments on folk intuitions. *Midwest Studies in Philosophy*, 31(1), pp.214–242.

Nahmias, E., Morris, S., Nadelhoffer, T. and Turner, J., 2005. Surveying Freedom: Folk Intuitions about free will and moral responsibility. *Philosophical Psychology*, 18(5), pp.561–584.

Nahmias, E., Morris, S.G., Nadelhoffer, T. and Turner, J., 2006. Is incompatibilism intuitive? *Philosophy and Phenomenological Research*, 73(1), pp.28–53.

Nahmias, E., Shepard, J. and Reuter, S., 2014. It's OK if 'my brain made me do it': People's intuitions about free will and neuroscientific prediction. *Cognition*, 133(2), pp.502–516.

Najdowski, C. and Bottoms, B., 2009. Jurors' perceptions of juvenile defendants: The influence of intellectual disability, abuse history, and confession evidence. *Behavioral Sciences & Law*, 27(3), pp.401–430.

Napier, J.L. and Jost, J.T., 2008. Why are conservatives happier than liberals? *Psychological Science*, 19(6), pp.565–572.

Newman, G.E., Bloom, P. and Knobe, J., 2014. Value judgments and the true self. *Personality & Social Psychology Bulletin*, 40(2), pp.203–16.

Newman, G.E., De Freitas, J. and Knobe, J., 2015. Beliefs about the true self explain asymmetries based on moral judgment. *Cognitive Science*, 39(1), pp.96–125.

Nichols, S., 2011. Experimental philosophy and the problem of free will. *Science*, 331(6023), pp.1401–1403.

Nichols, S. and Knobe, J., 2007. Moral responsibility and determinism: The cognitive science of folk intuitions. *Nous*, 41(4), pp.663–685.

Nisbett, R. and Wilson, T., 1977. Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3), pp.231–259.

Nozick, R., 1981. *Philosophical Explanations*. Cambridge, MA: Harvard University Press.

O'Brien, L.T. and Major, B., 2005. System-justifying beliefs and psychological well-being: The roles of group status and identity. *Personality and Social Psychology Bulletin*, 31(12), pp.1718–1729.

Open Science Collaboration, ., 2015. Estimating the reproducibility of psychological science. *Science*, 349.

Otto, K. and Dalbert, C., 2005. Belief in a just world and its functions for young prisoners. *Journal of Research in Personality*, 39(6), pp.559–573.

Paulhus, D.L. and Carey, J.M., 2011. The FAD-Plus: Measuring lay beliefs regarding free will and related constructs. *Journal of Personality Assessment*, 93(1), pp.96–104.

Pfeffer, K., Maxwell, M. and Briggs, A., 2012. Attributions for youth crime, accountability and legal competence. *Journal of Criminal Psychology*, 2(2), pp.127–139.

Pitts-Taylor, V., 2010. The plastic brain: Neoliberalism and the neuronal self. *Health*, 14(6), pp.635–652.

Pizarro, D., Uhlmann, E. and Salovey, P., 2003. Asymmetry in judgments of moral blame and praise: The role of perceived metadesires. *Psychological Science*, 14(3), pp.267–272.

Pizarro, D.A. and Tannenbaum, D., 2012. Bringing character back: How the motivation to evaluate character influences judgments of moral blame. In: M. Mikulincer and P.R. Shaver, eds., *The Social Psychology of Morality: Exploring the Causes of Good and Evil*. Washington, DC, US: American Psychological Association, pp.91–108.

Polderman, T.J.C., Benyamin, B., de Leeuw, C.A., Sullivan, P.F., van Bochoven, A., Visscher, P.M. and Posthuma, D., 2015. Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics*, 47(7), pp.702–709.

Pratt, T.C. and Cullen, F.T., 2000. The empirical status of Gottfredson and Hirschi's general theory of crime: A meta-analysis. *Criminology*, 38(3), pp.931–964.

Pronin, E. and Kugler, M., 2010. People believe they have more free will than others. *Proceedings of the National Academy of Sciences*, 107(52), pp.22469–22474.

van Prooijen, J.-W., 2006. Retributive reactions to suspected offenders: The importance of social categorizations and guilt probability. *Personality & Social Psychology Bulletin*, 32(6), pp.715–726.

van Prooijen, J.-W. and van den Bos, K., 2009. We blame innocent victims more than I do: Self-construal level moderates responses to just-world threats. *Personality & Social Psychology Bulletin*, 35(11), pp.1528–1539.

van Prooijen, J.-W. and Lam, J., 2007. Retributive justice and social categorizations: The perceived fairness of punishment depends on intergroup status. *European Journal of Social Psychology*, 37(6), pp.1244–1255.

Protzko, J., Ouimette, B. and Schooler, J., 2016. Believing there is no free will corrupts intuitive cooperation. *Cognition*, 151, pp.6–9.

Quinton, A., 1954. On punishment. *Analysis*, 14, pp.1933–1942.

Rafter, N., 2008. *The Criminal Brain: Understanding Biological Theories of Crime*. New York, NY: New York University Press.

Raine, A., 2013. *The Anatomy of Violence: The Biological Roots of Crime*. New York, NY: Pantheon Books.

Raine, A. and Glenn, A.L., 2014. *Psychopathy: An Introduction to Biological Findings and their Implications*. New York, NY: New York University Press.

Raine, A., Laufer, W.S., Yang, Y., Narr, K.L., Thompson, P. and Toga, A.W., 2012. Increased executive functioning, attention, and cortical thickness in white-collar criminals. *Human Brain Mapping*, 33(12), pp.2932–2940.

Raine, A., Lencz, T., Bihrle, S., LaCasse, L. and Colletti, P., 2000. Reduced prefrontal gray matter volume and reduced autonomic activity in antisocial personality disorder. *Archives of General Psychiatry*, 57(2), pp.119–127.

Rand, D.G., Armao IV, J.J., Nakamaru, M. and Ohtsuki, H., 2010. Anti-social punishment can prevent the co-evolution of punishment and cooperation. *Journal of Theoretical Biology*, 265(4), pp.624–632.

Read, J., Haslam, N., Sayce, L. and Davies, E., 2006. Prejudice and schizophrenia: A review of the 'mental illness is an illness like any other' approach. *Acta Psychiatrica Scandinavica*, 114(5), pp.303–318.

Reichle, B., Schneider, A. and Montada, L., 1998. How do observers of victimization preserve their belief in a just world cognitively or actionally? In: L. Montada and M.J. Lerner, eds., *Responses to Victimization and Belief in a Just World*. New York, NY: Plenum Press, pp.55–86.

Retz, W., Retz-Junginger, P., Supprian, T., Thome, J. and Rösler, M., 2004. Association of serotonin transporter promoter gene polymorphism with violence: Relation with personality disorders, impulsivity, and childhood ADHD psychopathology. *Behavioral Sciences & the Law*, 22(3), pp.415–425.

Rhee, S. and Waldman, I., 2002. Genetic and environmental influences on antisocial behavior: A meta-analysis of twin and adoption studies. *Psychological Bulletin*, 128(3), pp.490–529.

Rigoni, D., Kühn, S., Gaudino, G., Sartori, G. and Brass, M., 2012. Reducing self-control by weakening belief in free will. *Consciousness and Cognition*, 21(3), pp.1482–1490.

Rigoni, D., Kühn, S., Sartori, G. and Brass, M., 2011. Inducing disbelief in free will alters brain correlates of preconscious motor preparation. *Psychological Science*, 22(5), pp.613–618.

Rigoni, D., Pourtois, G. and Brass, M., 2015. 'Why should I care?' Challenging free will attenuates neural reaction to errors. *Social Cognitive and Affective Neuroscience*, 10(2), pp.262–268.

Roberts, J. V., 2009. The recidivist premium: Revisiting the recidivist sentencing premium. In: A. von Hirsch and A. Ashworth, eds., *Principled Sentencing: Readings on Theory and Practice*, 3rd ed. Oxford, UK: Hart Publishing, pp.148–162.

Roberts, J. V. and Hough, M., 2011. Custody or community? Exploring the boundaries of public punitiveness in England and Wales. *Criminology and Criminal Justice*, 11(2), pp.181–197.

Roberts, J. V, Crutcher, N. and Verbrugge, P., 2007. Public attitudes to sentencing in Canada: Exploring recent findings. *Canadian Journal of Criminology and Criminal Justice*, 49(1), pp.75–107.

Robinson, P.H. and Darley, J.M., 1997. The utility of desert. *Northwestern University Law Review*, 91, pp.453–499.

Rogers, R. and Zimbarg, R., 1987. Antisocial backgrounds of defendants evaluated for insanity: A research note. *International Journal of Law and Psychiatry*, 10(1), pp.75–80.

Roper v. Simmons, 543 U.S. 551 [2005].

Rose, D. and Nichols, S., 2013. The lesson of bypassing. *Review of Philosophy and Psychology*, 4(4), pp.599–619.

Rose, N., 2000. The biology of culpability: Pathological identity and crime control in a biological culture. *Theoretical Criminology*, 4(1), pp.5–34.

Roskies, A. and Nichols, S., 2008. Bringing moral responsibility down to earth. *The Journal of Philosophy*, 105(7), pp.371–388.

Ross, L., 1977. The intuitive psychologist and his shortcomings. In: L. Berkowitz, ed., *Advances in Experimental Social Psychology (Volume 10)*. New York, NY: Academic Press, pp.173–220.

Rucker, D.D., Polifroni, M., Tetlock, P.E. and Scott, A.L., 2004. On the assignment of

punishment: The impact of general-societal threat and the moderating role of severity. *Personality & Social Psychology Bulletin*, 30(6), pp.673–684.

Saks, M.J., Schweitzer, N.J., Aharoni, E. and Kiehl, K.A., 2014. The impact of neuroimages in the sentencing phase of capital trials. *Journal of Empirical Legal Studies*, 11(1), pp.105–131.

Sampson, R. and Laub, J., 2003. Life course desisters? Trajectories of crime among delinquent boys followed to age 70*. *Criminology*, 41(3), pp.555–92.

Sampson, R.J., 1997. Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science*, 277(5328), pp.918–924.

Sanderson, C.A., Zanna, A.S. and Darley, J.M., 2000. Making the punishment fit the crime and the criminal: Attributions of dangerousness as a mediator of liability. *Journal of Applied Social Psychology*, 30(6), pp.1137–1159.

Sarkissian, H., Chatterjee, A. and Brigard, F. De, 2010. Is belief in free will a cultural universal? *Mind & Language*, 25(3), pp.346–358.

Schweitzer, N.J., Baker, D.A. and Risko, E.F., 2013. Fooled by the brain: Re-examining the influence of neuroimages. *Cognition*, 129(3), pp.501–511.

Schweitzer, N.J. and Saks, M.J., 2011. Neuroimage evidence and the insanity defense. *Behavioral Sciences & the Law*, 29(4), pp.592–607.

Schweitzer, N.J., Saks, M.J., Murphy, E.R., Roskies, A.L., Sinnott-Armstrong, W. and Gaudet, L.M., 2011. Neuroimages as evidence in a mens rea defense: No impact. *Psychology, Public Policy, and Law*, 17(3), pp.357–393.

Scott, E.S., Reppucci, N.D., Antonishak, J. and DeGennaro, J.T., 2006. Public attitudes about the culpability and punishment of young offenders. *Behavioral Sciences & the Law*, 24(6), pp.815–832.

Scurich, N. and Appelbaum, P., 2016. The blunt-edged sword: Genetic explanations of misbehavior neither mitigate nor aggravate punishment. *Journal of Law and the Biosciences*, 3(1), pp.140–157.

Scurich, N. and Monahan, J., 2016. Evidence-based sentencing: Public openness and opposition to using gender, age, and race as risk factors for recidivism. *Law and Human Behavior*, 40(1), pp.36–41.

Selke, W.L., 1980. The impact of higher education on crime orientations. *Journal of Criminal Justice*, 8(3), pp.175–184.

Shalvi, S., Dana, J., Handgraaf, M.J.J. and De Dreu, C.K.W., 2011. Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior. *Organizational Behavior and Human Decision Processes*, 115(2), pp.181–190.

Shalvi, S., Eldar, O. and Bereby-Meyer, Y., 2012. Honesty requires time (and lack of justifications). *Psychological Science*, 23(10), pp.1264–1270.

Shariff, A.F., Greene, J.D., Karremans, J.C., Luguri, J.B., Clark, C.J., Schooler, J.W., Baumeister, R.F. and Vohs, K.D., 2014. Free will and punishment: A mechanistic view of human nature reduces retribution. *Psychological Science*, 25(8), pp.1563–1570.

Shaw, C.R. and McKay, H.D., 1942. *Juvenile Delinquency and Urban Areas*. Chicago:

Chicago University Press.

Shelley, T.O., Waid, C.A. and Dobbs, R.R., 2011. The influence of criminal justice major on punitive attitudes. *Journal of Criminal Justice Education*, 22(4), pp.526–545.

Shepard, J. and Reuter, S., 2012. Neuroscience, choice, and the free will debate. *AJOB Neuroscience*, 3(3), pp.7–11.

Shepherd, J., 2012. Free will and consciousness: Experimental studies. *Consciousness and Cognition*, 21(2), pp.915–927.

Siegel, D., 2014. *Brainstorm: The Power and Purpose of the Teenage Brain*. London, UK: Penguin.

Smith, K.B. and Green, D.N., 1984. Individual correlates of the belief in a just world. *Psychological Reports*, 54(2), pp.435–438.

Snead, O., 2007. Neuroimaging and the 'complexity' of capital punishment. *New York University Law Review*, 82(5), pp.1265–1339.

Soon, C.S., Brass, M., Heinze, H.-J. and Haynes, J.-D., 2008. Unconscious determinants of free decisions in the human brain. *Nature Neuroscience*, 11(5), pp.543–545.

Spiranovic, C.A., Roberts, L.D., Indermaur, D., Warner, K., Gelb, K. and Mackenzie, G., 2011. Public preferences for sentencing purposes: What difference does offender age, criminal history and offence type make? *Criminology and Criminal Justice*, 12(3), pp.289–306.

Sripada, C.S., 2010. The Deep Self Model and asymmetries in folk judgments about intentional action. *Philosophical Studies*, 151(2), pp.159–176.

Sripada, C.S., 2012a. Mental state attributions and the side-effect effect. *Journal of Experimental Social Psychology*, 48(1), pp.232–238.

Sripada, C.S., 2012b. What makes a manipulated agent unfree? *Philosophy and Phenomenological Research*, 85(3), pp.563–593.

Stalans, L.J. and Lurigio, A.J., 1990. Lay and professionals' beliefs about crime and criminal sentencing: A need for theory, perhaps schema theory. *Criminal Justice and Behavior*, 17(3), pp.333–349.

Steffensmeier, D. and Demuth, S., 2001. Ethnicity and judges' sentencing decisions: Hispanic-black-white comparisons. *Criminology*, 39(1), pp.145–178.

Steinberg, L., 2009. Should the science of adolescent brain development inform public policy? *American Psychologist*, 64(8), pp.739–750.

Steinberg, L., 2012. Should the science of adolescent brain development inform public policy? *Issues in Science and Technology*, 28(3), pp.67–78.

Steinberg, L., 2013. The influence of neuroscience on US Supreme Court decisions about adolescents' criminal culpability. *Nature Reviews Neuroscience*, 14(7), pp.513–518.

Steinberg, L. and Scott, E., 2003. Less guilty by reason of adolescence: Developmental immaturity, diminished responsibility, and the juvenile death penalty. *American*

Psychologist, 58(12), pp.1009–1018.

Stevenson, M.C., 2009. Perceptions of juvenile offenders who were abused as children. *Journal of Aggression, Maltreatment & Trauma*, 18(4), pp.331–349.

Stevenson, M.C., Bottoms, B.L. and Diamond, S.S.S., 2010. Jurors' discussions of a defendant's history of child abuse and alcohol abuse in capital sentencing deliberations. *Psychology, Public Policy, and Lawand Law*, 16(1), pp.1–38.

Strohminger, N., Knobe, J. and Newman, G., 2017. The true self: A psychological concept distinct from the self. *Perspectives on Psychological Science*, 12(4), pp.551–560.

Sutton, R.M. and Douglas, K.M., 2005. Justice for all, or just for me? More evidence of the importance of the self-other distinction in just-world beliefs. *Personality and Individual Differences*, 39(3), pp.637–645.

Sutton, R.M. and Winnard, E.J., 2007. Looking ahead through lenses of justice: The relevance of just-world beliefs to intentions and confidence in the future. *British Journal of Social Psychology*, 46(3), pp.649–666.

Sykes, G.M. and Matza, D., 1957. Techniques of neutralization: A theory of delinquency. *American Sociological Review*, 22(6), pp.664–670.

Tetlock, P.E., 2002. Social functionalist frameworks for judgment and choice: Intuitive politicians, theologians, and prosecutors. *Psychological Review*, 109(3), pp.451–471.

Tetlock, P.E., Self, W.T. and Singh, R., 2010. The punitiveness paradox: When is external pressure exculpatory – and when a signal just to spread blame? *Journal of Experimental Social Psychology*, 46(2), pp.388–395.

Thapar, A., Langley, K., Fowler, T., Rice, F., Turic, D., Whittinger, N., Aggleton, J., Van den Bree, M., Owen, M., O'Donovan, M., Raine, A., Moffitt, T.E., Wechsler, D., Dolan, R.J., Anderson, S.W., Anderson, S.W., Winterer, G., Gogos, J.A., Huotari, M., Lewis, D.A., Thapar, A., Brennan, P.A., Rhee, S.H., Tibbetts, S., Botting, N., Brooks, A.A., Rutter, M., Angold, A., Walton, K.A., Hsu, T.M., Rutter, M., Taylor, E., Egan, M.F., Malhotra, A.K., Weinberger, D.R., Mattay, V.S., Diamond, A., Fossella, J., Tsai, S.J., Goldberg, T.E., Bilder, R.M., Mills, S. and Caspi, A., 2005. Catechol O-Methyltransferase gene variant and birth weight predict early-onset antisocial behavior in children with Attention-Deficit/Hyperactivity Disorder. *Archives of General Psychiatry*, 62(11), p.1275.

The American Law Institute, ., 1962. *Model Penal Code: Sentencing*. Philadelphia, PA.

Thomson, J., 1985. The trolley problem. *The Yale Law Journal*, 94(6), pp.1395–1415.

Tonry, M.H., 2011. Emotion, intuition and determinism. In: M.H. Tonry, ed., *Why Punish? How Much? A Reader on Punishment*. Oxford, UK: Oxford University Press.

Topalli, V., 2005. When being good is bad: An expansion of neutralization theory. *Criminology*, 43(3), pp.797–836.

Trust for London and New Policy Institute, ., 2016. *Overview of London boroughs*. [online] London's Poverty Profile. Available at: <http://www.londonspovertyprofile.org.uk/key-facts/overview-of-london-boroughs/> [Accessed 11 Jan. 2017].

Turkheimer, E., 2000. Three laws of behavior genetics and what they mean. *Current Directions in Psychological Science*, 9(5), pp.160–164.

Turner, J. and Tajfel, H., 1982. *Social Identity and Intergroup Relations*. Cambridge, UK: Cambridge University Press.

Twenge, J., Baumeister, R. and Tice, D., 2001. If you can't join them, beat them: Effects of social exclusion on aggressive behavior. *Journal of Personality and Social Psychology*, 81(6), pp.1058–1069.

Twenge, J.M. and Foster, J.D., 2008. Mapping the scale of the narcissism epidemic: Increases in narcissism 2002–2007 within ethnic groups. *Journal of Research in Personality*, 42(6), pp.1619–1622.

Tyler, T. and Boeckmann, R., 1997. Three strikes and you are out, but why? The psychology of public support for punishing rule breakers. *Law and Society Review*, 31(2), pp.237–266.

Vohs, K.D. and Schooler, J.W., 2008. The value of believing in free will: Encouraging a belief in determinism increases cheating. *Psychological Science*, 19(1), pp.49–54.

Walen, A., 2016. *Retributive Justice*. [online] The Stanford Encyclopedia of Philosophy. Available at: <https://plato.stanford.edu/archives/win2016/entries/justice-retributive/> [Accessed 20 Mar. 2018].

Walker, I. and Read, J., 2002. The differential effectiveness of psychosocial and biogenetic causal explanations in reducing negative attitudes toward "mental illness". *Psychiatry: Interpersonal and Biological Processes*, 65(4), pp.313–325.

Waller, B.N., 2011. *Against Moral Responsibility*. Cambridge, Massachusetts: The MIT Press.

Waller, R.R., 2014. The threat of effective intentions to moral responsibility in the zygote argument. *Philosophia*, 42(1), pp.209–222.

Warr, M., Meier, R.F. and Erickson, M.L., 1983. Norms, theories of punishment, and publicly preferred penalties for crimes. *The Sociological Quarterly*, 24(1), pp.75–91.

Warren, J.I., Murrie, D.C., Chauhan, P., Dietz, P.E. and Morris, J., 2004. Opinion formation in evaluating sanity at the time of the offense: An examination of 5175 pre-trial evaluations. *Behavioral Sciences & The Law*, 22(2), pp.171–186.

Wegner, D.M., 2003. The mind's best trick: How we experience conscious will. *Trends in Cognitive Sciences*, 7(2), pp.65–69.

Weiner, B., Perry, R. and Magnusson, J., 1988. An attributional analysis of reactions to stigmas. *Journal of Personality and Social Psychology*, 55(5), pp.738–748.

Weisberg, D., Taylor, J. and Hopkins, E., 2015. Deconstructing the seductive allure of neuroscience explanations. *Judgment and Decision Making*, 10(5), pp.429–441.

Weisberg, D.S., Keil, F.C., Goodstein, J., Rawson, E. and Gray, J.R., 2008. The seductive allure of neuroscience explanations. *Journal of Cognitive Neuroscience*, 20(3), pp.470–477.

Wikström, P. and Treiber, K., 2009. Violence as situational action. *International Journal of Conflict and Violence*, 3(1), pp.75–96.

Woolfolk, R.L., Doris, J.M. and Darley, J.M., 2006. Identification, situational constraint, and social cognition: Studies in the attribution of moral responsibility. *Cognition*, 100(2), pp.283–301.

Wright, A.J., Sutton, S.R., Hankins, M., Whitwell, S.C.L., Macfarlane, A. and Marteau, T.M., 2012. Why does genetic causal information alter perceived treatment effectiveness? An analogue study. *British Journal of Health Psychology*, 17(2), pp.294–313.

Wright, J.P., Beaver, K.M., DeLisi, M., Vaughn, M.G., Boisvert, D. and Vaske, J., 2008. Lombroso's legacy: The miseducation of criminologists. *Journal of Criminal Justice Education*, 19(3), pp.325–338.

Wright, J.P. and Cullen, F.T., 2012. The future of biosocial criminology: Beyond scholars' professional ideology. *Journal of Contemporary Criminal Justice*, 28(3), pp.237–253.

Yang, Y., Raine, A., Narr, K.L., Lencz, T., LaCasse, L., Colletti, P. and Toga, A.W., 2007. Localisation of increased prefrontal white matter in pathological liars. *The British Journal of Psychiatry*, 190(2), pp.174–175.

Zonana, H. V., Wells, J.A., Getz, M.A. and Buchanan, J., 1990. Part I: The NGRI registry: Initial analyses of data collected on Connecticut insanity acquittees. *Journal of the American Academy of Psychiatry and the Law*, 18(2), pp.115–128.

Glossary

The Acting Self. 'The narrow set of outcome-directed proximal desires, means-end beliefs, and intentions, that are the immediate causal source of the action' (Sripada, p.7).

The assumption of the conscious brain. The philosophical assumption made by most neuroscientists that brain activity is responsible for conscious thoughts, desires and plans.

The assumption of determinism. The philosophical assumption made by most neuroscientists that brain activity is the product of uncontrollable prior events.

The assumption of the social brain. The philosophical assumption made by most neuroscientists that brain activity is the product of uncontrollable social (not just genetic) factors.

The biosocial offender. The offender whose criminal behaviour was described in relation to genetic and social influences on behaviour.

The capacity for rational choice. The general capacity of the offender to form intentions from a basic knowledge of reality and illegality, and to act upon those intentions through self-control.

The caused offender. The offender whose criminal behaviour was described as the causal effect of the brain or genes and the environment.

Compatibilism. The view that determinism is compatible with the existence of free will.

The compatibilist. A person who believes that determinism is compatible with the existence of free will.

Corrupted information. The perceived 'corrupting of the information used by' the offender to form intent (Sripada, p.7).

The Deep Self. A person's stable traits, i.e., 'a much larger set of more stable, enduring and fundamental attitudes' (Sripada, p.7).

Deep Self discordance. The perceived 'discordance' of the offending behaviour from the 'deep self' of the offender (Sripada, p.7).

The Deep Self Model. In the Deep Self Model, Sripada (p.7) predicts that people blame offenders for their immoral actions when their criminally Acting Self – 'the narrow set of outcome-directed proximal desires, means-end beliefs, and intentions, that are the immediate causal source of the action' – concords with the stable traits of their Deep Self – 'a much larger set of more stable, enduring and fundamental attitudes'.

Determinism. The idea that all behaviour is the inevitable product of prior uncontrollable events.

The deterministic incompatibilist. A person who believes all behaviour is the inevitable product of prior uncontrollable events and that this fact rules out the existence of free will.

The determinist. A person who believes all behaviour is the inevitable product of prior uncontrollable events.

The distally caused offender. The offender whose criminal behaviour originates from factors in the distant past.

The Explanation Hypothesis. The idea that people attribute moral responsibility to actors when the most salient explanation for their behaviour is a unique property of the self (Björnsson and Persson, 2013).

Incompatibilism. The view that determinism is incompatible with the existence of free will.

The incompatibilist. A person who believes determinism is incompatible with the existence of free will.

The incompatibilist neurocriminologist. An academic who publishes (empirical or library) research on the relationship between the brain and criminal behaviour and also believes determinism is incompatible with the existence of free will.

Moral alarm. The anxiety experienced in the process of causing harm.

The neurobiological offender. The offender whose criminal behaviour was described in relation to the neurobiological influences on behaviour.

The neurocriminologist. An academic who publishes (empirical or library) research on the relationship between the brain and criminal behaviour.

Neurocriminology. The study of how the volume and activity of different brain regions correlate with criminal behaviours.

Neuroscience. The study of the brain.

Psychopathy. The mental disorder characterised by a lack of moral alarm and more generally, a lack of empathy for the suffering of others.

The real caused offender. The offender whose real-life criminal behaviour was described as the causal effect of the brain or genes and the environment.

The real self. The self resulting from the normal processes of chance implicated in biosocial development.

Appendices

i. Appendix A: Experimental materials and variables for the theatre study

The full script for the play can be found at the end of Appendix A. Below are the main excerpts from this script that cover the neuroscientific content.

The adolescent brain

You say to me your brain is broken. It's like an adult's brain but it doesn't work properly. It's like you're in a city you've never been to, and you don't have a map, and you don't know what you're doing, and you keep taking the wrong turns. You say, listen to me, don't worry, one day you'll be OK, probably. Your brain will start working properly. One day your brain will be just like mine, and then you'll be OK, but until then, you've got to try and be more, like me.

I say to you, my brain isn't broken. It's beautiful. I'm in a city I've never been to and I see bright lights and new ideas and fear and opportunity and a thousand million roads all lit up and flashing. I say, my brain isn't broken, it's like this for a reason. I'm like this for a reason. I'm becoming who I am and I'm scared and you're scared, because who I am might not be who you want me to be, or who you are, and I don't know why, but I don't say it's all going to be OK. There are so many things I don't say to you. I don't know why. I want to say them, but I can't. I pick up my plate, put it in the kitchen, and go upstairs.

The prefrontal cortex

There's this bit at the front of the brain called the prefrontal cortex. It's right here, right in the front of your brain. It's the bit that controls your planning and consequences and stuff like that. It's the really sensible bit. It's the bit that stops you doing stupid stuff.

The bit that says, ‘Don’t do that, or that will happen’. It’s a bit like your Dad. [Some time later.] The prefrontal cortex is one of the last bits of the teenage brain to develop. It takes age and ages to work properly.

The limbic system

There’s this part of the brain called the Limbic System. It’s in the middle of the brain and it’s made up of loads of different bits…and there’s one part called the Reward Centre that makes you want to have fun. It’s the bit of the brain that makes you do stuff you shouldn’t. When you’re a teenager your Limbic System is really hypersensitive. It goes wild whenever you do something fun or exciting or risky. I actually did a project at school about this. Shall I show you? So this is the brain and it’s divided into different parts, this is the diagram that I drew, um, it’s… So basically the Limbic System is a complex system of nerves and networks in the brains, so that includes, err, the olfactory bulbs, hippocampus, amygdala, that’s really important that’s like the main bit and the bit that does rewards and stuff and then there’s hold on I’ll just read these out, there’s the…the anterior thalamic nuclei, fornix, columns of fornix, mammillary body, septum pellucidum, habenular commissure, cingulate gyrus, parahippocampal gyrus, limbic cortex, and limbic midbrain areas. [Some time later.] The Limbic System is the bit that gives you the kick out of taking risks.

Example envelope

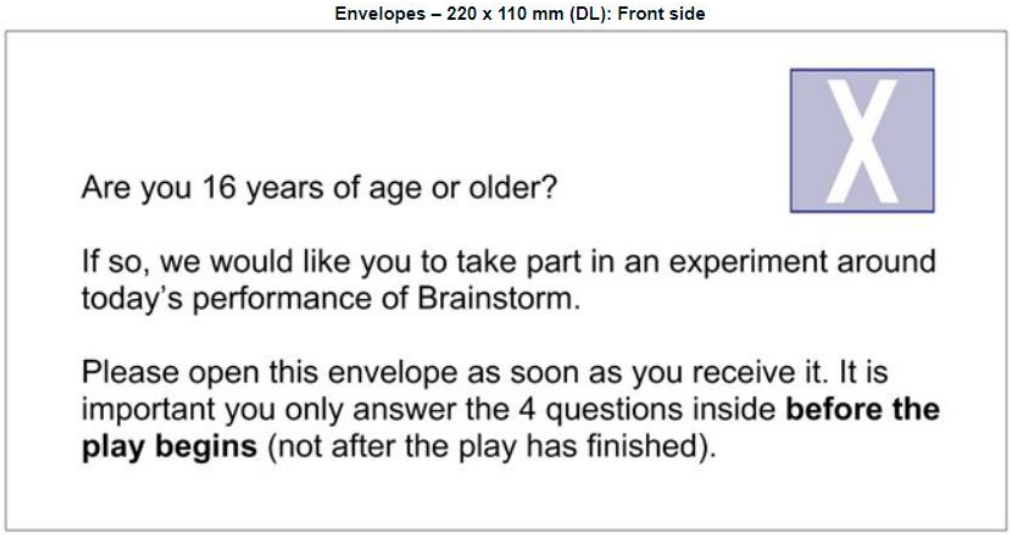


Figure A1. Example front of the envelope.

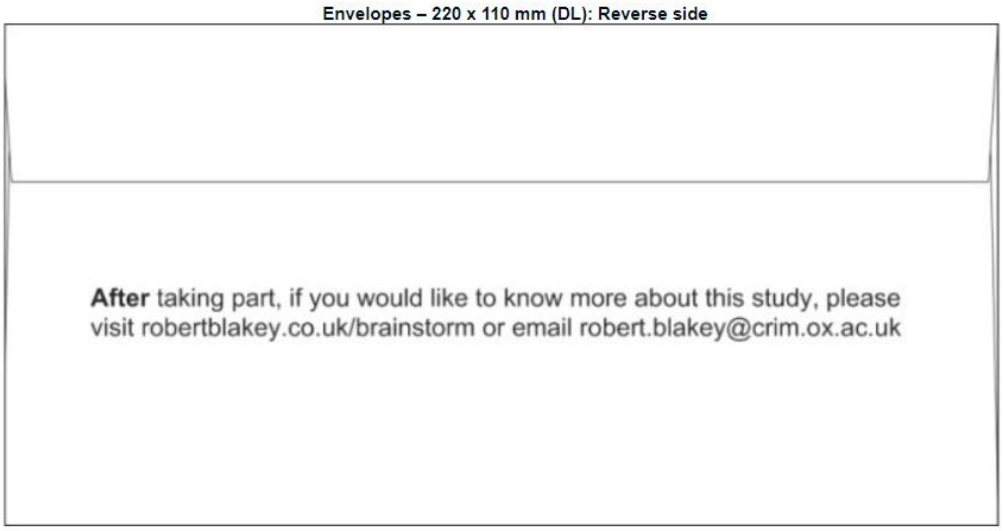
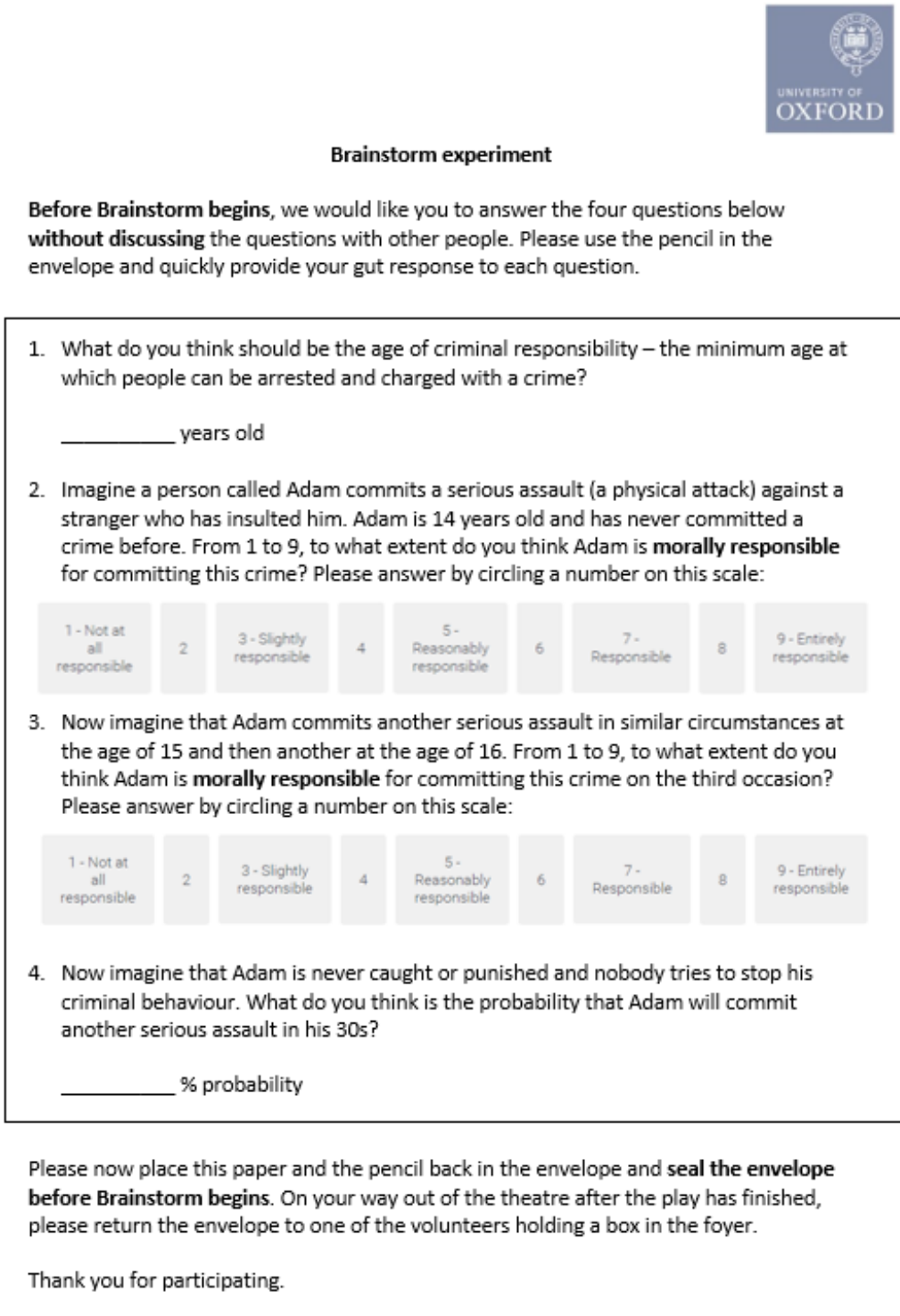


Figure A2. Example back of the envelope.

Example survey



UNIVERSITY OF OXFORD

Brainstorm experiment

Before Brainstorm begins, we would like you to answer the four questions below **without discussing** the questions with other people. Please use the pencil in the envelope and quickly provide your gut response to each question.

1. What do you think should be the age of criminal responsibility – the minimum age at which people can be arrested and charged with a crime?

   _____ years old

2. Imagine a person called Adam commits a serious assault (a physical attack) against a stranger who has insulted him. Adam is 14 years old and has never committed a crime before. From 1 to 9, to what extent do you think Adam is **morally responsible** for committing this crime? Please answer by circling a number on this scale:

1 - Not at all responsible	2	3 - Slightly responsible	4	5 - Reasonably responsible	6	7 - Responsible	8	9 - Entirely responsible

3. Now imagine that Adam commits another serious assault in similar circumstances at the age of 15 and then another at the age of 16. From 1 to 9, to what extent do you think Adam is **morally responsible** for committing this crime on the third occasion? Please answer by circling a number on this scale:

1 - Not at all responsible	2	3 - Slightly responsible	4	5 - Reasonably responsible	6	7 - Responsible	8	9 - Entirely responsible

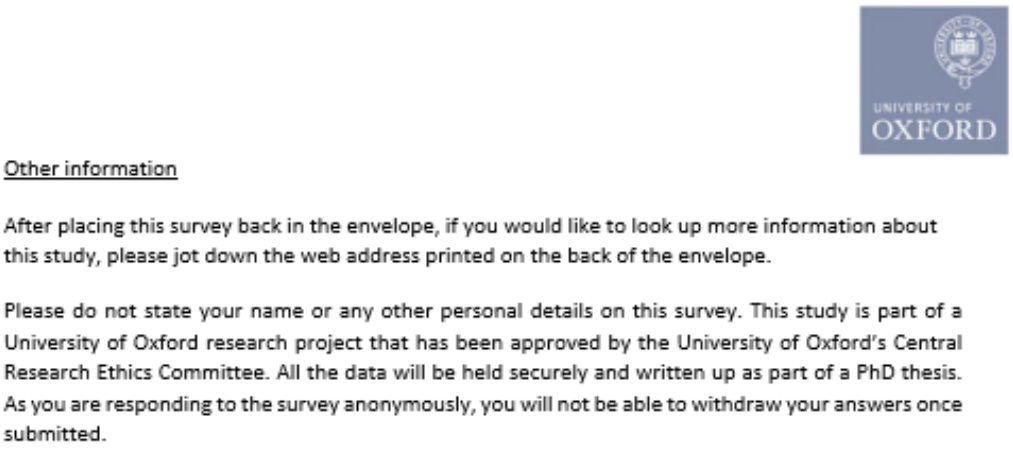
4. Now imagine that Adam is never caught or punished and nobody tries to stop his criminal behaviour. What do you think is the probability that Adam will commit another serious assault in his 30s?

   _____ % probability

Please now place this paper and the pencil back in the envelope and **seal the envelope before Brainstorm begins**. On your way out of the theatre after the play has finished, please return the envelope to one of the volunteers holding a box in the foyer.

Thank you for participating.

Figure A3. Example front page of the theatre study survey.



Other information

After placing this survey back in the envelope, if you would like to look up more information about this study, please jot down the web address printed on the back of the envelope.

Please do not state your name or any other personal details on this survey. This study is part of a University of Oxford research project that has been approved by the University of Oxford's Central Research Ethics Committee. All the data will be held securely and written up as part of a PhD thesis. As you are responding to the survey anonymously, you will not be able to withdraw your answers once submitted.

Figure A4. Example rear page of the theatre study survey.

The full script for Brainstorm

[The play script originally presented here cannot be made freely available via ORA due to copyright reasons. The script was sourced from Company Three and can be purchased online at Amazon.co.uk].

ii. Appendix B: Experimental materials and variables for the online study

This study vignette is below and based on Stimulus C of Schweitzer et al. (2011).

Overview

Donny Adams, age 32, has been found guilty of Inflicting Grievous Bodily Harm on his neighbour, 38-year-old Robert Samuels.

Mr Samuels was working in his garden when his lawnmower launched a small stone into Mr Adams's window, cracking it. The offender, Mr Adams, immediately ran from his house and struck Mr Samuels with a hammer, causing him to fall to the ground unconscious.

The emergency hospital doctor found that Mr Samuels had been knocked out by a single heavy blow to his head and suffered moderate brain injuries as a result.

The prosecutor

Mr Samuels was working in his back garden with a lawnmower and heard a loud noise as if the lawnmower had hit a stone. After a few seconds, the offender, Mr Adams,

emerged from his house and ran directly at Mr Samuels while screaming and swinging a hammer.

Mr Adams already had the hammer in his hand. He had been using the hammer to knock a screw into a wall to hang up a picture at the time the stone cracked his window.

Mr Adams struck Mr Samuels forcefully in the head with the hammer. Mr Samuels collapsed to the ground. The offender, Mr Adams, ran back into his house and closed the door, while two witnesses rang the emergency services.

The prosecutor (continued): a 'rage attack'

The type of attack committed by Mr Adams is known as a 'rage attack'. A 'rage attack' is an attack in which the intensity of violence appears to be out of proportion with the intensity of the triggering event (in this case, the window cracking).

The previous day

The prosecutor argued that events the previous day could help explain this 'rage attack'. The day before the stone cracked his window, Mr Adams checked his girlfriend's phone and found that Mr Samuels had been sending her flirty messages. By the time the window cracked, Mr Adams had not confronted his girlfriend or Mr Samuels about the messages.

Consequently, Mr Adams became enraged when the stone from Mr Samuels's lawnmower cracked his window. In retaliation, Mr Adams attacked Mr Samuels.

The prosecutor (continued)

The victim Mr Samuels sustained moderate brain damage from the forceful blow to his head. He stayed in a coma for 20 days. He has since come out of the coma and returned

to his home. However, Mr Samuels continues to have difficulty remembering many words and controlling fine motor movements (such as holding pencils or typing).

Mr Samuels is married with 2 young children. He is a teacher in a local high school.

The survey questions

(Unless otherwise stated, participants responded on the following scale: *strongly disagree*, *disagree*, *somewhat disagree*, *somewhat agree*, *agree*, *strongly agree*.)

The judge asks you to recommend the number of years that Mr Adams should spend in prison.

The maximum prison sentence for Inflicting Grievous Bodily Harm is 16 years. How much time would you require Mr Adams to spend in prison?

Number of years

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

Prison 

Prison is not the only option for dealing with Mr Adams. A second option is secure accommodation that aims to reform (rather than punish) offenders.

Secure accommodation is designed to be as pleasant as possible, while also ensuring that nobody can escape. In comparison to a normal prison, Mr Adams will receive far more services that aim to reduce reoffending by addressing the underlying causes of his offending behaviour.

Mr Adams can be sent to secure accommodation for a maximum of 16 years after he has completed any prison sentence. Therefore, the combined intervention of prison and secure accommodation can last a maximum of 32 years.

Now that you have the options of prison and/or secure accommodation: how much time would you require Mr Adams to spend in prison? Following his prison sentence, how much time would you require Mr Adams to spend in secure accommodation?

Number of years

	0 1 2 3 4 5 6 7 9 10 11 12 13 14 15 16
Prison	
Secure accommodation	

How much do you agree or disagree with the following statements?

It is Mr Adams's fault that he attacked Mr Samuels.

If nobody tries to stop him, Mr Adams is likely to commit another 'rage attack'.

One purpose of secure accommodation is to prevent the offender re-offending by addressing the underlying causes of the offending behaviour - in other words, to reform the offender.

How much do you agree or disagree with the following statement?

It is possible to reform Mr Adams so that he stops committing 'rage attacks'.

One purpose of prison is to deter the offender from re-offending by making the offender fear being caught and sent to prison again.

How much do you agree or disagree with the following statement?

With a long enough prison sentence, Mr Adams could be deterred from committing 'rage attacks' in the future.

As you read earlier, Mr Adams normally uses half as much inhibition as the average person, meaning that Mr Adams normally exercises half the amount of control over his impulses as the average person.

How much do you agree or disagree with the following statements?

It is Mr Adams's fault that he normally exercises half the amount of control over his impulses as the average person.

As you read earlier, Mr Adams's frontal brain lobes normally use half as much oxygen as the average person's frontal brain lobes, meaning that Mr Adams normally exercises half the amount of control over his impulses as the average person.

How much do you agree or disagree with the following statements?

It is Mr Adams's fault that he normally exercises half the amount of control over his impulses as the average person.

Before the day he attacked Mr Samuels, Mr Adams was consciously aware that he normally exercises less control over his impulses than the average person.

The average person normally exercises the full amount of control over his/her impulses.
How much do you agree or disagree with the following statement?

It would take Mr Adams a lot more effort than the average person to exercise this full amount of control over his impulses.

Now we would like you to think about what happened from the moment the stone cracked Mr Adams's window.

How much do you agree or disagree with the following statements?

From the moment the stone cracked his window, Mr Adams could have resisted the impulse to attack Mr Samuels.

At some point after the stone cracked his window, Mr Adams made a conscious decision to attack Mr Samuels.

How much do you agree or disagree with the following statements?

Mr Adams did not really want to attack Mr Samuels.

Mr Adams's decision to attack Mr Samuels reflects his personality.

How important do you think each of the following factors was in causing Mr Adams to attack Mr Samuels?

His free choice

(*Very unimportant*, *unimportant*, *somewhat unimportant*, *somewhat important*, *important*, *very important*)

Evil intent

His brain

His conscious mind

Unconscious processes

His emotions

His social life experiences (i.e., experiences with other people)

His genes

The government is considering a new policy. This policy would involve two steps:

- Applying the same tests that were used on Mr Adams to the general population to reliably identify anyone who normally exercises less than half the amount of control over their impulses as the average person.
- Taking these identified people to secure accommodation, even if their lack of control has not yet caused anyone harm.

These identified people would be required to live in secure accommodation until they started to use more than half the amount of control as the average person. Inside this accommodation, the main goal would be to increase the person's use of control through special interventions.

How much do you agree or disagree with the following statement?

The government should introduce this new policy.

Imagine that since birth, Mr Adams had the exact same genes and life experiences as you have (instead of his own genes and his own life experiences). As you read before, Mr Samuels has been messaging Mr Adams's partner and a stone from Mr Samuel's lawnmower cracks Mr Adams's window.

How much do you agree or disagree with the following statement?

With your genes and your life experiences (instead of his own), Mr Adams would respond to the cracked window by attacking Mr Samuels.

Imagine that since birth, you had the exact same genes and life experiences as Mr Adams has (instead of your own genes and your own life experiences). You therefore

also live in Mr Adams's house, Mr Samuels has been messaging your partner and a stone from Mr Samuels's lawnmower cracks your window.

How much do you agree or disagree with the following statement?

With Mr Adams's genes and life experiences (instead of your own), you would respond to the cracked window by attacking Mr Samuels.

Brain transplant

Leading scientists predict that advances in technology will make it possible to transplant a brain from one person to another person in the near future.

Imagine that it is already possible to transplant brains from one person to another. Just before the stone cracks his window, Mr Adams's brain is replaced with your brain and your brain is replaced with Mr Adams's brain. In other words, you and Mr Adams swap brains.

So Mr Adams now has your brain.

How much do you agree or disagree with the following statement?

With your brain (instead of his own), Mr Adams would respond to the cracked window by attacking Mr Samuels.

Brain transplant (continued)

Remember you now have Mr Adams's brain. Imagine you also live in Mr Adams's house, Mr Samuels has been messaging your partner and a stone from Mr Samuels's lawnmower cracks your window.

How much do you agree or disagree with the following statement?

With Mr Adams's brain (instead of your own), you would respond to the cracked window by attacking Mr Samuels.

What is your gender?

What is your age?

What is your nationality?

Please select the country that represents your nationality.

What is the highest level of education that you have completed?

(*Less than secondary school*, *secondary school*, *sixth form or college*, *university: bachelor's (undergraduate) degree*, *university: master's degree*, *university: doctorate degree*)

iii. Appendix C: Experimental materials and variables for the museum study

Below is the vignette presented in the museum study.

The offence

What tempted you? I was getting off an airplane. I had been away with friends from university to celebrate us having finished our degrees. We were the last to check in for the return flight so we all had to sit separately from each other. I was sitting at the very back of the plane so I was literally the last person to get off the airplane when we landed. As I walked down the aisle to the plane door, I noticed a purse had been left on one of the seats.

How did you respond? I grabbed the purse. I was going to hand it in but in the end I took it home instead. My girlfriend at the time really wanted to go to a particular luxury holiday resort but I couldn't afford it. So I tried to use one of the credit cards in the purse to buy the holiday. It worked. I bought the holiday. I never told my girlfriend about the card – I cut it up and threw it away as soon as I had used it.

Written by: Adam.

The science

For the experimental manipulation, below is the exact text that participants read, with the three variations indicated:

Every condition

Ertan scientists have discovered why some Ertans have better self-control than others. The answer is about the [conscious decisions/genes and environments/brains] of Ertans. Whenever an Ertan is faced with a temptation, the Ertan must consciously decide whether to control his/her urge to give into temptation or not. [The biosocial and neurobiological conditions continued as follows.]

The biosocial condition

The scientists have discovered that this conscious decision is completely caused by two factors: 1. The specific genes that the Ertan has inherited from his/her parents, 2. The specific influences in the environments that the Ertan has experienced across his life, including family and friends. The scientists have discovered that 50% of the differences between how much self-control different Ertans use can be explained by differences in their genes. The other 50% can be explained by differences in their environmental experiences.

The neurobiological condition

The scientists have discovered that this conscious decision is completely caused by the specific chemical reactions and electrical activity occurring in the Ertan's brain. The scientists have discovered that 100% of the differences between how much self-control different Ertans use can be explained by differences in the chemical reactions and electrical activity that occur in their brains.

[On the subsequent page, participants read an application of the explanation to the specific offender, Adam, as shown below.]

Every condition

One day Adam is faced with the temptation described to you on the piece of paper. He consciously decides to give in to temptation and buy the holiday with the stolen credit card.

[For the biosocial and neurobiological conditions, the text continued as follows.]

The biosocial condition

[In the real-world condition: According to scientific research,] his conscious decision is completely caused by his specific genes and environmental experiences.

The neurobiological condition

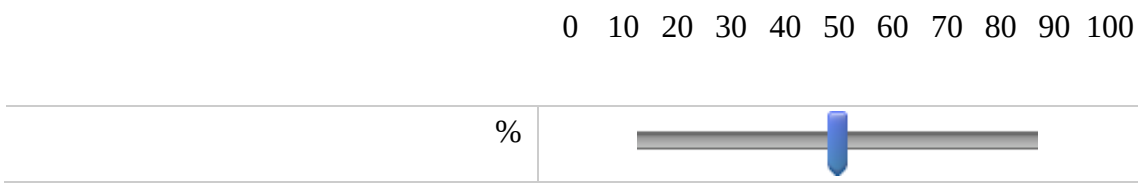
[In the real-world condition: According to scientific research,] his conscious decision is completely caused by the specific chemical reactions and electrical activity occurring in his brain.

The survey questionsHow much do you agree or disagree with the statement below?

It is Adam’s fault that he decided to give in to the temptation.

Imagine that every visitor to the Science Museum was asked to agree or disagree that Adam’s decision to give in to temptation was his fault.

What percent (%) of visitors do you think would agree that the decision was Adam's fault?



How much do you agree or disagree with each of the statements below?

Adam wanted to give in to this temptation.

Adam's decision to give into temptation reflects him as a person.

How much do you agree or disagree with each of the statements below?

Adam could have decided to resist the temptation.

If nobody tries to stop him, Adam is likely to give into similar temptations in the future.

You somewhat agreed/agreed/strongly agreed that Adam is likely to give into similar temptations in the future.

How much do you agree or disagree with the statement below?

Adam can be changed so that he stops giving into similar temptations in the future.

When Adam gave into temptation, another person lost money as a result. This person was the victim of Adam's actions.

How much do you agree or disagree with the statement below?

The victim was capable of preventing Adams from giving in to this temptation.

Imagine Adam's actions were brought to court and you acted as the judge in that court. As the judge, you are asked to recommend an appropriate punishment for Adam. The only available option is prison.

For how many weeks would you send Adam to prison?

Normally in England, the maximum punishment that a judge can choose for this type of theft is 36 weeks (in other words, 9 months) in prison.

	0 4 8 12 16 20 24 28 32 36
Number of weeks	

Now we would like to ask you the same questions that we previously asked about Adam the Ertan. This time please answer about Adam the person who visited the Science Museum in July this year.

It is important you answer carefully because we hope to publish the results of this experiment in a national prison newspaper read by prisoners across the country. / It is important you answer carefully because we hope to publish the results of this experiment in a national newspaper read by people across the country.

[The questions capturing participants' attitudes towards Adam were repeated.]

Now we would like you to complete an anonymous version of the task that Adam completed. First, take a piece of paper from the box labelled 'Blank paper' and bring it back here.

On this piece of paper we will ask you to write about a time when you faced a temptation and either resisted or gave in to that temptation. We will then ask you to place this piece of paper in the closed box labelled 'Today's temptations'.

From tomorrow onwards, everyone who participates in this experiment will be able to read the pieces of paper collected in today's box. Therefore, future visitors to the museum will read what you write. However, visitors will not know it was you who wrote your particular temptation because you will not write down your name.

Please choose just one temptation from your life. It be can something tiny from today or something big from a long time ago. Under the appropriate titles on the piece of paper, now write down:

- The temptation
- How you responded to that temptation

Please do not place the paper in the box yet.

Please type the number at the bottom of the paper into the box below.

_____

Did your response to the temptation involve resisting or giving into temptation?

Resisting

Giving in

To what extent do you think your response to the temptation was a morally good thing to do?

(*Not at all good, somewhat good, quite good, good, very good, extremely good*)

To what extent do you think your response to the temptation was a morally bad thing to do?

(*Not at all bad, somewhat bad, quite bad, bad, very bad, extremely bad*)

Now place your piece of paper in the box labelled 'Today's temptations' and come straight back here.

Some of the final questions in the experiment may appear to ask for sensitive information. Therefore, if you would prefer not to answer some of these questions, simply leave the answer spaces blank and move on to the next question.

All your responses will be completely anonymous and kept confidential – we do not ask for your name and your answers will never be associated with any other identifying information.

Below is a list of things you may have done. How often have you done each of these things in the last 5 years?

How often have you...

...downloaded music, TV shows, films or other files from internet sites that may be illegal?

(*Never*, *once*, *twice*, *3 or 4 times*, *5 times or more*)

...agreed to pay or be paid in cash to avoid taxes?

...when selling a second-hand item, hid or not disclosed faults in what you were selling?

...bought an item, used or worn it and then tried to return it to the store for a refund?

...during an insurance claim, added items which had not been lost, damaged or stolen, or increased the value of any items claimed?

...tried to claim for replacement items or compensation you knew you were not entitled to?

...received a speeding fine?

...travelled by train without a valid train ticket?

In this experiment you read two explanations:

An explanation of why some Ertans have better self-control than others

An explanation of why some humans have better self-control than others

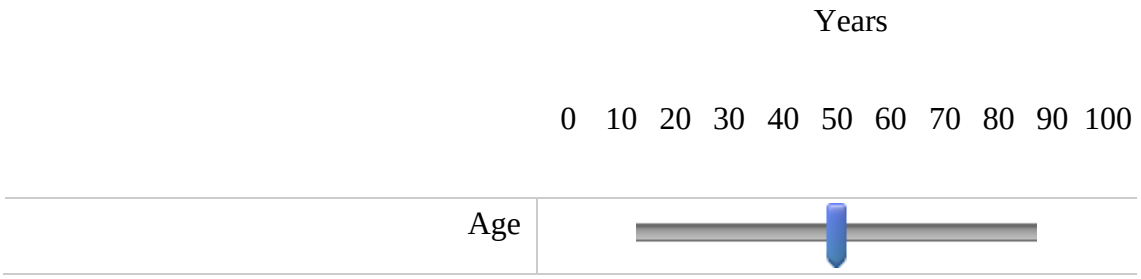
To what extent did you believe the explanation of why some humans have better self-control than others?

(*Strongly disbelieved*, *disbelieved*, *somewhat disbelieved*, *somewhat believed*, *believed*, *strongly believed*)

What is your nationality?

Please select the country that represents your nationality.

What is your age?



What is your gender?

We would like to send you a 5-minute survey on the same topic in a few months' time. This would greatly benefit our research. Would you be happy for us to send you this survey?

(*Yes*, *No*)

Would you like us to send you a summary of the results after the experiment is complete?

(*Yes*, *No*)

If you answered 'Yes' to either of the above questions, please type your email address into the box below.

Now click the red arrow to complete the experiment. You may then read the temptations that participants wrote down yesterday and before yesterday in the box labeled ‘Yesterday’s temptations’.

Follow-up to the museum study

This survey has only one page.

How much do you agree or disagree with the following statements?

Prisoners should have access to televisions or gym facilities.

If prison has to be used, it should be used sparingly and only as a last option.

My general view towards offenders is that they should be treated harshly.

People always have free will.

How people’s lives unfold is completely up to them.

People ultimately have complete control over their decisions and their actions.

iv. Appendix D: Experimental materials and variables for the Facebook study

Below are the scales used in the Facebook study.

Psychopathy

To what extent do you agree or disagree with the following statements? (Try to answer spontaneously.)

(1 = *strongly disagree*, 6 = *strongly agree*)

I like to get revenge on authorities.

I avoid dangerous situations.

People often say I’m out of control.

It’s true that I can be mean to others.

People who mess with me always regret it.

I have never gotten into trouble with the law.

I enjoy having sex with people I hardly know.

I’ll say anything to get what I want.

Free will

To what extent do you agree or disagree with the following statements? (Try to answer spontaneously.)

(1 = *strongly disagree*, 7 = *strongly agree*)

People always have the ability to do otherwise.

People always have free will.

How people’s lives unfold is completely up to them.

People ultimately have complete control over their decisions and their actions.

People have free will even when their choices are completely limited by external circumstances.

Determinism

To what extent do you agree or disagree with the following statements? (Try to answer spontaneously.)

(1 = *strongly disagree*, 7 = *strongly agree*)

Everything that has ever happened had to happen precisely as it did, given what happened before.

Every event that has ever occurred, including human decisions and actions, was completely determined by prior events.

People's choices and actions must happen precisely the way they do because of the laws of nature and the way things were in the distant past.

A supercomputer that could know everything about the way the universe is now could know everything about the way the universe will be in the future.

Given the way things were at the Big Bang, there is only one way for everything to happen in the universe after that.

Dualism

Please answer the following questions about your duplicate. (Try to answer spontaneously.)

(1 = *definitely no*, 7 = *definitely yes*)

Does your duplicate have the same eye colour as you?

When your duplicate behaves badly, does your duplicate experience the same amount of anxiety as you when you behave badly?

Is the immune system in your duplicate the same strength as the immune system in you?

Does your duplicate display the same amount of good and bad behaviour as you?

Is your duplicate's nose the same shape as your nose?

Is the moral alarm in your duplicate the same strength as the moral alarm in you?

The survey questions

The crying baby

Please imagine this scenario:

Enemy soldiers have taken over your village. They have orders to kill all remaining civilians. You and some of your townspeople have sought refuge in the cellar of a large house.

Outside you hear the voices of soldiers who have come to search the house for valuables. Your baby begins to cry loudly. You cover his mouth to block the sound. If you remove your hand from his mouth his crying will summon the attention of the soldiers who will kill you, your child, and the others hiding out in the cellar. To save yourself and the others you must smother your child to death.

Would you smother your child in order to save yourself and the other townspeople?

(*Yes*, *No*)

How much do you agree or disagree with the following statement?

I feel guilty about my response to the question on the previous page.

The trolley and the switch

Please imagine this scenario:

A runaway trolley is headed for five people who will be killed if it proceeds on its present course. The only way to save the five people is to hit a switch that will turn the trolley onto an alternate set of tracks where it will kill one person instead of five.

Would you hit the switch to turn the trolley in order to save five people at the expense of one?

(*Yes*, *No*)

How much do you agree or disagree with the following statement?

I feel guilty about my response to the question on the previous page.

The trolley and the bridge

Please imagine this scenario:

A trolley threatens to kill five people. You are standing next to a large stranger on a footbridge that spans the tracks, in between the oncoming trolley and the five people.

The only way to save the five people is to push this stranger off the bridge, onto the tracks below. The stranger will die if you do this, but his body will stop the trolley from reaching the five people on the tracks.

Would you save the five people on the tracks by pushing this stranger to his death?

(*Yes*, *No*)

How much do you agree or disagree with the following statement?

I feel guilty about my response to the question on the previous page.

Everyone who takes part in this study will be entered into a lottery to win between £10 and £60 (so between €11.50 and €69). Now it is time to determine how much money you could win.

To do this, we will ask you to roll a virtual dice three times. The outcome of your first roll (1, 2, 3, 4, 5 or 6) will determine the amount of money you could win for taking part in this study. The higher the number you roll, the more money you could win.

Specifically, if you roll a 1, you will have the chance to win £10. If you roll a 2, you will have the chance to win £20. If you roll a 3, you will have the chance to win £30, 4 = £40, 5 = £50 and 6 = £60.

Please note this survey software can only record the answers you click or type into the survey. Therefore, we will ask you to report the outcome of your dice rolls at the end of this task.

Please now roll the virtual die by clicking the red arrow.

Please wait while the die is being rolled.

The virtual die has been rolled. The outcome of your first roll is:

2

When you are ready to roll the die a second time, please click the red arrow below.

Please wait while the die is being rolled.

The virtual die has been rolled. The outcome of your second roll is:

6

When you are ready to roll the die a third time, please click the red arrow below.

Please wait while the die is being rolled.

The virtual die has been rolled. The outcome of your third roll is:

3

Please click the red arrow below in order to enter the outcome of your first roll (the amount of money you could win).

You will be automatically moved on to the remaining pages of the study in 30 seconds. Before this happens, please indicate below the outcome of your first roll. Then click the red arrow.

(If you do not respond within 30 seconds, we will be forced to enter you for the smallest prize, £10.)

(*1 (£10), 2 (£20), 3 (£30), 4 (£40), 5 (£50), 6 (£60)*)

As you know, at the end of this study, you will be entered into a lottery to win money. The winner of this lottery will be a certain numbered person to take part in the study (e.g., the 5th person to take part). Since the winning number has already been randomly chosen, we can award the money as soon as the winner completes the study.

Below are 2 options for when you will receive the money if you win. Please choose an option. I would like to:

(*...find out immediately whether I have won the money, ...find out 3 months later whether I have won the money. If you choose this option, we will add £60 to the prize money*)

In this study, you read an explanation of how the moral alarm reduces the desire to behave badly.
To what extent did you believe this explanation of the moral alarm?

(*Strongly disbelieved, disbelieved, somewhat disbelieved, somewhat believed, believed, strongly believed*)

In this study, you were given feedback about the strength of your own moral alarm from an analysis of your Facebook likes.

To what extent did you believe that this feedback accurately described the strength of your moral alarm?

(*Strongly disbelieved, disbelieved, somewhat disbelieved, somewhat believed, believed, strongly believed*)

Were you suspicious of any part of this study?

If no, please move onto the next page.

If yes, please describe your suspicions below.

_____

_____

_____

_____

_____

The final questions in the study may appear to ask for sensitive information. Therefore, if you would prefer not to answer some of these questions, simply leave the answer spaces blank and move on to the next question. All your responses will be completely anonymous and kept confidential.

What is your nationality?

Please select the country that represents your nationality.

What is your age?

_____

What is your gender?

(*Female, male, other*)

Which device did you use to participate in this study?

(*Phone, tablet, laptop*)

What is the highest level of education that you have completed?

(*Less than secondary school, secondary school, sixth form or college, university: bachelor's (undergraduate) degree, university: master's degree, university: doctorate degree*)

Have you studied (or are you studying) any of the following subjects at university?
If not, please leave all options blank.
You may tick more than one box if you have studied more than one subject at university.

(*Law, Psychology, Neuroscience, Biology, Any other science, Sociology, Criminology, Any other social science, Any other subject from the arts or humanities*)

Would you like us to send you a summary of the results from this study after the study is complete?

(*Yes, No*)

Would you like to be entered into the lottery to win the prize money?

(*Yes, No*)

If you answered 'Yes' to either of the above questions, please type your email address into the box below. (Please note this email address will be stored separately to all your other responses in this study to ensure your answers remain anonymous.)

v. Appendix E: Experimental materials and variables for the postcard study

Example postcards

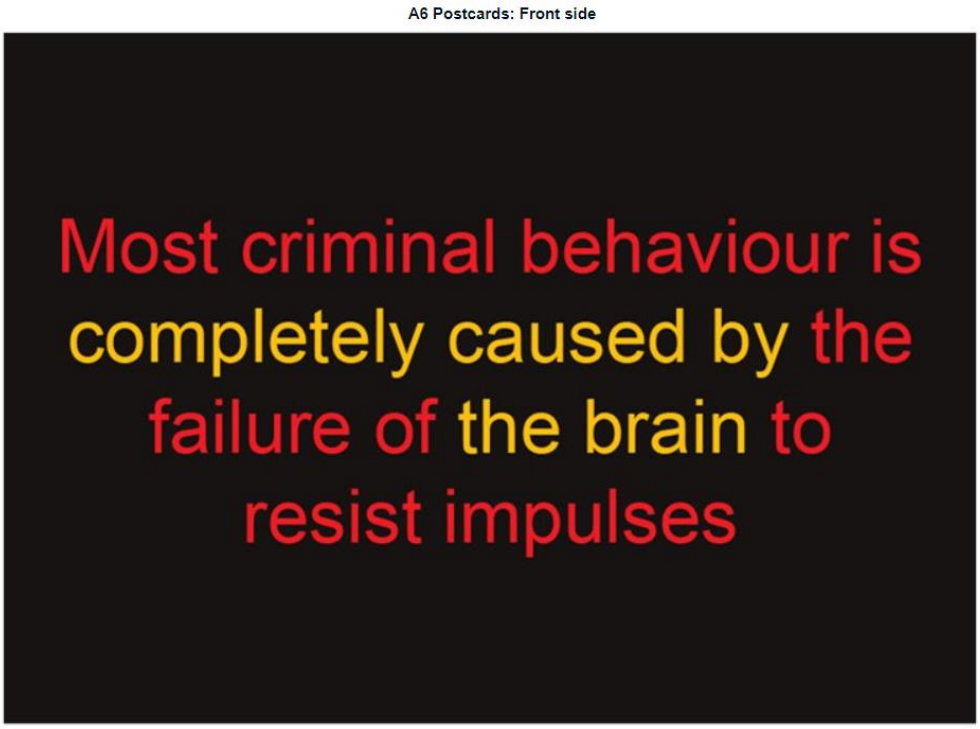


Figure A5. Example front page of the postcard.

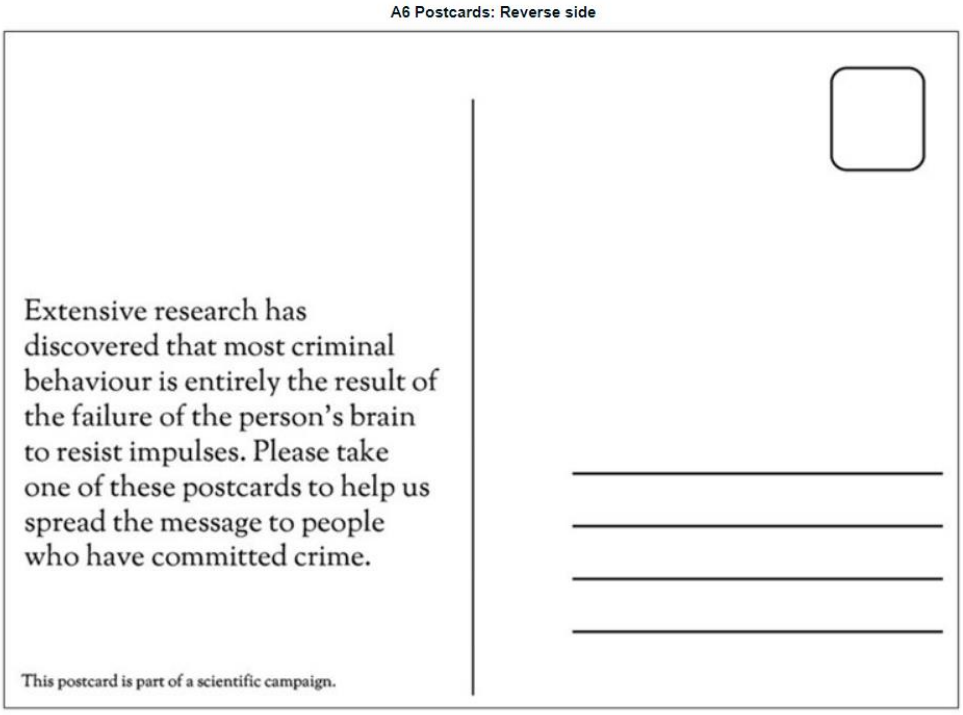


Figure A6. Example rear page of the postcard (without the handwritten text).