BRIATHAR DÉ MO LÓCHRANN

UNIVERSITY of LIMERICK

OLLSCOIL LUIMNIGH

"It's wrong but I can't explain why!"

Moral Dumbfounding and Moral Judgement: How Failure to Justify Moral Judgements can Inform our Understanding of How they are Made

Cillian McHugh

Thesis submitted to

Mary Immaculate College ~ University of Limerick

For the degree of Doctor of Philosophy

Supervisors: Dr Marek McGann, Dr Eric R. Igou, Dr Elaine L. Kinsella

Submitted to Mary Immaculate College, University of Limerick

May 2018

Abstract

Moral dumbfounding occurs when people fail to justify a strongly held moral judgement with supporting reasons. The discovery of moral dumbfounding coincided with a growth in intuitionist and dual-process theories of moral judgement over rationalist theories, and its existence has directly informed their development (e.g., Haidt, 2001; Prinz, 2005; Bucciarelli, Khemlani, & Johnson-Laird 2008; Dwyer, 2009; Cushman, Young, & Greene 2010). Despite the influence of moral dumbfounding on the morality literature, the phenomenon is poorly understood. Direct evidence in support of dumbfounding is limited to a single study (Haidt, Björklund, & Murphy, 2000), which had a final sample of 30 participants and was never published in peer-reviewed form. The aim of the current project is to examine the phenomenon of moral dumbfounding directly, firstly, to test if it is a real phenomenon, and secondly to evaluate how the existence (or absence) of moral dumbfounding can inform theories of moral judgement. Three studies demonstrate that dumbfounding is a genuine phenomenon that can be reliably elicited in a laboratory setting, and develop methods for studying dumbfounding. Two studies address specific challenges to dumbfounding, and demonstrate that (a) people do not reliably articulate reasons that may be governing their judgement, and (b) moral principles are not consistently applied across differing contexts. A final set of studies tested two hypothesised explanations of moral dumbfounding associated with dual-process theory (e.g., Cushman, 2013; Crockett, 2013), and model theory (Bucciarelli et al., 2008). Using a range of manipulations across seven studies, the observed evidence for these explanations is weak. That dumbfounding is poorly explained by existing theories of moral judgement presents a significant limitation of current theories of moral judgement. To address this limitation, a possible alternative theoretical approach that provides an explanation for moral dumbfounding is explored.

Declaration of Originality

College: Mary Immaculate College ~ University of Limerick

Department: Psychology

Degree: Ph.D.

Name of Candidate: Cillian McHugh

ID Number: 0858994

Title of Thesis: Moral Dumbfounding and Moral Judgement: How Failure to Justify Moral Judgements can Inform our Understanding of How they are Made

Declaration: I hereby declare that this thesis is the result of my own original research and does not contain the work of any other individual. All sources that have been consulted have been identified and acknowledged in the appropriate way.

Signature of Candidate: ____

Cillian McHugh

Acknowledgements

Firstly, I would like to thank Marek, Eric, and Elaine, without whom this thesis would never have been written. All three of you were incredibly generous with your expertise and your time over the past few years. Marek you were always around for a chat (whether you were expecting one or not!). You got the project started, and were always available to answer my questions, big and small. Your answers were usually accompanied by direction towards relevant reading. This often involved simply taking a book off your shelf and handing it to me – I now have a stack of books belonging to you at home. Also, your written feedback was always a healthy combination of insight, guidance, and humour! Eric, I met you first in the upgrade panel when I transferred from the Masters to the Ph.D. register. Following this, you came on board as a second supervisor, bringing fresh eyes and a new perspective to the project. Your input really helped in designing the studies conducted. You continually challenged me to think outside my comfort zone, introducing me to a much broader literature in the process. Elaine, your input has been invaluable. Your keen eye for detail brought an extra level of rigour to the designing of studies, and gave additional clarity and structure to the overall project. You ensured that each step in the process was grounded in a clear articulable rationale. On top of this, you were always so encouraging in your feedback; often giving me a confidence boost when it was most needed.

I would also like to thank the staff in the Department of Psychology in MIC, both past and present. It has been a pleasure to work with you these past few years in both a research and a teaching capacity. I am especially grateful to Michelle Glasheen for all your help with the practicalities of collecting data, particularly in the early days when I needed to learn how to use the various software packages (with your help). You ensured I had everything I needed for all the studies (even the ones that didn’t make it into the final thesis). Beyond this, I really appreciate your continuous support and encouragement, you always stopped for a chat to see how the project was coming along. In addition, I would like to thank Sandra O’ Brien, in UL for your help with the later studies, and working with online platforms.

A special thanks my colleagues in the Academic Learning Centre, who listened to regular updates on the progress of the project. I am lucky to have worked with you all, and I have learned loads from you. To all the staff in MISU, you have been amazing. I didn’t expect to spend as much time in there as I did, but hey!

A very special thanks to my family. You have supported me through ups and downs of the past few years. You supported me in every way imaginable, and I would not have made it to this point without you. Even if I was not always “present”, spending a lot of time hiding behind my laptop (particularly the most recent Christmas holidays, when write-up was in full swing), it was always nice to come home, even if it was only for dinner.

Finally, a massive thanks to my friends and colleagues (you know who you are), who offered me enthusiastic support and encouragement while also ensuring that I remember that there is also life beyond the thesis.

One last word of thanks to Dr John Perry and Professor Roger Giner-Sorolla who made the examination process a pleasant, (almost relaxing) experience. Your insightful feedback was helpful and helped to improve the quality of the thesis as a whole.

Table of Contents

Introduction – Overview and Summary

On the 26th January 2016, the Scottish public petitions committee rejected a petition to legalise incest between consenting adults. The rejection was unanimous and took less than a minute. There was no discussion on the substantive content of the petition and the rationale for rejection did not extend beyond there is “no public interest” in pursuing it further. The need to avoid discussion on the petition's subject matter was also apparent in the newspaper coverage at the time, which focused on the “loophole” that allowed such a petition to reach the public petitions committee in the first place (‘MSPs throw out incest petition’, 2016).

According to Richard Morris, the petition's author, Scottish law is discriminatory and infringes on the autonomy of the individual. It has also been argued elsewhere that Scottish law may be incompatible with the European Convention on Human Rights (Roffee, 2014). In rejecting the petition without debate, these issues were not addressed, and any need to provide a rationale for why the state should legislate against incest was successfully avoided. Whether or not the MSPs (Members of the Scottish Parliament) would have been able to provide such a rationale, or if an appropriate rationale can be provided at all, is therefore unknown. The possibility that there might not be a reasoned justification of the law was ignored by both the media and the MSPs on the public petitions committee.

The above anecdote is an example of coherence between national law and a moral norm. The role of the law in upholding the moral standards of a society is well established. However, what the above example highlights is that occasionally, particular laws, and by extension, related moral norms are not always easy to justify. If an individual member of the committee was pressed as to why they were rejecting the petition, he/she would likely struggle to provide a reason. They may appeal to emotions as justifications, indeed the chair of the committee is reported as describing the idea as “abhorrent” (MacNab, 2016). If pressed, the committee member may have demonstrated a phenomenon in moral psychology known as “moral dumbfounding”.

Moral dumbfounding occurs when people fail to provide reasons for a strongly held moral judgement (Haidt, 2001; Haidt, Björklund, & Murphy, 2000). It typically occurs when people encounter taboo behaviours that do not result in harm (Haidt, 2001). As a phenomenon, it provides a unique insight into the making of moral judgement, and it has been cited as supporting evidence for various theories of moral judgement. Despite the possible implications of moral dumbfounding for the morality literature, little is really known about the phenomenon. There is limited empirical evidence demonstrating dumbfounding, and very few current theories of moral judgement offer an explanation.

The aim of this thesis is to address the limited understanding of moral dumbfounding in moral psychology. Chapter 1 presents what is currently known about dumbfounding, and traces the influence that moral dumbfounding has had on the morality literature. Chapter 2 discusses the limited explanations of moral dumbfounding and associated challenges to dumbfounding, both of which can be attributed to a lack of empirical evidence for dumbfounding. Chapter 3 examines whether or not dumbfounding is a real phenomenon, and materials and methods for measuring and studying moral dumbfounding are developed. New evidence for moral dumbfounding was found. Chapter 4 applies the methods developed in Chapter 3 to address specific challenges to the existence of dumbfounding. It was found that (a) people do not reliably articulate reasons for a judgement; (b) principles that may be guiding people’s judgements are not consistently applied across differing contexts. A possible dual-process explanation of moral dumbfounding is identified in Chapter 5 and the associated prediction that manipulations of cognitive load may have an effect on the prevalence of moral dumbfounding is tested. Chapter 6, testing another prediction of a dual-process explanation of moral dumbfounding, investigates if dumbfounding can be reduced by facilitating analytical thinking, or prompting participants with a reason. Chapter 7 draws on all of the empirical studies reported in previous chapters to review our current state of understanding of the phenomenon of moral dumbfounding in terms of extant theories of moral psychology. Limitations of the resulting understanding are identified. In response to limitations in our current understanding of moral dumbfounding identified in Chapter 7, an alternative theoretical outlook is broached, and possible avenues of future work are examined in Chapter 8.

1 Chapter 1 – Moral Dumbfounding and Moral Psychology

Moral dumbfounding is "the stubborn and puzzled maintenance of a judgment without supporting reasons" (Haidt et al., 2000, p. 2). It typically manifests as a state of confusion or puzzlement coupled with (a) an admission of not having reasons or (b) the use of unsupported declarations (e.g., "It's just wrong!") as justification for a judgement (Haidt et al., 2000; Haidt & Hersh, 2001), particularly, when people encounter taboo behaviours that do not result in any harm.

This chapter will provide the background to moral dumbfounding as a phenomenon in moral psychology. A brief account of the origins of, and evidence for dumbfounding will be provided. Specific issues arising from the paucity of empirical evidence for dumbfounding will then be outlined. The role of moral dumbfounding in the shaping of the morality literature will then be discussed alongside a critical summary of a number of key theories of moral judgement.

This discussion will begin with perhaps the most notable change in the morality literature since the discovery of moral dumbfounding, namely the growth of intuitionism over rationalism. Evidence for intuitionism over rationalism, beyond discussions of moral dumbfounding will be presented. Secondly, aside from the intuitionist-rationalist debate, interest in the linguistic analogy/universal moral grammar has also grown in recent years. Limitations of this approach highlighting its unsuitability for the study of moral dumbfounding will be identified.

Thirdly, the highly influential work of both Haidt (2001), and Greene will be discussed. Haidt's work led to the discovery of moral dumbfounding and his social intuitionist model of moral judgement was developed in direct response to this discovery. In this way the discovery of moral dumbfounding and the development of Haidt's social intuitionist model (SIM) of moral judgement may be seen as marking the beginning of the recent growth in intuitionist theories of moral judgement. Greene built on Haidt's work, proposing the earliest explicitly dual-process theory of moral judgement. Greene's work paved the way for the study of moral judgement to be aligned with dual-process theories of cognition more generally. The contributions of both Haidt and Greene had arguably the most significant influence on the development of theories of moral judgement over the past two decades. The theories as originally presented do not reflect developments of recent years and limitations of both theories will be discussed.

Finally, a number of more recent theories (dual-process theories e.g., Crockett, 2013; Cushman, 2013; model theory, Bucciarelli, Khemlani, & Johnson-Laird, 2008; skill and expertise approaches, e.g., Hulsey & Hampson, 2014; Narvaez & Lapsley, 2005; and categorisation approaches, e.g., Harman, Mason, & Sinnott-Armstrong, 2010; Prinz, 2005; Stich, 1993) will be discussed, with a particular focus on the implications of the existence of moral dumbfounding for these theories. The relative merits and weaknesses of these theories will also be discussed.

1.1 Moral Dumbfounding – Background, Evidence, and Issues

1.1.1 A brief history of moral dumbfounding. The earliest evidence for moral dumbfounding emerged indirectly as a result of a study by Haidt, Koller, and Dias (1993). This was a cross-cultural study examining the variability of the moral judgements of participants depending on age, socio-economic status, and nationality (USA or Brazil). Participants were presented with a range of moral scenarios, some of which were offensive, but harmless; for example, cutting up a national flag (Brazil or USA, matched to sample) and using it to clean the bathroom; a family eating their dog after it was killed by a car; and, a brother and sister kissing each other on the mouth. When asked to justify their condemnation of certain actions, some participants (from both countries) used unsupported declarations as a reason; for example, “Because it’s wrong to eat your dog” or “Because you’re not supposed to cut up the flag” (Haidt et al., 1993, p. 632). This study was not a direct study of moral dumbfounding, rather it was investigating differences in the way people reason about moral scenarios. The use of unsupported declarations in response to some moral scenarios was noted among a range of responses (Haidt et al., 1993).

A later study, by Haidt, Björklund, and Murphy (2000), directly investigated the phenomenon of moral dumbfounding. In their study, two moral scenarios (*Incest* and *Cannibal*: see Appendix A) designed to elicit strong emotional reactions, but with no identifiable harmful consequences (emotional intuition scenarios), were contrasted with a traditional moral judgement scenario (*Heinz*) that involved balancing the interests of two people (reasoning scenario). They observed differences in responses between the two types of scenarios, participants were better at defending their judgement for the reasoning scenario than for the emotional intuition scenarios. It appeared that these emotional intuition scenarios could elicit dumbfounding as evidenced by significant increases in (a) admissions of having no reasons for a judgement, or (b) the use of unsupported declarations (e.g., “it’s just wrong”) as a justification for a judgement (Haidt et al., 2000, p. 12). Although interesting, that study (consisting of a final sample of thirty participants) has not been published in peer reviewed form and has not been replicated.[1]

[1] The original Haidt, Björklund, Björklund and Murphy (2000) study has been published as a non-peer reviewed research report by Lund University as Björklund, Haidt, and Murphy, (2000). In the present paper we will follow the practice of the majority of authors discussing dumbfounding in focusing on the unpublished Haidt et al. manuscript, as it is freely available to download from the University of Virginia.

The following year, Haidt and Hersh (2001) investigated differences between conservatives and liberals, across a range of responses to moral issues, and found that conservatives produced more dumbfounded type responses (e.g., stuttering, stating “I don’t know”, admitting they could not explain their answers; Haidt & Hersh, 2001, p. 200), than liberals when discussing particular issues. Although this study did not investigate dumbfounding directly, the findings indicate that there may be individual differences that drive moral judgements which have not yet been fully investigated.

The phenomenon of moral dumbfounding has been widely discussed in the moral psychology literature (e.g., Cushman, 2013; Cushman, Young, & Greene, 2010; Cushman, Young, & Hauser, 2006; Hauser, Cushman, Young, Kang-Xing Jin, & Mikhail, 2007; Prinz, 2005; Royzman, Kim, & Leeman, 2015), but there is limited available empirical information about the nature of moral dumbfounding and the reliability with which it can be elicited in everyday human behaviour. Some authors have argued that moral dumbfounding does not really exist (Gray, Schein, & Ward, 2014; Jacobson, 2012; Sneddon, 2007; Wielenberg, 2014; see also Royzman et al., 2015).[2] This thesis is a detailed investigation of moral dumbfounding, and aims to address each of the limitations identified above. It aims to assess the contribution of dumbfounding towards the development of theories of moral judgement. It will empirically test for the existence of moral dumbfounding, and it will draw on existing theories of moral judgement to provide possible explanations of

[2] These are largely theoretical arguments offering explanations of dumbfounding that are consistent with a rationalist perspective (e.g., Kohlberg, 1971; Topolski, Weaver, Martin, & McCoy, 2013). However, Royzman, Kim, and Leeman (2015) present some empirical evidence in support of this position. This is discussed in more detail below, and in Chapter 4.

dumbfounding. These explanations will then be systematically tested across multiple studies.

1.1.2 Empirical evidence for dumbfounding. The existence of moral dumbfounding has had a considerable influence on the evolution of the moral psychology literature over the past two decades, shaping the development of theories of moral judgement (e.g., Cushman et al., 2010; Haidt, 2001; Hauser, Young, & Cushman, 2008; Prinz, 2005). Despite this pervasive influence, there remains limited empirical evidence for the existence of dumbfounding. The most cited demonstration (Haidt et al., 2000) only contained a final sample of N = 30, and has not been published in peer reviewed form. There has not been a direct replication of the original study,[3] though there has been a study which purports to demonstrate individual differences in susceptibility to dumbfounding, among a range of other responses (Haidt & Hersh, 2001). The inability to articulate principles consistent with a judgement has also been demonstrated (Cushman et al., 2006).

The limited empirical evidence for dumbfounding is problematic for three related reasons. Firstly, there does not appear to be any systematic means to elicit dumbfounding in a rigorous transparent way, and there is no defined agreed measure of moral dumbfounding. Secondly, and directly related, it is unclear how reliable dumbfounding is as a phenomenon. Thirdly, and following directly from this second problem, the existence of moral dumbfounding is widely accepted by many moral theorists and has informed the development of theories of moral judgement (e.g., Bucciarelli et al., 2008; Cushman et al., 2010; Haidt, 2001; Prinz, 2005). This means

[3] Recent work by Royzman, Kim, and Leeman (2015) includes a demonstration of dumbfounding using the incest scenario. This work is an attempt to identify possible reasons that may be guiding the judgement of participants and in limiting its focus to a single scenario (*Incest*), it is not classed here as a direct replication of the original work by Haidt et al. (2000).

that, whether or not dumbfounding is real, has serious implications for the moral psychology literature more generally. However, these implications cannot be addressed without addressing the first problem: measuring dumbfounding.

1.1.3 Defining and measuring moral dumbfounding. Definitions of moral dumbfounding vary within the moral psychology literature. It was originally defined as “the stubborn and puzzled maintenance of a judgment without supporting reasons” (Haidt et al., 2000, p. 2; see also, Haidt & Hersh, 2001, p. 194; Haidt & Björklund, 2008, p. 197). Some authors cite the original definition verbatim (e.g., Jacobson, 2012; Royzman et al., 2015); others include the maintenance of a moral judgement despite the absence of supporting reasons, but omit any reference to stubbornness or puzzlement (e.g., Cushman et al., 2006; Dwyer, 2009; Gray et al., 2014; Haidt, 2007; Wielenberg, 2014); and some refer to confidence in the judgement, but again, omit any reference to stubbornness or puzzlement (e.g., Cushman et al., 2010; Hauser et al., 2007, 2008; Pizarro & Bloom, 2003; Sneddon, 2007).

It is apparent from the literature that there is no single, agreed definition of moral dumbfounding. That said, an absence of reasons for, or an inability to justify or defend, a moral judgement, is consistently identified across definitions. However, even despite this apparent consistency, there remains considerable variation in the language used to describe this “failure to provide reasons for a moral judgement”. Indeed, the lack of definitional specificity has led to differing interpretations of moral dumbfounding. It also allows for the possibility of disagreement relating to the implications, both theoretical and practical, of moral dumbfounding.

According to the original definition, moral dumbfounding is “the stubborn and puzzled maintenance of a judgment without supporting reasons” (Haidt et al., 2000, p. 2). This definition contains four separate elements: (i) stubbornness; (ii) puzzlement; (iii) maintaining of the judgement; and (iv) the absence of supporting reasons. Of these individual elements, stubbornness and puzzlement, arguably, emerge as consequences of the combination of the maintenance of the judgement in the absence of supporting reasons. If a person maintains a judgement in the absence of reasons (and this absence of reasons has been pointed out to them) they will be perceived as stubborn; and, if a person becomes aware that they do not have reasons for their judgement, they may become puzzled.

Following this, and in line with the wider literature, the combination of elements (iii) and (iv), the maintenance of the judgement in the absence of supporting reasons are here identified as essential elements of dumbfounding. This does not mean that stubbornness and puzzlement should be ignored entirely; accounting for them may be useful in differentiating between a failure to provide reasons and a refusal to provide reasons. However, viewing stubbornness and puzzlement as consequences of the maintenance of a judgement in the absence of supporting reasons, indicates that they are subsequent to, and not a necessary part of, moral dumbfounding.

This view of dumbfounding includes the elements of the phenomenon that are mentioned the most frequently within the wider literature. It is also consistent with the way dumbfounding is described in the original study by Haidt et al. (2000). They report interesting variation in a number of non-verbal behaviours that may be linked with stubbornness or puzzlement, but beyond these, they do not offer a specific indication of how stubbornness and puzzlement are operationalised. Furthermore, other than appearing in the introductory definition for dumbfounding, in the abstract, (Haidt et al., 2000, p. 2), the terms “stubborn” and “puzzled” do not appear again for the remainder of the paper.

Haidt et al. (2000) report a range of responses that may illustrate a state of dumbfoundedness (admissions of not having reasons and unsupported declarations), however, they do not provide details of the numbers of participants they classified as dumbfounded, or a specific response that may be used to make such a classification. The numbers of participants who provided admissions of not having reasons are reported, however it is unclear whether or not this may be taken as a specific measure of dumbfounding or even if such a measure exists. This vagueness in the initial operationalisation of dumbfounding is reflected in the wider literature, whereby evidence of, or, illustrations of, dumbfounding include unsupported declarations (Haidt, 2001, p. 817; Prinz, 2005, p. 101), and tautological reasons (‘because it’s incest’; Mallon & Nichols, 2011, p. 285; see also Russell & Giner-Sorolla, 2011b for discussion of tautological reasons and disgust responses). The current research aims to identify specific measurable responses that may be used as indicators of dumbfounding.

In the work of Haidt et al. (2000), and the wider literature, the absence of supporting reasons appears to present in two distinct ways. Firstly, and non-controversially, participants may become aware that they do not have reasons and acknowledge this (admissions of not having reasons). Secondly, participants may fail to provide reasons. Measuring this failure to provide reasons is more problematic; if a participant does not admit to not having reasons, they attempt to disguise their failure to identify reasons. The use of unsupported declarations or tautological reasons as justifications for a judgement may be identified as a failure to provide reasons. Stating “it’s just wrong” or “because it’s wrong” does not answer the question “do you have a reason for your judgement?” (Mallon & Nichols, 2011,

p. 285).

Despite the limited evidence for moral dumbfounding, and the lack of clarity surrounding how dumbfounding should be measured, moral dumbfounding remains an important phenomenon in moral psychology. It is discussed in relation to, and has been cited as evidence for various theories of moral judgement (e.g., Bucciarelli et al., 2008; Cushman, 2013; Cushman et al., 2010; Dwyer, 2009; Haidt, 2001; Prinz, 2005). Furthermore, the discovery of moral dumbfounding (Haidt et al., 2000) coincided with, and arguably contributed to, the growth of intuitionist theories (e.g., Haidt, 2001) of moral judgement over rationalist theories of moral judgement (e.g., Kohlberg, 1969, 1985).

1.2 Moral dumbfounding and the Growth of Intuitionism over Rationalism.

The moral psychology literature has been long been characterised by a tension between intuitionism and rationalism (e.g., Cameron, Payne, & Doris, 2013; Hume, 2000/1748; Kant, 1959/1785; Nussbaum & Kahan, 1996). According to an intuitionist approach, our moral judgements are grounded in an emotional or intuitive automatic response rather than slow deliberate reasoning (Cameron et al., 2013; Crockett, 2013; Cushman, 2013; Cushman et al., 2010; Greene, 2008; Haidt, 2001; Prinz, 2005). In contrast, rationalism (as described by Haidt, 2001) posits that our moral judgements are grounded in reason, or discernible moral principles (Fine, 2006; Kennett & Fine, 2009; Kohlberg, 1971; Royzman et al., 2015).

The existence of moral dumbfounding is presented by Haidt (2001) as evidence against a rationalist perspective, in that, if moral judgements were grounded in reasons people would be able to provide reasons for their judgements. Some authors argue that a failure to articulate reasons does not necessarily provide evidence for the absence of reasons (e.g., Sneddon, 2007). Indeed, the linguistic analogy/universal moral grammar (e.g., Dwyer, 2009; Mikhail, 2007; discussed in more detail below) offers an explanation of moral dumbfounding based on this reasoning. Intuitionist approaches propose that the source of moral judgements lies in unconscious or automatic intuitions, such that the reasons for a judgement are not necessarily accessible to a person making a given judgement (Haidt, 2001; Haidt & Björklund, 2008). Moral dumbfounding serves as a demonstration of the inaccessibility of reasons for a judgement and such that the existence of moral dumbfounding is viewed as supporting evidence for intuitionist theories of moral judgement over rationalist theories (Cushman et al., 2010; Haidt, 2001; Hauser et al., 2008; Prinz, 2005).

In recent years there has been a growing acceptance of intuitionist approaches of moral judgement over rationalist approaches (Cameron et al., 2013). This growth of intuitionism over rationalism is supported by a large body of evidence beyond moral dumbfounding. Moral dumbfounding however, is more than just evidence for an intuitionist perspective, it is a clear and applied illustration of the intuitive nature of moral judgements. It provides a real life example of some of the practical implications of intuitionism. This illustrative power of moral dumbfounding was recognised and utilised by Haidt (2001). In his original paper introducing and defending his social intuitionist model of moral judgement he opens with the Julie and Mark scenario (Appendix A) and a discussion of moral dumbfounding (Haidt, 2001, p. 814). It is clear from reading Haidt (2001; Haidt & Björklund, 2008; Haidt & Hersh, 2001) that the existence of moral dumbfounding played a role in the development of his social intuitionist model.

1.2.1 Evidence for intuitionism over rationalism. Beyond moral dumbfounding, and the influential work of Haidt (2001 discussed in more detail below), there has been a growing acceptance of intuitionist approaches of moral judgement over the type of rationalist approaches described by Haidt (2001) in the past fifteen years or so (Cameron et al., 2013). In order for moral judgements to be viewed as rationalist, they must be stable and resistant to change from contextual influences other than reasons. Any study that shows variability or another type of contextual influence on moral judgement may thus be seen as support for intuitionist theories over rationalist theories. A large number of such studies has accumulated over the last decade and a half, amassing to a significant body of evidence supporting intuitionist theories over rationalist theories.

1.2.1.1 Context effects on moral judgements. Over the past number of years, various context effects on moral judgement have been identified. Studies of moral judgements often involve presenting participants with a scenario describing a behaviour and asking the participant to judge the behaviour. Two variants of the "Trolley" dilemma are particularly popular. Consider a trolley hurtling down a track towards five people, such that it will kill them all on impact. In one variant of this scenario (*Switch*), people are asked if it is permissible to flip a switch that will divert the trolley onto a side track. There is another person on this side track, who will be killed by the trolley if the switch is changed. In another variant of the scenario (*Push*), participants are asked if it is permissible to push a large man off a bridge, to intercept the runaway trolley. The impact will kill this man, but the trolley will be stopped and the five people will be saved. In both versions of this scenario the net result is the same: one person will die in the process of saving five lives. However, people are much more likely to agree with achieving this result by flipping a switch than by pushing a man (Cushman, 2013; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001).

In a third variation (*Loop*), a switch can be flipped to divert the trolley onto a separate track that loops back to the main track, such that successfully preventing the deaths of the five people by flipping the switch requires that there is an obstacle on the diverted section of track that will stop the trolley. In this variant, there is a bystander on the diverted section of track whose weight will stop the trolley. The net result, and the action committed, in both *Loop* and in *Switch* are the same, however, people are less likely to flip the switch in *Loop* than in *Switch* (Doris, 2010). It turns out that causing harm as a means to achieve a goal is generally regarded as more wrong than causing harm as a side-effect of achieving a goal. This is known as the doctrine of double effect.

In discussions of trolley dilemmas, some authors refer to a conflict between a utilitarian position and a deontological position (e.g., Greene, 2008). According to utilitarianism, the moral choice is the choice that maximises the positive outcomes (or minimises negative outcomes). Thus, for any variant of the trolley dilemma, this is the choice that minimises the net number of deaths, i.e., saving five people at the cost of one person. Deontology involves the following of specific rules (deontological positions). One such rule may be “do not kill/do not engage in an act that results in killing”. This means that the deontological choice for the trolley dilemma is the one that avoids an action that directly results in the killing of another person. Utilitarianism leads to action when the net result is more favourable, while deontology leads to inaction.

The variability in the making of moral judgements observed in studies of *Switch*, *Push*, and the doctrine of double effect is inconsistent with a rationalist perspective. Along with instances of moral dumbfounding, the variability noted in these scenarios provide evidence for, and have contributed to the emergence of intuitionism. If a person makes a judgement based on a moral principle (e.g., do not kill/minimise negative outcomes), this principle should be applied consistently. Contextual factors that are unrelated to the moral principles that purportedly govern moral judgements should not influence the application of these principles. Context effects on moral judgement that illustrate this limitation of rationalism are not limited to the doctrine of double effect. Three further classes of contextual factors (order effects, wording/framing/language effects, and emotional influences) that reliably influence the making of moral judgements are discussed below.

Beginning with order effects, a study by Lanteri, Chelini, and Rizello (2008) presented participants with both the *switch* (*lever*) and *push* (*stranger*) versions of the trolley dilemma. The order of presentation was varied and it was found that presenting *Push* first influenced judgements on *Switch* with fewer participants endorsing the pulling of the lever when than when the lever dilemma was presented first. The order of presentation had no effect on responses to the push version of the dilemma. Similar results were found by Lombrozo (2009), Petrinovich and O'Neill (1996), and by Nichols and Mallon (2006).

A study by Liao, Wiegmann, Alexander, and Vong (2012) investigated order effects on responses to the loop version of the trolley dilemma. They found that participants were more likely to judge action as acceptable in *Loop* when it was preceded by *Switch* than by *Push*. Wiegmann, Okan, and Nagel (2012) offer an insightful explanation of these findings suggesting that actions normally judged as acceptable are susceptible to order effects whereas actions that are normally judged as wrong are resistant to order effects. They conducted an experiment using a range of variants of the trolley dilemma and found this to be the case (Wiegmann et al., 2012), that actions normally judged as wrong are not as susceptible to order effects as actions that are normally judged as acceptable.

Schwitzgebel and Cushman (2012) demonstrated that philosophy professors are also susceptible to the influence of order effects. Furthermore, they also showed that the order of presentation of moral dilemmas influenced philosophy professors' subsequent endorsing of particular moral principles. That the making of moral judgements can vary by presenting moral dilemmas in a different order provides evidence in support of intuitionist approaches over a rationalist approaches. Such variability is inconsistent with a rationalist approach – if judgements were based on principles, they would be unaffected by the order of scenario presentation.

The way in which a question or moral dilemma is worded has been shown to influence the judgements made by participants. In a study by Petrinovich and O'Neill (1996), studying the switch variant of the trolley dilemma, two possible wordings of the question (one advocating action, and one advocating inaction) were used to include mention of either “death” or “saved”; e.g., (1) “Throw the switch, which will result in the death of the one innocent person on the side track,” and (2) “Do nothing, which will result in the death of the five innocent people” contrasted with (1) “Throw the switch, which will result in the five innocent people on the main track being saved,” and (2) Do nothing, which will result in the one innocent person being saved” (Petrinovich & O’Neill, 1996, p. 149). They found that participants were more likely to agree with statements containing “saved” than containing “death” for both types of statements, advocating action or inaction (Petrinovich & O’Neill, 1996).

The level of abstraction of the information provided to participants influences the judgements they make. In a study investigating attitudes towards people benefiting from genetic advantages Freiman and Nichols (2011) found that, when framed in an abstract way (e.g., "Suppose that some people make more money than others solely because they have genetic advantages" Freiman & Nichols, 2011, p. 127) participants did not report this was fair or deserved. However, when framed in terms of a concrete example, e.g., (a) comparing two jazz singers whereby one naturally has a better range due to genetics; or (b) comparing two jugglers, one of whom has better hand eye coordination due to genetics, participants reported it to be both fair and deserved for the genetically advantaged person to receive more money. A similar influence of level of abstraction was found by Nichols and Knobe (2007) on judgements of blame.

Other than minor manipulations of the wording of scenarios or questions, an interesting phenomenon known as "the foreign language effect" has also been identified, whereby people's judgements vary depending on whether they read a scenario in their first language or in a second language. Various authors have demonstrated that people appear to make more utilitarian judgements when they are presented with a scenario in their second language than if the scenario is presented in their native language (Costa et al., 2014; Geipel, Hadjichristidis, & Surian, 2016; Hayakawa, Tannenbaum, Costa, Corey, & Keysar, 2017). Again, and as with moral dumbfounding, this variability provides evidence for intuitionist theories over rationalist theories. If the making of judgements was grounded in moral principles they would not be susceptible to variability depending on changes in the wording, level of abstraction, or language of presentation.

The most widely identified contextual factor that influences moral judgement, and perhaps the most interesting for discussions of moral dumbfounding, is emotion. There are theories that link specific types of emotion to specific types of moral judgement (e.g., Chapman, 2018; Giner-Sorolla, 2018; Royzman, Atanasov, Parks, & Gepty, 2014; Rozin, Lowery, Imada, & Haidt, 1999; Russell & Giner-Sorolla, 2011a, 2013). Prinz (2005) draws on the work of Rozin et al. (1999) in proposing that moral dumbfounding occurs as a result of the disgusting nature of the behaviours in question. However, as yet, there is no rigorous way to empirically test this claim. There are also theories that identify a particular emotional component of specific types of moral judgement (e.g., Greene, 2008; Greene et al., 2001; Nakamura, 2013). There is also a large body of evidence documenting the role of incidental emotion, specifically incidental disgust, on the making of moral judgements (Greene, 2008; Haidt, 2001; May, 2014; Valdesolo & DeSteno, 2006; Wheatley & Haidt, 2005).

In a now classic study, Wheatley and Haidt (2005) hypnotically linked disgust with particular trigger words. Vignettes depicting moral scenarios were then presented to the participants and they were asked to make moral judgements on the characters/behaviours presented. It was found that judgements were harsher when the trigger word was present in the vignette. It was even found that the presence of a trigger word caused participants to moralise a morally neutral scenario, attributing the behaviour described to deceitful, self-serving motives.

Other studies have yielded similar effects for disgust. For example, Eskine, Kacinik, and Prinz (2011) provided participants with either a sweet beverage, a bitter beverage or water and asked to rate a series of moral transgressions. Again disgust was found to influence the judgements. These are just a sample from the many studies documenting the influence of incidental disgust in the making of moral judgements (Borg, Lieberman, & Kiehl, 2008; Cameron et al., 2013; David & Olatunji, 2011; Eskine et al., 2011; Rozin, Haidt, & MacCauley, 2009; Schnall, Haidt, Clore, & Jordan, 2008; Valdesolo & DeSteno, 2006). The reverse effect has also been found with a clean scented room promoting charity, reciprocity and trust (Eskine et al., 2011; Liljenquist, Zhong, & Galinsky, 2010; Zhong & Liljenquist, 2006; Zhong, Strejcek, & Sivanathan, 2010). As with the previous context effects, the emotional influences on the making of moral judgements identified above are problematic for rationalist approaches and provide evidence for intuitionist approaches over rationalist approaches.

1.2.1.2 Resistance to reasons. According to the rationalist perspective, people make moral judgements based on reason. A corollary of this is that people should not make judgements that they cannot justify through reasons. Furthermore, when presented with new evidence or reasons in support of an alternative position, people should revise their judgement. Neither of these appears to occur in real life.

Firstly, moral dumbfounding itself offers a clear case of participants defending a judgement that they cannot justify through reason. Typical examples of scenarios that lead to dumbfounding involve harmless taboos. Consider a researcher, Jennifer, working in a medical lab with bodies that have been donated for the general use and disposal at the discretion of researchers in the lab. Jennifer finds a body that is due to be incinerated the following day. Is it wrong for her to take home and eat a piece of meat from the body? Typically people judge this as wrong, however some people struggle to justify their judgement. This provides evidence that people have intuitions regarding what is right or wrong, and that these intuitions are not necessarily grounded in reason.

A recent study by Stanley, Dougherty, Yang, Henne, and De Brigard (2017) provides a clear case of the resistance of moral judgements to change based on alternative reasons. In their study, participants made a judgement on one of two moral dilemmas. They were then presented with either affirming reasons, opposing reasons, or both affirming reasons and opposing reasons. Following the presentation of reasons, participants presented with opposing reasons were the most likely to change their initial decisions. However, the numbers of participants who changed their judgements ranged from 2% or 5% of the sample depending on which dilemma was presented. This meant that even the participants who viewed opposing reasons were more likely to maintain their judgement than to change it following reading the opposing reasons. This finding provides further support for intuitionist theories over rationalist theories.

The resistance of moral judgements to change based on reasons, along with the extensive contextual variation described above suggests that moral judgements are intuitive rather than deliberative or rational. Over the past two decades a range of intuitionist theories of moral judgement have been proposed (Eden & Tamborini, 2016; Haidt, 2001; Sinnott-Armstrong, Young, & Cushman, 2010), however, extant theories of moral judgement at the time of the discovery of moral dumbfounding extended beyond the intuitionist-rationalist debate. One such theory of note is the linguistic analogy (Daniels, 1989; Dwyer, 2009; Hauser et al., 2008; Rawls, 1971) or universal moral grammar (Harman, 2000; Hauser, 2006a; Mikhail, 2007). This approach draws on apparent parallels between the emergence of language and the emergence of moral norms. Moral dumbfounding even provides an illustration of one of the parallels between morality and language, whereby people frequently apply grammatical rules that they cannot articulate.

1.3 Linguistic Analogy and Universal Moral Grammar

The linguistic analogy or universal moral grammar was originally proposed by Rawls (1971) but has been expanded by various theorists in recent years (e.g., Dwyer, 2009; Harman, 2000; Hauser, 2006a; Hauser et al., 2008; Mikhail, 2000, 2007). Drawing on Chomsky's work on generative linguistics (Chomsky, 1965, 1976, 2000) it is claimed that our capacity for moral judgement emerges through the same processes as the emergence of language. It is important to consider this approach in the study of moral dumbfounding for two reasons. Firstly, as a research programme it pre-dates the discovery of moral dumbfounding, and the intuitionist theories that followed this discovery (e.g., Haidt’s social intuitionist model; Haidt, 2001). Haidt (2001) does cite Rawls (1971; though this is part of a general comment on the prevalence of rationalism in modern philosophy Haidt, 2001, p. 816), and independently draws parallels between language and morality (Haidt, 2001, p. 826). However, the linguistic analogy or universal moral grammar is not acknowledged or directly discussed by Haidt (2001; Haidt & Björklund, 2008). This means that the possible contribution of this approach in the understanding moral dumbfounding may have been neglected. Secondly, and as noted by Dwyer (2009), the linguistic analogy does provide a possible explanation for moral dumbfounding.

It could be argued that the existence of moral dumbfounding provides evidence for the universal moral grammar (UMG), and that moral judgement can be explained by using the linguistic analogy (LA). Native speakers of a language successfully apply many complex grammatical rules. However, successfully applying a rule in context does not necessarily mean that these speakers can articulate the rule (Dwyer, 2009). Applied to the moral domain, this would imply that people can apply a moral rule (make a judgement) without being able to articulate why they made a particular judgement; this is what is observed in moral dumbfounding.

Moral dumbfounding can be explained well using the linguistic analogy (Dwyer, 2009). However, this explanation of dumbfounding relies on an implicit acceptance of what is essentially a rationalist perspective, i.e., that there are rules or principles that underlie our moral judgements. The contextual variation in moral judgements, described above, is inconsistent with this claim casting doubt on UMG as an approach to understanding moral dumbfounding.

The appeal of UMG rests on apparent parallels between the emergence of moral judgements and the emergence of grammar. However, despite apparent similarities, Dupoux and Jacob have identified a number of dis-analogues between morality and language (Dupoux & Jacob, 2007). They argue that (a) morality is evaluative, not generative (like language); (b) grammatical rules in language are domain specific, however, subtle contextual cues can place a particular action within or apart from the moral "domain" (deciding to flip a switch to divert a trolley is not normally a moral decision, however it can become a moral decision depending on the possible outcomes of flipping the switch); (c) the role of emotion is different in morality and in language. Regarding (c), the relative role of emotion in language versus in morality, according to Dupoux and Jacob (2007), a key claim in UMG made by Hauser (2006a, 2006b) is that "moral judgements cause emotions, but not vice versa" (Dupoux & Jacob, 2007, p. 376). This view is inconsistent with the research demonstrating the influence of incidental emotions on moral judgement, (Cameron et al., 2013; David & Olatunji, 2011; Huebner, Dwyer, & Hauser, 2009; May, 2014; Valdesolo & DeSteno, 2006). Dupoux and Jacob also draw on the work of Blair (1995; Blair, Jones, Clark, & Smith, 1997; Blair, Peschardt, Budhani, Mitchell, & Pine, 2006) to argue that empathy appears to have a causal role in the making of moral judgements.

Prinz (2008b) argues strongly against many facets of UMG. Firstly, Prinz notes that morality does not appear to have a "critical period" in the same way that language does (Prinz, 2008b, p. 158) citing case studies of children raised in isolation who do not present moral deficits in later life (Prinz, 2008b). Prinz (2008b) also highlights the differing roles of feedback in the learning of language versus morality. The learning of morals relies heavily on reward and punishment while this is not the case for language. Two other features of morality associated with UMG are innateness and universality. Prinz (2008a, 2008b) and Machery and Mallon (2010; Mallon, 2008) reject these claims noting that there is almost no evidence to suggest that morality is universal. In contrast, the variability of moral norms, and variability regarding what issues are considered “moral” has been widely observed, suggesting that the universality claim is false (Machery & Mallon, 2010; Prinz, 2008b). This lack of universality places considerable doubt on the innateness claim (Machery & Mallon, 2010, p. 34).

Finally, the linguistic analogy as described by Hauser et al. (Hauser et al., 2008), Dwyer (2009; Dwyer & Hauser, 2008), and Mikhail (Mikhail, 2007) appears to be almost exclusively grounded in Chomsky's work (Chomsky, 1965, 1976, 2000). This uncritical adopting of Chomsky's framework does not reflect the nuances of the wider linguistics literature and the various criticisms of Chomsky's framework (e.g., Christiansen & Chater, 2008; Hinzen, 2012; Tomasello, 2003, 2014). In uncritically adopting this theoretical framework for analogy, without acknowledging its limitations, UMG clearly has limited use in explaining the cognitive processes and underlying moral judgement.

Despite the importance and relevance of UMG/LA for discussions of moral dumbfounding the areas of concern described above limit its usefulness for the study of moral dumbfounding. The strength of evidence for UMG/LA has been challenged (Dupoux & Jacob, 2007, 2008; Machery & Mallon, 2010; Mallon, 2008; Prinz, 2008b). Furthermore, the approach relies too heavily on an uncritical acceptance of the work of Chomsky (1965, 1976, 2000) neglecting various other developments in linguistic theory (Christiansen & Chater, 2008; Hinzen, 2012; Tomasello, 2003, 2014).

1.4 Influence of Moral Dumbfounding on the Morality Literature

While the discovery of moral dumbfounding coincided with a renewed interest in UMG/LA it is unclear if this renewed interest in UMG/LA is related to the discovery of moral dumbfounding. On the other hand, the growth of intuitionist theories of moral judgement over the type of rationalist theories described by Haidt (2001) can clearly be attributed, at least in part, to the discovery of moral dumbfounding. This is most clearly evident in the development of Haidt’s social intuitionist model (SIM, Haidt, 2001; Haidt & Björklund, 2008). Haidt’s SIM is one of the two most influential theories from the early 2000s, the other being Greene's dual-process theory of moral judgement (Greene, 2008; Greene et al., 2001). Limitations of both Haidt’s SIM and Greene's dual-process theory of moral judgement are detailed below however, even in spite of these limitations, Haidt and Greene made arguably the most significant contribution to the moral psychology literature in recent years. Following the discovery of moral dumbfounding, Haidt initiated the growth in intuitionism that is still seen today (e.g., Eden & Tamborini, 2016; Gigerenzer, 2008; Jacobson, 2008; Sauer, 2017; Sinnott-Armstrong, 2008a; Sinnott-Armstrong et al., 2010), making specific reference to moral dumbfounding in introducing, illustrating, and defending SIM (Haidt, 2001). Greene (2008; Greene et al., 2001) does not discuss dumbfounding directly, however he draws on SIM and on Haidt’s work in defending his dual-process theory (Greene, 2008), the first theory of moral judgement that allowed for the morality literature to be aligned with dual-process theories of cognition more generally. These theories are not without their limitations, however they each made a significant contribution to the development of the moral judgement literature over the past two decades. This contribution is of particular importance in discussions of moral dumbfounding, given the extent to which dumbfounding may be seen as leading to their development. Each is taken in turn below.

1.4.1 Haidt's social intuitionist model (SIM). Haidt's primary claim in presenting his social intuitionist model is that moral judgements are caused by intuitions, and that moral reasoning is generally a post-hoc rationalisation of a judgement that has already been made (Haidt, 2001; Haidt & Björklund, 2008). The model itself is a descriptive account of the ordered sequence by which different factors influence a given judgement. The role of each of these links in generating moral judgements is described, however the details surrounding the cognitive processes involved and the underlying mechanisms are quite vague.

An overview of SIM is shown in Figure 1.1 (taken from Haidt, 2001, p. 815). The numbers refer to what Haidt calls links, where each link represents a different process. There are six of these links: (1) the intuitive judgement link, (2) the post-hoc reasoning link, (3) the reasoned persuasion link, and (4) the social persuasion link, (5) the reasoned judgement link and (6) the private reflection link (Haidt, 2001, p. 815). The order of these links reflects the order in which they occur.

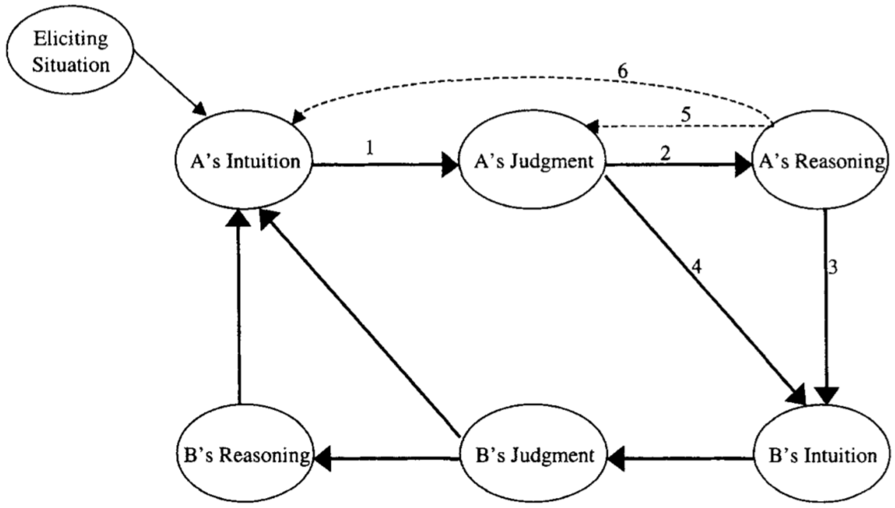

Figure 1.1: Links in SIM taken from Haidt (2001, p. 815)

The first two links are located within the individual. The intuitive judgement link (link 1), is the making of a moral judgement through intuition. Once a judgement has been made, it can be reasoned about, or rationalised: the post-hoc reasoning link (link 2).

Links 3 and 4 introduce the social element of SIM, describing two ways in which social influences that may affect moral judgements. Link 3, the reasoned persuasion link, relates to the use of reasons by one person to influence the judgement of another person. Link 4, the social persuasion link, relates to all other ways in which social factors may influence a person's judgement (e.g., conformity).

Links 5 and 6 relate the way in which reasoning may influence a person's judgement. Firstly, link 5, the reasoned judgement link, people may revise a judgement based on reasoning. Then, in link 6, the private reflection link, this revised judgement may become a revised intuition.

1.4.1.1 SIM as overly descriptive. The focus of SIM appears to be a coherent description of the variability of moral judgements rather than on providing a well developed theory of moral judgement, that accounts for the underlying mechanisms and provides testable predictions regarding judgements. This can be seen in an over-reliance on analogy and metaphor (e.g., intuitions as phonemes Haidt, 2001, p. 827; ‘the brain has a kind of gauge’, a ‘like-ometer’ Haidt & Björklund, 2008, p. 187, comparing moral rules and moral foundations to ‘cuisines’ and ‘taste receptors’ 2008, p. 202) where details of the underlying mechanisms and cognitive processes at play would be more appropriate.

Another example of the descriptive rather than explanatory approach apparent in SIM can be seen in the discussion on moral dumbfounding. Moral dumbfounding is presented as an interesting phenomenon that provides supporting evidence for SIM over rationalist theories, in that if reasons were guiding participants' judgements they would not be dumbfounded. However, Haidt does not offer any further explanation of dumbfounding and how or why it occurs.

1.4.1.2 The role of emotion in SIM. It is unclear whether the intuitions discussed in SIM are equivalent to emotions or distinct from emotions. Moral intuitions are frequently equated with moral emotions, often being referred to together to make the same point: “moral intuitions and emotions such as empathy and love ... and shame, guilt, and remorse” (Haidt, 2001, p. 825). The emotional content of the scenarios in the dumbfounding paradigm is cited as an explanation of moral dumbfounding. However, intuitions and emotions are also referred to separately, and therefore viewed as distinct, e.g., “Moral intuition, then, appears to be the automatic output of an underlying, largely unconscious set of interlinked moral concepts” (Haidt, 2001, p. 825). The confusion between emotion and intuition is particularly stark in the account outlining the development of the intuitions. The development of moral intuitions is consistently discussed with reference to emotions (e.g., teaching a child why a behaviour is wrong Haidt & Björklund, 2008, pp. 184–185). Haidt draws on the (now disputed, see Carter & Smith Pasqualini, 2004; Dunn, Dalgleish, & Lawrence, 2006; Maia & McClelland, 2004, 2005) somatic marker hypothesis (Bechara & Damasio, 2005; Bechara, Damasio, Tranel, & Damasio, 2005; Damasio, 1994), and on the "affect as information" hypothesis, in providing an account of the development of intuitions. Beyond appeal to emotion, and claims relating to socialisation, Haidt does not provide an adequate account of the underlying mechanisms that give rise to the emergence of intuitions. Yet, despite this apparent equivalence, intuitions also appear to be viewed as distinct from emotions, e.g., "Years of such implicit learning, coupled with explicit discussion, should gradually tune up intuition" (Haidt, 2001, p. 829).

In equating intuitions with emotions, the SIM as described Haidt and Björklund (Haidt, 2001; 2008) over-states the role of emotion in the making of moral judgements. They explain moral dumbfounding with reference to the emotional content of the intuition scenarios. There is a large body of evidence implicating emotion in the making of moral judgements (Cameron et al., 2013; Cannon, Schnall, & White, 2011; Eskine et al., 2011; Jones & Fitness, 2008; Schnall et al., 2008; Valdesolo & DeSteno, 2006). However, in recent years, caution has been advised in interpreting this (May, 2014), with some authors arguing for a clear distinction to be made between intuitions which have causal influence and emotions which do not (Huebner et al., 2009). Furthermore, a recent meta-analysis has called the effect into question entirely (Landy & Goodwin, 2015). Given these developments, the equivalence of moral intuitions and moral emotions apparent in SIM seems inappropriate. A more measured approach should account for the influence of emotions, while maintaining a clear separation of intuitions and emotions. This account should provide detail on the underlying mechanisms governing these intuitions and a coherent account of their development.

1.4.1.3 The role of reason in SIM. The SIM presents an inconsistent account of the role of reason in the emergence of moral judgements. Reasoning is described as almost exclusively post-hoc, with an extensive defence of the claim that reason does not play a causal role in moral judgements. Moral dumbfounding is cited in support of this claim. However in Link 5, the reasoned judgement link, it is suggested that logic can play a causal role in a person's judgement, “overriding their initial intuition” (Haidt, 2001, p. 819). Various authors have argued in favour of this latter interpretation of the role of reasoning moral judgement, (Fine, 2006; Kennett & Fine, 2009; Liao, 2011).

In discussing persuasion, Haidt is still committed to the claim that reasoning or reasons do not cause or change intuitions. Haidt refers to Martin Luther King Jr.’s “I have a dream” speech and argues that the success of the speech can be attributed the use of metaphor and imagery rather than logic. This use of metaphor and imagery enabled King to “trigger new intuitions” (Haidt, 2001, p. 823). This is echoed in a later work in which Haidt and Björklund describe the reasoned persuasion link in SIM, referring to persuasion as an attempt to “trigger the right intuitions in others” (Haidt & Björklund, 2008, p. 191). From this it appears that according to the SIM, successful persuasion is not grounded in reason, but rather the triggering of relevant intuitions. However, Haidt and Björklund also describe a case of dyadic reasoning leading to “new and better conclusions” in cases where “people are at least a little bit responsive to the reasons provided” (Haidt & Björklund, 2008,

p. 193).

The inconsistency surrounding the role of reason in SIM has implications regarding the mapping of SIM onto dual-process theories (e.g., Zajonc, 1980). Haidt draws on dual-process theories to support his claims regarding the intuitive nature of moral judgement. However as noted by Saltzstein and Kasachkoff (2004), reason plays a much greater role in these theories than in Haidt's SIM, where its primary role is the post-hoc rationalisation of intuitions.

1.4.1.4 Social influence. There is a tendency for the social aspect of SIM to be presented as a "catch-all" solution to any issues that are not addressed in the other links. This view is even encouraged by Haidt and Björklund (2008), stating: "Please don't forget the social part of the model, or you will think that we think that morality is just blind instinct, no smarter than lust" (p. 181). Despite the implication that the social links address any perceived weaknesses in the other links, the social aspect contains weaknesses of its own. For example, the claim that the social influences of coherence and relatedness bias our judgements has been challenged. Liao (2011), notes that for the most part, Haidt's discussion of social influence relates to the mutual influence between friends, whereby people are motivated to agree with friends, and the resultant tendency to agree with friends on moral issues (p. 10). Liao suggests that the nature of friendship means that people often have reasons to trust the judgements of friends, and that a tendency to agree with friends does not necessarily constitute bias. Liao argues that in proposing that moral judgements are biased by motivations to agree with friends, Haidt (2001) has conflated two distinct motivations to agree with friends: (a) a motivation to agree with friends to maintain a harmonious relationship; (b) a motivation to agree with friends because we tend to trust their judgements in general (Liao, 2011, p. 20). According to Liao, (a) may be considered biased, whereas (b) does not constitute bias.

A more serious issue with the social aspect of SIM has been noted by both Jacobson (2008) and Narvaez (2008) whereby, according to SIM, being moral involves adopting the morals of those around you. Saltzstein and Kasachkoff also note that social influence according to SIM is reduced to “overt compliance” (Saltzstein & Kasachkoff, 2004, p. 273). Jacobson (2008, p. 228) cites Haidt and Björklund's slogan: “A fully enculturated person is a virtuous person” (Haidt & Björklund, 2008, p. 216). In this view, habituated conformity and enculturation has been conflated with morally virtuous behaviour. Narvaez (2008) refers to instances and immoral conformity (Nazi soldiers) to cast doubt on this claim.

In spite these weaknesses detailed here, the overall contribution of SIM to the morality literature must be appreciated. The introduction of SIM revived and legitimised an intuitionist approach to the study of moral psychology. This prompted a growth in intuitionism which has resulted in a range of important discoveries about the nature of moral judgement (e.g., variability, the influence of emotion on judgements, wording/order effects). As noted previously, the influence of moral dumbfounding on the development of SIM is clear. Later theories do not draw as heavily on moral dumbfounding, though it is widely acknowledged as consistent with, and seen as evidence for intuitionism, either through explicit reference to the original study (e.g., Cushman et al., 2010; Prinz, 2005), or through reference to Haidt’s (2001) seminal paper (e.g., Cushman, 2013; Greene, 2008).

Where the influence of SIM may be viewed as contributing to the emergence of general intuitionist approaches (Eden & Tamborini, 2016; Sauer, 2017; Sinnott-Armstrong, 2008a; Sinnott-Armstrong et al., 2010), the work of Greene (2008; Greene et al., 2001) may be seen as giving rise to the emergence of dual-process theories of moral judgement. Given the dominance of dual-process theories of moral judgement in the modern moral psychology literature (Brand, 2016; Crockett, 2013; Cushman, 2013; see also: Doris, 2010; Sinnott-Armstrong, 2008b, 2008c, 2008d; Christensen, Flexas, Calabrese, Gut, & Gomila, 2014), the contribution Greene made is of particular importance. Though Greene does not draw on moral dumbfounding directly, there are clear parallels between Haidt's SIM and Greene's dual-process theory. Furthermore, Greene (2008) draws extensively on Haidt's work, both theoretical and empirical (explicitly admitting to drawing on Haidt's 'insights'; Greene, 2008, p. 36). Given the influence of Haidt on Greene's work, and the degree to which Haidt's insights were informed by the existence of moral dumbfounding, it is reasonable to argue that the development of Greene's dual-process theory of moral judgement was shaped, at least in part, by the existence of moral dumbfounding. Greene's dual-process theory (Greene, 2008; Greene et al., 2001) is certainly consistent with the existence of moral dumbfounding, and may even provide an explanation for it.

1.4.2 Greene's dual-process theory of moral judgement. In many ways, Greene's dual-process theory of moral judgement (Greene, 2008; Greene et al., 2001) may be seen as an improvement of SIM (Haidt, 2001; Haidt & Björklund, 2008), in that a number of the weaknesses of SIM outlined above are not present in Greene's theory. Greene does not conflate emotion and intuition, rather, Greene commits to the claim that some judgements are grounded in emotion, while others are grounded in "cognition" (reason). In his model, the emergence of moral intuitions is not attributed to evolutionary modules and the role of reason is clearly defined. There is also no attempt to attribute unexplained phenomena to generalised social influence.

The thrust of Greene's work can be traced to an important insight offering an empirically grounded explanation of the trolley problem (Foot, 1967; Thomson, 1976, 1986). The trolley problem refers to the interesting phenomenon in moral psychology, whereby people make different judgements on similar scenarios in which the eventual outcome of both scenarios is the same. Recall the *Switch* and *Push* variants of the trolley dilemma described in section 1.2.1.1. The possible outcomes in each is the same: action saves five people but kills one; inaction allows five people to be killed by a runaway trolley but allows one person to live. The difference between these dilemmas is the type of action required; in *Switch* the action is the flipping of a switch to divert the trolley, whereas in *Push* the action is the pushing of a large man off a bridge to stop the runaway trolley. In both versions of this scenario the net result is the same: one person will die in the process of saving five lives, however, people are much more likely to agree with the actions in *Switch* than in *Push* (Cushman, 2013; Greene et al., 2001). Greene (2008; Greene et al., 2001) explained this variation in terms of the relative emotional content of the two scenarios, where the content of *Push* is more emotionally loaded than that of *Switch*. From this, the *Push* version of the dilemma was identified as, moral-personal, while the *Switch* version was identified as moral-impersonal, in that, *Push* is more “up close and personal”, involving direct contact with the victim (Greene, 2008, p. 43).

Drawing on this moral-personal/moral-impersonal distinction that appeared to be guiding the decisions of participants in the trolley dilemma, Greene et al. (2001) tested if it generalised across different scenarios. Greene et al. (2001) compiled a battery of 60 vignettes, categorised as moral/non-moral, and up close and personal/intuitively impersonal. They found that for the moral personal dilemmas, participants took longer to endorse a utilitarian (consequentialist) action than to reject it (e.g., took longer to agree to pushing the man off the footbridge than to reject pushing the man). From this they identified the moral-personal/moral-impersonal distinction as a key variable in influencing how the judgement is made. They claimed that emotion is implicated in the moral-personal dilemmas, whereas “cognition” is implicated in moral-impersonal dilemmas. Furthermore, the type of judgement made is linked to whether emotion or cognition is involved, with emotion leading to deontological judgements, and cognition leading to consequentialist judgements.

Moral dumbfounding may be understood in terms of this emotion-cognition distinction. According to this interpretation, the moral dumbfounding scenarios elicit a strong emotional reaction, leading participants to make a deontological judgement (condemning the behaviour because the behaviour is wrong) rather than a utilitarian/consequentialist judgement (rating the behaviour as not wrong because there were no negative consequences). As such, Greene's dual-process theory of moral judgement is both consistent with, and provides a possible explanation of moral dumbfounding.

There are three primary weaknesses in Greene's model. Firstly the mapping of deontology and consequentialism to emotion and cognition is contrived, inconsistent, and most likely due to coincidence rather than underlying mechanisms. Secondly, the emotion-cognition distinction is overly simplistic and does not reflect the wider dual-processes literature. Finally, and most worryingly, the moral-personal/moral-impersonal distinction provides a foundation for Greene's other claims is not supported by his own data (McGuire, Langdon, Coltheart, & Mackenzie, 2009). Let us take each of these in turn.

1.4.2.1 Deontological/utilitarian distinction. According to Greene's model, when we make judgements grounded in emotion, we perform as deontologists, and when we use “cognition” to make a judgement we perform as consequentialists. The theoretical and practical implications of this finding are unclear, and an explanation is not provided. A crude reading of Greene's claims is that the type of rationalism rejected by Haidt is correct, but that there are two competing sets of moral principles that guide our moral decision making, and Greene has identified emotional content as the moderating factor influencing which set of principles governs a given judgement. According to this view, emotion is the single contextual variable that influences moral judgements, however this is clearly not the case (e.g., Freiman & Nichols, 2011; Nadelhoffer & Feltz, 2008; Petrinovich & O’Neill, 1996; Schwitzgebel & Cushman, 2012). An alternative interpretation, is that this is simply an interesting coincidence arising from the different ways in which particular types of moral principles are treated and taught in our society.

Consider, for example, a moral expertise account of moral judgement (e.g., Dreyfus & Dreyfus, 1990; Hulsey & Hampson, 2014; Narvaez & Lapsley, 2005). According to this view, moral judgements are grounded in the past experience and learning history of the individual. People become skilled at making judgements that are made consistently. Specific associated contextual factors may become linked with the making of a particular judgement as a result of consistent co-occurrence of these factors and a particular judgement. This approach may offer an explanation of Greene's mapping of emotion onto deontological principles whereby throughout the learning of a deontological principle, it is consistently associated with an emotional component. It becomes a socialisation feedback loop. A child may be told that a particular behaviour is “disgusting”, and that it is wrong to engage in that behaviour (a deontological proposition). The child may associate this behaviour with shame having caused disgust in parents. In an effort to avoid causing shame again, the child may adopt the position of the parents, that the behaviour is wrong and disgusting. This cycle repeats then when the child grows up and has children of his/her own. On the other hand, the learning of utilitarian positions (e.g., minimising net number of deaths) may involve more abstract discussion of issues that are removed from direct experience; and this may occur at a later age. A deontological position may emerge as “grounded in emotion” and a utilitarian position may emerge as “grounded in cognition”, and this occurs as a result of the way in which a given principle is learned. It is possible that some (possibly less well known) deontological positions may become grounded in “cognition” and that some utilitarian positions may become grounded in emotion. For example, the utilitarian “fair distribution of resources” may become grounded in emotion for a person that grew up in a large family, where they had to fight over sweets, toys, or other resources siblings may have to divide amongst themselves. Such a situation would undermine Greene's emotion/cognition – deontological/utilitarian distinction.

1.4.2.2 Emotion/cognition distinction. Greene's emotion-“cognition” distinction frames System 1 (intuitive) type processes as grounded in emotion. This is problematic for the same reasons it was problematic in SIM. Unlike Haidt, Greene is at least consistent in identifying intuitions as emotional. However, the second criticism, that the role of emotion in SIM is over-stated, also applies to Greene, particularly in view of developments in recent years. May (2014) argues that the effects described in studies of incidental emotion and moral judgement do not support provide evidence that incidental emotions can change a judgement. In one example (Wheatly & Haidt, 2005), incidental disgust was led to small differences in responses on a Likert scale. However these differences did not cross the midpoint of the scale, and therefore the there was no real difference in the valence of participants' judgements between experimental (incidental disgust) and control (no incidental disgust) groups. Furthermore, Landy and Goodwin (2015) conducted a meta-analysis on studies of incidental disgust and moral judgements and found that when controlling for publication bias, the influence of disgust on moral judgements disappeared. The evidence for the role of incidental emotion in the making of moral judgements therefore does not support Greene's (2008) position that the making of moral judgements can be attributed to emotion.

1.4.2.3 Re-analysing Greene's data. Perhaps the most striking evidence against Greene's dual-process theory of moral judgement comes from an independent re-analysis of his own data by McGuire et al. (2009). Recall that Greene's theory is grounded in the moral-personal/moral-impersonal distinction identified in Greene et al. (2001). However the reliability and significance of this distinction has come under threat in recent years (Christensen et al., 2014; Christiansen & Chater, 2008; McGuire et al., 2009; Mikhail, 2007).

A number of issues with the materials used by Greene et al. (2001) are identified by McGuire et al. (2009). The consistency in use of emotive language in personal/impersonal dilemmas is questioned, along with questions relating to the ambiguity of the questions participants answered (were actions 'appropriate' Borg, Hynes, van Horn, Grafton, & Sinnott-Armstrong, 2006; McGuire et al., 2009).

On inspection of the vignettes used, McGuire et al. (2009) identified a number of items that did not appear to truly constitute dilemmas, in they were consistently poorly endorsed by participants. Items endorsed by less than 5% of participants were removed from the analysis. Following this the effect initially described by Greene et al. (that moral personal scenarios led to deontological judgements, and moral impersonal scenarios led to utilitarian judgements, 2001) completely disappeared. It is apparent from this that the results presented by Greene et al. (2001) can be attributed to a small number of outlier vignettes rather than on a fundamental distinction between moral-personal and moral-impersonal.

The emotion/cognition distinction and the associated mapping to deontological/consequentialist moral principles in Greene's dual-process theory (Greene, 2008; Greene et al., 2001) was grounded in the moral-personal/moral-impersonal distinction. However, given that this moral-personal/moral-impersonal distinction does not appear to generalise to moral judgements beyond the Trolley dilemma, the assumptions resting on this distinction appear unsupported.

1.4.2.4 Beyond the personal/impersonal distinction. The moral-personal/moral-impersonal distinction is one specific contextual influence that may affect moral judgement in specific circumstances. It is increasingly apparent that Greene was probably mistaken to develop his a theory based on this single influence. Since the publication of Greene et al.'s influential paper (2001), various other factors affecting moral judgement have also been identified. For example Mendez, Anderson, and Shapira (2005) and Valdesolo and DeSteno (2006) identified level of physical contact involved as influencing judgements. Cushman et al. (2006) have shown that people are sensitive to whether or not harm is a foreseen side-effect of an action or a means to a particular end. Following this, Christensen and Gomila (2012) re-categorised the vignettes used by Greene et al. (2001) to include three additional factors: intentionality (was harm intended as a means or foreseen consequence), evitability (was harm avoidable), benefit recipient (did the harmed party benefit). In a later study (Christensen et al., 2014) these three factors were combined with the original personal-impersonal distinction and it was found that participants were indeed sensitive to each of the factors.

Greene's dual-process theory of moral judgement is clearly limited by focusing on a single factor identified as influencing moral judgement. Each new factor that is discovered poses an additional challenge to the Greene's theory. Importantly, a theory of moral judgement grounded in a content specific influence does not provide a coherent theory of moral judgements more generally. Theories of moral judgement should instead investigate the underlying mechanisms that give rise to the making of moral judgements, with an awareness that these judgements may be susceptible to a range of contextual influences.

1.5 The Development and Incorporation of Intuitionism in Moral Theories

Despite the limitations identified above in both Haidt's (Haidt, 2001; Haidt & Björklund, 2008) and Greene's (Greene, 2008; Greene et al., 2001) work, their contribution to the moral judgement literature should not be overlooked. Haidt, following from his discovery of moral dumbfounding, pioneered the growth of intuitionist theories of moral judgement still evident today (e.g., Eden & Tamborini, 2016; Gigerenzer, 2008; Jacobson, 2008; Sauer, 2017; Sinnott-Armstrong, 2008a; Sinnott-Armstrong et al., 2010). Greene, building on the work of Haidt, offered the first theory of moral judgement that allowed for the aligning of the morality literature dual-processes in cognition more generally. This alignment allows for the possibility of a dual-process explanation of moral dumbfounding.

1.5.1 Dual-process theories of moral judgement. Since Greene, dual-process theories have become a standard in moral psychology (e.g., Brand, 2016). Throughout the development of the various dual-process theories, the influence of both Haidt and Greene remained present. Furthermore, the existence of moral dumbfounding continues to be consistent with the various dual-process theories of moral judgement that have evolved over the years. Some theorists continue to cite dumbfounding as evidence for (e.g., Cushman, 2013; Triskiel, 2016), though the immediate influence of moral dumbfounding on these later theories is not as strong.

Where Greene's (2008) dual-process theory centred around a distinction between emotion and cognition, both Cushman (2013) and Crockett (2013) distinguish between "model-based" and "model-free" (Crockett, 2013, p. 363; Cushman, 2013, p. 277) processes. This model-based/model-free distinction may be interpreted in terms of existing distinctions in the dual-process literature more generally. Such characterisations include: intuitive or heuristic versus analytic (e.g., Chaiken, 1980; Evans, 1989, 2006, 2007; Hammond, 1996) automatic versus controlled (e.g., Schneider & Shiffrin, 1977) experiential/rational (e.g., Epstein, 1994; Epstein & Pacini, 1999) implicit or tacit/explicit (Evans & Over, 2013; Reber, 1989), or associative versus rule-based (for review, see Evans, 2008; Sloman, 1996; Smith & DeCoster, 2000).

The model-based/model-free characterisation of dual-processes proposed by both Cushman (2013) and Crockett (2013) is not subject to the weaknesses of SIM (Haidt, 2001) described above. Where SIM was identified as overly descriptive, and failing to account for underlying mechanisms, both Cushman (2013) and Crockett (2013) provide an account for the learning of both model-based and model-free responses. Furthermore, where intuitions were conflated with emotions in SIM, model-free processes are clearly distinct from emotions. In placing model-free processes as clearly distinct from emotions, the over-reliance on emotion present in both SIM (Haidt, 2001) and Greene's dual-process theory (2008) is not present in

Cushman's (2013) and Crockett's (2013) theories. This also means that the role of reason (model-based processes) is much more clear in these approaches than in SIM. Finally, neither Cushman (2013), nor Crockett (2013) presents "social influence" as a the type of catch-all it appears as in SIM. These strengths, along with the aligning of this approach with dual-process theories of cognition more generally make this model-based/model-free dual-process theory of moral judgement a useful theory for the study of moral dumbfounding.

Using this model-based/model-free distinction, Cushman (2013) and Crockett (2013) offer an alternative interpretation of the trolley problem. Where Greene (Greene, 2008; Greene et al., 2001) mapped emotion to deontological and judgements and cognition to utilitarianism, Cushman (2013) and Crockett (2013) map model-free processes onto action based decisions, and model-based processes onto outcome based decisions. Applying this to the trolley problem, they propose that action of pushing has been learned, through reinforcement history, to be seen as wrong. Conversely, the act of flipping a switch does not have the same reinforcement history identifying it as wrong, which means that, in the switch variant of the trolley dilemma, people are not confronted with an act that is morally loaded. According to both Cushman and Crockett (Crockett, 2013; Cushman, 2013), this means that the model-free system does not interfere with the judgement of the model based system, which allows people to make judgements based on outcomes.

1.5.1.1 Limitations of the action/outcome distinction. For the purposes of the current discussion, there are two related issues with the approaches proposed by Cushman (2013) and Crockett (2013). The first relates to a key limitation of the action/outcome distinction regarding the doctrine of double effect. The second relates to a more general failure to account for other influences on moral judgement.

On the surface, this distinction between actions and outcomes appears to work well. However, it is possible that it presents as an over-generalisation, that, while appropriate in the majority of cases, does not reflect every reality. To illustrate this, consider the doctrine of double effect (Cushman et al., 2006; Doris, 2010; Mikhail, 2000), whereby causing harm as a means to achieve a goal is regarded as more wrong than causing harm as a side-effect of achieving a goal. This occurs even when the action involved in each scenario is the same.

Mikhail (2000) presented participants with two looped versions of the trolley dilemma (a switch diverted the trolley onto a loop of track that rejoined the main track and continued towards the five people). In one version, there was a large man standing on the track who would stop (slow) the trolley if it hit him, in the other version the large man is standing in front of a large object that will slow/stop the trolley. The outcome in each case is the same, and the action in each case is the same, however in version one the death of large man serves as the means to save the people, whereas in version two the death of the large man is an unavoidable side effect of saving the people. It was found that people view the means version as worse than the side effect version. It is clear from this illustration that the action/outcome distinction does not account for the doctrine of double effect.

Recall that Greene's moral-personal/moral-impersonal distinction was problematic because it failed to account for various other influences on moral judgement. The action/outcome distinction is problematic for the same reason. There are at least five known factors that influence moral judgements: (1) intentionality (Christensen et al., 2014; Christensen & Gomila, 2012) or doctrine of double effect (Cushman et al., 2006; Doris, 2010; Mikhail, 2000); (2) evitability (Christensen et al., 2014; Christensen & Gomila, 2012); (3) benefit recipient

(Christensen et al., 2014; Christensen & Gomila, 2012); (4) the personal-impersonal distinction (Christensen et al., 2014; Christensen & Gomila, 2012; Greene et al., 2001); and (5) level of physical contact (Mendez et al., 2005; Valdesolo & DeSteno, 2006). It seems then that, rather than studying in isolation the influence of the action/outcome distinction (Crockett, 2013; Cushman, 2013) on moral judgements, it should be added to this list of factors that are known to influence moral judgements, and studied as “one of many” factors.

One of the key strengths of Cushman's (2013) and Crockett's (2013) work is that they can be mapped onto dual-process theories of judgement more generally. This means that the vast body of research on dual-process theories of judgement can be applied to the moral domain. This can offer important insights in terms of both explanatory power and identifying testable predictions for furthering understanding. Interestingly, particularly for the current purposes, the existence of moral dumbfounding is consistent with both approaches, with Cushman (2013) making explicit reference to the dumbfounding paradigm.

1.5.1.2 Limitations of a dual-process approach. A theory of moral judgement that is grounded in dual-processes, is subject to the same criticisms levelled at dual-process theories of cognition (e.g., Mugg, 2015). One such criticism, stems from a criticism of a dual-systems interpretation of cognition, whereby automatic and controlled processes reside in different systems. Mugg cites a growing body of evidence suggesting that the distinguishing features of System 1 and System 2 “crosscut” each other (Mugg, 2015, p. 2) such that identifying System 1 and System 2 as distinct kinds has become problematic. A softer dual-processes interpretation has been widely adopted in response to this crosscutting of features (Evans, 2008, 2011; Mugg, 2015). Mugg, however, argues that this re-labelling as processes as opposed to systems does not address the issue (Mugg, 2015).

A reinterpretation of the distinctions between the processes may alleviate this criticism. Recall the various characterisations of dual-processes identified above: intuitive or heuristic versus analytic, automatic versus controlled, experiential/rational, implicit or tacit versus explicit, associative versus rule-based (Evans, 2008). An alternative habitual/deliberative distinction may be identified by drawing on Barsalou's research on the development of automaticity in categorisation (Barsalou, 1999, 2003, 2005, 2008). Under this interpretation, responses that appear to be grounded in automatic or implicit processes are responses that have become highly skilled or habitual as a result of prior learning and rehearsal through experience. This learning and rehearsal may occur either explicitly or implicitly. As automaticity develops the rehearsal becomes increasingly implicit, and the response becomes habitual. Barsalou developed his account with specific reference to the emergence of categorical knowledge, however the view of categorical knowledge adopted by Barsalou (dynamical, contextualised, and goal-derived as opposed to stable, taxonomic and hierarchically organised, e.g., Barsalou, 2003) is so broad that seemingly all knowledge may be framed in terms of this classification of categorical knowledge (this is explored in more detail with specific reference to moral knowledge in Chapter 8). According to this approach, our intuitions constitute knowledge that has been acquired over time, to the point where it has become automatic or implicit, or habitual, while responses that appear to be grounded in controlled, rational, or rule-based processes are responses for which deliberation is required. Deliberation is implicated when novelty is encountered, e.g., an unfamiliar situation, or a situation that elicits competing or conflicting intuitions (habitual responses). Furthermore, it may also be involved in situations where a judgement is required to be defended (Haidt, 2001; recall Haidt's reasoned persuasion link Haidt & Björklund, 2008). This habitual/deliberative interpretation of dual-processes places habitual responding and deliberative responding at opposing ends of a continuum as opposed to distinct separable processes.

The habitual/deliberative distinction addresses the concerns raised by Mugg (2015) in two ways. Firstly, this distinction does not posit different systems, or different processes, rather, the habitual/deliberative distinction positions two types of responding on opposite ends of a continuum. Secondly, deliberative responding is always supported by habitual responding to some degree, in that our prior knowledge (intuitions) support deliberation. This distinction between habitual and deliberative responses is consistent with the distinction between automatic and controlled processes that is made in the dual-process literature more generally (Evans, 2008; Schneider & Shiffrin, 1977). This means existing research on dual-processes may be interpreted using the habitual/deliberative distinction, and as such, acknowledging the criticisms of the criticisms of Mugg (2015) does not entail a complete rejection of the wider dual-process literature.

1.5.2 Intuition versus reason. The original characterisation of intuitionism versus rationalism by Haidt has had a lasting influence on the wider morality literature. Haidt rejected reason, and by extension rejected the associated reasoning literature. Greene pitted emotion against "cognition" in his dual-process theory, severely limiting the scope and explanatory power of the theory. More recent theories draw on the wider literature on learning and automatic processes, in developing more sophisticated accounts of the development of intuitions. This work has received more prominence than work that incorporates aspects of the reasoning literature more generally into theories of moral judgement. Despite this, such work is being conducted (e.g., Bialek & Terbeck, 2016; Bucciarelli, 2009; Bucciarelli & Johnson-Laird, 2005; Bucciarelli et al., 2008; Johnson-Laird, 2006; Juhos, Quelhas, & Byrne, 2015). A true picture of how we make moral judgements should abandon the distinction that pits intuition against reasoning, and incorporate insights from both literatures.

1.5.3 Reasoning and moral judgement – model theory. One attempt to reconcile an intuitionist approach to morality with the reasoning literature has been made by Bucciarelli, Khemlani, and Johnson-Laird (2008; see also Bucciarelli & Daniele, 2015). In presenting their model theory they explicitly endorse many facets of both Haidt (2001), Greene's (2001) work, along with elements of the UMG (Mikhail, 2007). The importance of model theory for this thesis is twofold. Firstly, model theory adopts a dual-process perspective, and is therefore consistent with dual-process approaches more generally. This is important given the dominance of dual-process approaches to moral judgement (e.g., Brand, 2016; van den Bos, 2018), and the popularity of dual-process theories of cognition more generally (Chaiken & Trope, 1999; De Neys, 2006; Evans, 2010, 2011; Sun, Slusarz, & Terry, 2005). Secondly, in presenting model theory Bucciarelli et al. (2008) explicitly discuss how a phenomenon like dumbfounding may occur as a direct consequence of the way we reason. This discussion is supported with reference to Haidt (2001) and to moral dumbfounding, however of particular interest for the current work is that possibility of dumbfounding is presented as a prediction of model theory. Furthermore, Bucciarelli et al. (2008) predict that dumbfounding may occur beyond the moral domain, even offering examples of situations where non-moral dumbfounding might occur.

According to model theory, the making of moral judgements occurs through "reasoning from unconscious premises to conscious conclusions" (Bucciarelli & Daniele, 2015, pp. 268–269), which are represented as mental models. The four principles of moral theory as outlined by Bucciarelli et al. (2008) are as follows: (1) indefinability: there is no principle or definition that distinguishes moral issues from other deontic matters; (2) independent systems: deontic evaluations and emotions are based on independent systems that operate in parallel; (3) deontic reasoning: deontic evaluations depend on inferences in the form of unconscious intuitions or conscious reasoning; (4) moral inconsistency: moral beliefs are neither complete nor consistent. Each principle is described in more detail below.

The principle of indefinability states that "No simple principled way exists to tell from a proposition alone whether or not it concerns a moral issue as opposed to some other sort of deontic matter" (Bucciarelli et al., 2008, p. 125). Put simply, this means that there is no definable boundary to the moral domain. There are no features of a moral transgression that distinguish it from transgressions of non moral social norms (e.g., etiquette rules). Bucciarelli et al. (2008) provide a number of counter examples to purported definitional boundaries to the moral domain. For example, it has been claimed that the requirement of punishment is unique to the moral domain (e.g., Davidson, Turiel, & Black, 1983). Bucciarelli et al. (2008) counter that many immoral acts do not warrant punishment, and that this criterion utterly fails in identifying morally good actions. Bucciarelli et al. (2008) drawing on Nichols (2002) reject emotion as a possible criterion for distinguishing moral issues from non-moral issues, noting that in some situations, breaking etiquette norms can be more disgusting than stealing a paperclip.

One of the key implications of the principle of indefinability for the study of moral judgement is that it makes the notion of a dedicated moral mechanism within the brain highly unlikely. Bucciarelli et al. (2008) make the point that if there was a dedicated moral mechanism, then there would need to be a clear way to identify when this mechanism would apply. However no criterion has been identified yet, supporting the claim that moral reasoning occurs by the same mechanisms as reasoning about non-moral issues.

The principle of independent systems states that "Emotions and deontic evaluations are based on independent systems operating in parallel" (Bucciarelli et al., 2008, p. 126). This means that the claims that (a) the making of moral judgements is grounded in emotion (e.g., Greene, 2008; Haidt, 2001; Prinz, 2005), or (b) that moral evaluations give rise to emotions (e.g., Hauser, 2006a), are both incomplete. Incidental emotions have been widely shown to influence moral judgements (e.g., Cameron et al., 2013; Wheatley & Haidt, 2005), and moral judgements have also been shown to elicit emotions (e.g., Royzman, Atanasov, et al., 2014; Rozin et al., 1999). Locating emotion and moral judgement in independent but parallel systems allows for this bidirectional relationship.

The principle of deontic reasoning states that "all deontic evaluations including those concerning matters of morality depend on inferences, either *unconscious intuitions* or *conscious reasoning*" (Bucciarelli et al., 2008, p. 127 emphasis added). This principle places model theory firmly within the broader suite of dual-process theories more generally. According to this principle, people have unconscious intuitions that give rise to moral judgements, and people can use conscious reasoning to arrive at a moral judgement. This is particularly evident when people are presented with dilemmas that lead to conflicting intuitions. When faced with conflicting intuitions people consciously reason towards a judgement (Bucciarelli & Daniele, 2015).

The final principle, moral inconsistency, states that "the beliefs that are the basis of moral intuitions and conscious moral reasoning are neither complete nor consistent" (Bucciarelli et al., 2008, p. 128). Evidence for this principle has been discussed above in the rejection of rationalism. Recall the influence of contextual factors on judgements with the same outcomes (trolley dilemmas). Furthermore, Bucciarelli et al. (2008) have shown that people are able to identify modifications to dilemmas that would (a) lead them to change their judgements, and (b) render the dilemmas unresolvable.

The key strength of model theory over dual-process theories of moral judgement more generally is that it goes beyond a descriptive account of moral judgements being grounded in two processes. Where other theories investigate contextual factors that influence intuitions or the selection of one intuition over a conflicting intuition. Model theory offers a detailed account of the process of conscious reasoning not found in other theories. According to model theory, people reason about deontic propositions using mental models. A deontic proposition relates to how permissible/impermissible/obligatory an action may be. According to model theory, people represent deontic propositions as mental models of the associated permissible (or in some cases impermissible) state. Consider the following problem (taken from Bucciarelli et al., 2008, p. 126):

> You are permitted to carry out only one of the following two actions:
>
> Action 1: Take the apple or the orange, or both.
>
> Action 2: Take the pear or the orange, or both.
>
> Are you permitted to take the orange?

For Action 1, there are three permissible states that may be represented in a model: (i) taking the apple; (ii) taking the orange; and (iii) taking both the apple and the orange. Similarly, for Action 2 there are three permissible states: (i) taking the pear; (ii) taking the orange; and (iii) taking both the pear and the orange. Based on these models the intuitive response to the question "Are you permitted to take the orange?" is "Yes". However, this intuitive answer is in fact wrong because only one of the actions is permitted, and the taking of the orange is described in both Action 1 and Action 2. This error is predicted by model theory, and studies have shown that people make this error (Bucciarelli & Johnson-Laird, 2005; Bucciarelli et al., 2008). Framing deontic propositions in terms of mental models provides an account of conscious reasoning that predicts and explains specific variability in judgements.

Another strength of model theory, particularly for the current discussion, is that moral dumbfounding is not only explicitly addressed by it, but is also predicted by it. According to model theory the making of moral judgement occurs when we reason from unconscious premises. We may draw conscious conclusions from these unconscious premises however the premises remain unconscious. It is suggested by Bucciarelli et al. (2008) that these unconscious premises largely present as deontic propositions. When people make a moral judgement from these unconscious premises they simply apply the content of the premise (a deontic proposition) to the situation. For example, when reasoning about a scenario involving murder, a person's intuition is informed by the deontic proposition that "murder is wrong". That a person's intuition is informed by a deontic proposition implies that the content of the intuition is limited to the content of the deontic proposition. This means that if a person is asked for reasons for their intuition they will not necessarily be able to provide a reason, or as suggested by Bucciarelli et al. (2008, p. 127), they will be "dumbfounded".

Bucciarelli et al. (2008) predict the existence of dumbfounding beyond the domain of morality. They draw on Haidt (2001) and his discussion of moral dumbfounding in support of their prediction, however they also provide two plausible examples of how dumbfounding may occur in a non-moral context. Firstly, they refer to instances of unconscious knowing without being able to articulate why, with specific reference to music style and musicians. For example, a person might hear a piece of music and immediately identify it as being by Debussy. This person may be right, even without ever having heard the piece before yet they may be unable to articulate the specific features of the music that led them to conclude that it was by Debussy. Secondly, it is possible that people may struggle to provide a reason to the question "why shouldn't you eat peas with a knife?" (Bucciarelli et al., 2008).

Model theory discusses moral dumbfounding and predicts dumbfounding in the non-moral domain, however, the explanation of dumbfounding does not extend beyond attributing deontic propositions to unconscious premises. The discussion of these unconscious premises or intuitions is the key weakness of model theory. There is no attempt to provide an explanation for the emergence of these unconscious premises and therefore does not provide a full explanation of dumbfounding. This limitation is present in other dual-process approaches (Evans, 2010; Mallon & Nichols, 2011; Park, Levine, Kingsley Westerman, Orfgen, & Foregger, 2007). However, there are two sets of approaches attempt to provide an account for the emergence of the intuitions (or unconscious premises) that give rise to moral judgements. The first of these are expertise/skill based approaches. The second are categorisation approaches. These are discussed briefly below.

1.5.4 Moral intuitions and moral expertise. The intuitionist and dual-process theories discussed above do not offer a coherent picture of where the intuitions originate. Haidt (2001) suggests that they are innate. The aligning of the dual-process theories of moral judgement with dual-process theories of cognition more generally suggests that implicit learning plays a role in the development of moral intuitions (e.g., Berry & Dienes, 1993; Evans, 2010; Reber, 1989; Sun et al., 2005) though the specific mechanisms are unclear. Cushman (2013) and Crockett (2013) argue for independent learning systems but it is unclear how these would work.

A number of theorists have proposed skill based accounts for the development of morality. Dreyfus and Dreyfus (1990) provide a six stage account for the development of ethical expertise with analogy to learning to play chess and learning to drive. Narvaez and Lapsley (2005), and Hulsey and Hampson (2014) incorporate both moral decision making and moral behaviour into their accounts of the development of moral expertise. While these accounts have certain merit, they are limited in their usefulness for specifically understanding moral dumbfounding. Dreyfus and Dreyfus' (1990) account is very formal, providing an account for the learning of moral codes in a deliberate, and deliberative manner. However, the implicit learning of moral codes is not discussed. Narvaez and Lapsley (2005) emphasises the role of practice in the development of moral expertise; however she does not adequately distinguish between expertise in simply recognising behaviours as morally right or morally wrong, and what she calls moral expertise – whereby a person behaves morally all the time. Similarly, Hulsey and Hampson (2014) describe the development of moral knowledge as the development of expertise, which is linked to moral behaviour and moral identity.

1.5.5 Categorisation approaches to moral judgement. There have been at least two attempts to link the moral judgement literature with the research on categorisation. The first theory of interest is that of Stich (1993) which was elaborated on by Harman, Mason and Sinnott-Armstrong (2010). This approach rejects what Harman et al. term the "classical view of concepts" (Harman et al., 2010, p. 227). According to the classical view of concepts, every concept can be defined by a set of necessary and sufficient conditions (Harman et al., 2010, p. 227). Applying the classical view to the moral domain would lead to something of a rationalist approach, whereby the concepts of right and wrong are viewed in terms of a set of principles. In rejecting the classical view, Stich does not commit to which alternative approach should be adopted, rather that developments in both the categorisation literature and the morality literature should be considered in parallel. Harman et al. (2010) appear to adopt an exemplar approach, whereby a concept "is a set of stored (representations) of instances" (2010, p. 231).

While promising, the approach proposed by Stich (1993) pre-dates much of the influential contribution to the categorisation literature made by Barsalou (Barsalou, 1987, 1991, 1999, 2003, 2005, 2009). Harman et al. (2010) do not acknowledge Barsalou's contribution. Barsalou rejects exemplar theories in favour of situated simulation theory on the grounds that exemplar models purport to be modular, stable, and implicitly taxonomic in organisation, while the empirical research suggests that categorisation is non-modular, dynamical, and with an organisation that emerges as a consequence of goal directed behaviour (Barsalou, 2003). That neither Stich (1993), or Harmon et al. (2010) incorporate the important insights and contribution of Barsalou (e.g., Barsalou, 2003) into their theories means that these categorisation approaches to moral judgement do not reflect recent developments in the categorisation literature, limiting their value for the current project.

Prinz (2005) appears to present a more promising view of a categorisation approach to the study of moral judgement. Prinz (2005) describes the development of concepts and categories with specific reference to the influential work of Barsalou and attempts to extend this to the moral domain. However, Prinz proceeds to attribute moral judgements almost entirely to emotion: "Emotions, I will suggest, are perceptions of our bodily states. To recognize the moral value of an event is, thus, to perceive the perturbation that it causes" (Prinz, 2005, p. 99). This strong view of "moral categorisation as emotion" appears to be over-stating the role of emotion in the making of moral judgement, particularly in view of recent work advocating a more measured view of the role of emotion in the making of moral judgement (Huebner et al., 2009; Landy & Goodwin, 2015; May, 2014).

1.6 Conclusion

From the above discussion it is clear that moral dumbfounding provides an illustration of the intuitive nature of moral judgements. The discovery of moral dumbfounding contributed to the growth of intuitionist theories of moral judgement and it is increasingly accepted that intuition has some role in the making of moral judgements. Haidt's suggestion (Haidt, 2001; Haidt & Björklund, 2008) that moral intuitions are innate is not widely supported. It is also becoming increasingly apparent that reasoning plays a greater role in the making of moral judgement than Haidt allowed for (Haidt, 2001; Haidt & Björklund, 2008), prompting a range of theorist to adopt a dual-process view (Brand, 2016; Crockett, 2013; Cushman, 2013; Greene, 2008, 2013; Greene et al., 2001; Gubbins & Byrne, 2014; Mallon & Nichols, 2011), and various authors to study specifically the role of reasoning in the making of moral judgements (Bialek & Terbeck, 2016; Bucciarelli, 2009; Bucciarelli & Johnson-Laird, 2005; Bucciarelli et al., 2008; Byrne, 2015; Cowley & Byrne, 2005; Gubbins & Byrne, 2014; Johnson-Laird, 2006). The existence of moral dumbfounding remains consistent with these approaches, and it continues to be cited as evidence for them (Bucciarelli et al., 2008; Cushman, 2013; Triskiel, 2016).

Despite the prevailing influence of moral dumbfounding on the moral psychology literature, it remains poorly explained by existing theories. Furthermore, there remains uncertainty regarding whether or not dumbfounding is a real phenomenon, and the primary evidence in support of it is limited to a single unpublished manuscript which has not been directly replicated. The paucity of both, (a) empirical evidence for dumbfounding, and (b) theoretical explanations of dumbfounding, undermine the usefulness of drawing on dumbfounding in discussions of theories of moral judgement. The remainder of this thesis will attempt assess if drawing on moral dumbfounding can contribute meaningfully to the moral judgement literature by addressing each of these limitations in turn.

2 Chapter 2 – Intuitions and Moral Dumbfounding

Moral dumbfounding is consistent with both intuitionist perspectives and with dual-process theories of moral judgement and is cited in support of these approaches (Brand, 2016; Cushman, 2013; Doris, 2010; Greene, 2008; Haidt, 2001; Triskiel, 2016). Beyond this, it is also cited as evidence in support of both reasoning (Bucciarelli et al., 2008) and categorisation (Prinz, 2005) theories of moral judgement. However, moral dumbfounding remains poorly explained by these theories of moral judgement. Perhaps the most detailed discussion of possible causes of dumbfounding come from sceptics of the phenomenon (e.g., Jacobson, 2012; Royzman, Atanasov, et al., 2014; Sneddon, 2007). The absence of explanations of dumbfounding, and associated controversy surrounding its existence is related to the lack of empirical investigation into moral dumbfounding specifically.

The first section of this chapter will outline the explanations of dumbfounding associated with the primary theories of interest discussed in Chapter 1. The second section will detail the challenges to dumbfounding made by various authors. The final section will evaluate the strength of evidence the original demonstration of moral dumbfounding (Haidt et al., 2000) provides for the existence of moral dumbfounding.

2.1 The Limited Explanations of Moral Dumbfounding

Moral dumbfounding is widely-cited in the psychology literature as supporting evidence for various theories of moral judgement (e.g., Bucciarelli et al., 2008; Cushman et al., 2010; Dwyer, 2009; Haidt, 2001; Hauser et al., 2008; Prinz, 2005). However, any explanations of moral dumbfounding offered by these accounts are primarily descriptive. They do not offer insight into the specific mechanisms that underlie, or conditions that may lead to dumbfounding.

Dwyer (2009) draws on moral dumbfounding in defending the linguistic analogy or universal moral grammar. She outlines parallels between moral knowledge and linguistic knowledge, noting that, much like the moral dumbfounding paradigm, people can successfully apply a grammatical principle without being able to articulate it. However, the various criticisms of the linguistic analogy were outlined in Chapter 1, particularly the uncritical accepting of Chomsky's framework (Chomsky, 1965, 1976, 2000; Christiansen & Chater, 2008; Hinzen, 2012; Tomasello, 2003, 2014) mean that this framework is of limited value in understanding moral dumbfounding.

Consider now the unpublished report describing the original demonstration of moral dumbfounding (Haidt et al., 2000). In the foreword to this manuscript, Haidt acknowledges the limited theoretical contribution of the report. He describes it as a "description of an interesting phenomenon" (Haidt et al., 2000, p. 1), and cites the descriptive nature of the report as a reason for not submitting it for peer-reviewed publication. In discussing SIM (Haidt, 2001; Haidt & Björklund, 2008), Haidt describes the process of moral judgement as being "caused by quick moral intuitions and is followed (when needed) by slow, ex post facto moral reasoning" (Haidt, 2001, p. 817). Haidt presents moral dumbfounding to illustrate this view of moral judgement, however, beyond being used as an illustration, there is no deeper explanation of moral dumbfounding provided.

Greene (2008) does not offer an explanation of moral dumbfounding. Crockett (2013) does not refer to dumbfounding in outlining her approach to moral judgement. Stich (1993) and Harman et al. (2010) do not discuss moral dumbfounding. Dreyfus and Dreyfus's (1990) skill acquisition account of moral judgement pre-dates the discovery of moral dumbfounding. Neither Narvaez and

Lapsley (2005), nor Hulsey and Hampson (2014), discuss moral dumbfounding directly, and therefore neither of these approaches offer an explanation of moral dumbfounding.

Cushman, on the other hand, discusses dumbfounding directly in outlining his dual-process approach (Cushman, 2013), and elsewhere (e.g., Cushman et al., 2010, 2006). According to Cushman (2013), dumbfounding emerges as a consequence of the action/outcome distinction inherent in his model-free/model-based approach. Cushman suggests that people pass judgement using model-free mechanisms, focusing on actions, and that the search for reasons for the judgement uses model-based processes, and concerns outcomes. In this way, the harmless nature of scenarios that lead to dumbfounding prevents a person from successfully identifying outcome based reasons for their judgement. Despite the limitations discussed in Chapter 1, this explanation may provide a useful starting point testing a more general dual-process explanation of moral dumbfounding. Recall the habitual/deliberative distinction discussed in Chapter 1. Applying this to Cushman's explanation of dumbfounding (Cushman, 2013), identifies the making of a judgement as a habitual response, and the identification of reasons as deliberation. This means that one possible explanation of moral dumbfounding is that it emerges when deliberation yields a different response to a habitual response, resulting in conflict. This type of conflict has been well researched in dual-process research (Bonner & Newell, 2010; De Neys, 2012, 2014; De Neys & Glumicic, 2008). Identifying dumbfounding as conflict in dual-processes provides an explanation of moral dumbfounding that can be tested (e.g., manipulations of cognitive load have been shown to increase the rate of habitual type responding, see De Neys, 2006).

Prinz (2005) describes moral judgements as grounded in emotion. According to this view, when people make a moral judgement, they are interpreting an emotional reaction to a particular situation. Prinz (2005) identifies this emotional nature of moral judgements as the reason why they may be difficult to justify. Prinz suggests that in moral dumbfounding, a failure to provide reasons for a judgement may not be limited to harmless taboos, and proposes that questioning for reasons behind judgements of other behaviours, e.g., murder, may also lead to moral dumbfounding. As noted in Chapter 1 however, Prinz's approach places too much emphasis on the role of emotion in the making of moral judgement. This casts doubt on the claim that the emotional nature of moral judgements is what leads to moral dumbfounding.

Bucciarelli et al. (2008) present a theory of moral judgement that predicts dumbfounding. According to their model theory, moral knowledge concerns deontic propositions, such that our intuitions about the permissibility or impermissibility of a given action is grounded in propositional beliefs, e.g., "stealing is wrong". Knowing the content of a given deontic proposition is sufficient for successful moral reasoning. It is not necessary to know reasons for a given deontic proposition in order to successfully apply the proposition and reason successfully about the permissibility or impermissibility of a given behaviour. If a person knows that stealing is wrong, they will be able successfully identify stealing behaviour as wrong; and knowing reasons for judging stealing as wrong will have no bearing on the success with which they identify the behaviour as wrong. In other words, the making of a moral judgement does not require knowledge of the reasons for making that judgement, all that is necessary to make a judgement is knowledge of the relevant deontological principle. According to Bucciarelli et al. (2008), questioning a person on any given moral proposition may potentially lead to dumbfounding. This model theory explanation of moral dumbfounding can be tested.

Of the theories of moral judgement discussed in Chapter 1, there are two approaches, dual-process theories and model theory, that both stand up to scrutiny, and offer testable explanations of dumbfounding. According to a dual-process explanation, manipulations of cognitive load or psychological distance should influence responses in predictable ways (e.g., De Neys, 2006; Kross & Ayduk, 2008; Trope & Liberman, 2003). According to model theory, changes in the information provided may also influence responses in predictable ways (e.g., Bucciarelli et al., 2008; Johnson-Laird, 2006). These explanations of dumbfounding have not been tested. Testing them will further the development of theories of moral judgement, providing a greater insight into both the general making of moral judgements and the phenomenon of moral dumbfounding. However, before these explanations can be tested, the challenges to dumbfounding and related explanation should be addressed.

2.2 Challenges to Dumbfounding

The most detailed discussions of possible explanations of dumbfounding come from authors who challenge the accepted implication of dumbfounding, that moral judgements may not necessarily be grounded in reasons. Wielenberg (2014) argues that internalised moral principles guiding moral judgements do exist but may be "hidden" (Wielenberg, 2014, p. 99) in the same way that grammatical rules language are often easier to adhere to than to articulate.

Sneddon (2007) does not argue for hidden internalised moral principles, rather, he argues that these moral principles are located in the external social world. Sneddon contrasts expert knowledge with that of a layperson on a given subject. A layperson may not know how something like a computer works, however they know that there are basic principles that make it work, and that a computer engineer could probably articulate these principles. He suggests that dumbfounding is a similar phenomenon and that, when dumbfounded, people do not accept the possibility that their judgement is not based on reasons, rather that they just cannot articulate these reasons, in the same way that they cannot articulate the reasons that a computer works the way it does (Sneddon, 2007).

Some of the possible reasons that may underlie the judgements of dumbfounded participants have been suggested by various authors. For example, Gray, Schein, and Ward (2014) suggest that when judging moral scenarios people implicitly perceive harm even in scenarios that are construed as objectively harmless. If people perceive harm in the scenarios, then, even when the experimenter claims that they are harm free, this perception of harm still serves as a reason to condemn the behaviour. They conducted a series of experiments demonstrating that people do implicitly perceive harm in supposedly victim-less scenarios, e.g., “covering a bible with faeces”, or “having sex with a corpse” (Gray et al., 2014, p. 1603).

Similarly Jacobson (2012) argues that, in general, moral judgements are grounded in reasons and presents a number of plausible reasons why a person may condemn the actions of the characters in each of the dumbfounding scenarios. He also suggests that when participants appear to be dumbfounded they have simply given up on the argument and conceded to the experimenter who is in a position of authority.

Building on the work of Jacobson (2012), a recent series of studies by Royzman, Kim, and Leeman (2015), focusing on the *Incest* dilemma, identified two reasons that may be guiding participants' judgements. The reasons identified were: (a) potential harm – where participants believed that harm could arise as a result of the actions of the characters in the scenario despite the vignette stating that no harm arose; and (b) normative – where citing a moral norm is seen as sufficient justification for making a judgement consistent with that norm (regardless of the presence or absence of harm specifically). They found that people who rated the behaviour of Julie and Mark as wrong generally endorsed at least one of those reasons and suggest that this is evidence for a rationalist view of moral judgement (Royzman et al., 2015, p. 311). They argue that dumbfounded responding can be attributed to social pressure that exists in an interview setting, whereby participants accept the counter-arguments offered by the interviewer, even if they disagree, in order to avoid appearing uncooperative (Royzman et al., 2015, p. 299).

The explanations and challenges to the dumbfounding paradigm described above and challenges are largely grounded in the claim that people do have reasons for their judgements, and that dumbfounding can be interpreted in way that is consistent with this claim. Only one of these rationalist explanations has been tested (e.g., Royzman et al., 2015). Specific methodological issues with the Royzman et al. (2015) studies are identified and addressed in Chapter 4.

The controversy surrounding moral dumbfounding can at least partly be attributed to the lack of empirical evidence for dumbfounding. Evidence demonstrating moral dumbfounding is limited to the original Haidt et al. (2000) demonstration. It could be argued that the strength of evidence this study provides for the existence of the phenomenon does not justify the influence it has had on the morality literature.

2.3 The Paucity of Evidence for Moral Dumbfounding

In the Haidt et al. (2000) study, participants were presented with a moral reasoning scenario (*Heinz*), two moral intuition scenarios (*Incest* and *Cannibal*) and two behavioural intuition tasks (*Roach* and *Soul*). *Heinz* was taken from Kohlberg (1969) and depicted a scenario in which a man steals a drug to save his dying wife (see Appendix A). *Incest* was the Julie and Mark scenario described in Chapter 1. *Cannibal* described a woman, Jennifer, who worked with human cadavers in a medical school pathology lab, who took some human meat from the lab home and ate it. For *Roach* participants were asked to drink some apple juice/water that had had a sterilised cockroach dipped in it. For *Soul*, participants were offered two dollars to sign a piece of paper that contained the words "I, (participant's name), hereby sell my soul, after my death, to Scott Murphy [the experimenter], for the sum of two dollars". The bottom of the page included a disclaimer stating "this is not a legal or binding contract" (Haidt et al., 2000, p. 7).

Participants were presented with each of the scenarios/tasks and asked for their judgement/decision to complete the task. Following this, the experimenter argued against the judgement/decision of the participant. This discussion was video taped, and following the discussion for each scenario/task, participants completed a questionnaire in which they indicated their confidence in their judgement, how much their judgement was based on reason, and how much their judgement was based on gut feeling. To counterbalance for order effects, the scenarios/tasks were presented in one of two preselected orders. Finally a brief questionnaire that contained questions relating to basic demographics, and participants' political and religious views was completed. The primary analyses compared the responses to the intuition scenarios/tasks to the responses to the moral reasoning scenario.

There are four key areas of concern relating to this original study. Firstly, the report serves as a demonstration that certain moral scenarios elicit a response that appears to be more grounded in emotion than in reason. Moral dumbfounding is not rigorously tested, rather it is presented as an interesting phenomenon that illustrates the emotional nature of moral judgements. A second related concern, is that the scenarios that were compared elicited judgements of differing valences, such that people condemned the behaviour described in the intuition scenarios, however in the reasoning scenario they did not rate the behaviour of Heinz as wrong. To date, demonstrations of moral dumbfounding are limited to cases where participants rated a behaviour as wrong. This means that the valence of judgement may serve as a confounding variable in investigating variation in other responses associated with the different scenarios. A third concern, previously identified in Chapter 1, is that the report does not provide a clear definition of moral dumbfounding from which specific measures of dumbfounding can be gleaned. Instead, an array of variables, with varying operationalisations, that may or may not be indicative of a state of dumbfounding are presented. Finally, the final sample of the original study was thirty, this study was not published in peer reviewed form, and has not been directly replicated. These are discussed in turn below.

The study described in the original and most widely-cited demonstration of moral dumbfounding (Haidt et al., 2000) does not provide a rigorous test for the existence of moral dumbfounding. The analysis presented does not identify incidences of dumbfounding and compare them against instances where dumbfounding is not present. Instead, the responses to the different types of scenarios are compared. There is an implicit assumption in conducting this analysis that the “intuition scenarios” led to dumbfounding and the reasoning scenario did not, however this is not directly tested, again illustrating the absence of an explicit measure of moral dumbfounding.

The analysis of the variables of interest did not control for valence of judgement (whether participants judged the behaviour as right or wrong). Haidt et al. (2000) compared an array of responses regarding *Heinz* to similar responses regarding *Incest* and *Cannibal*. However, the valence of participants' judgements varied depending on scenario type. In general, participants rated *Cannibal* and *Incest* as wrong while generally rating *Heinz* as not wrong. This means that comparing responses depending on scenario type was also comparing a judgement of "wrong" against a judgement of "not wrong". Any differences observed may be due to the relative valence of judgements made as opposed to the type of scenario.

Haidt et al. (2000) define moral dumbfounding as "the stubborn and puzzled maintenance of a judgment without supporting reasons" (Haidt et al., 2000, p. 2). This definition is found in the abstract of the report and does not appear in the main body. Chapter 1 discussed the lack of definitional specificity regarding moral dumbfounding; also highlighting the consequence of this lack of definitional specificity, that there are currently no agreed measurable indicators of moral dumbfounding in the wider psychology literature. Drawing on the above definition, the discussion within the original report, and on the wider morality literature two responses that may be indicative of a state of dumbfounding were identified in Chapter 1: admissions of not having reasons, and unsupported declarations/tautological responses.

Each of these responses is discussed by Haidt et al. (2000) in the original paper, however it is not clear if these responses are taken as measures of dumbfounding. They are discussed in conjunction with a range of responses that may be related to dumbfounding. In addition to admissions of not having reasons, and the use of unsupported declarations, Haidt et al. (2000) also appear to take increased levels of confusion, higher ratings of judgements being grounded in “gut feeling”, an increase in “dead-ends” (whereby participants gave up on an argument), a higher frequency of saying “I don’t know” as indicative of dumbfounding. Each of these responses is presented as being related to dumbfounding, however no response is identified as essential for dumbfounding to be observed. Admissions of not having reasons are described as the “clearest evidence of dumbfounding” (Haidt et al., 2000, p. 14), yet this still falls short of identifying this response as indicative of moral dumbfounding. Rates of dumbfounding are not reported in the original paper, instead, rates of various responses are reported. Furthermore, the rates of these responses are reported relative to type of scenario as opposed to whether or not dumbfounding was observed.

Notwithstanding the challenges surrounding the measuring of dumbfounding discussed above, the report describes a single study that had a final sample of thirty participants. A single study with such a small sample size does not provide strong evidence for the phenomenon of moral dumbfounding. The report has not been published in peer-reviewed form and has not been directly replicated. This (particularly in view of the recent replication crisis in psychology, e.g., Open Science Collaboration, 2015) casts considerable doubt on the strength of evidence it provides for the existence of moral dumbfounding.

2.4 The Current Project

Two general research questions have been identified in response to the areas of concern identified in this chapter. Firstly, given the paucity of evidence for moral dumbfounding, the most pressing research question to address is: is moral dumbfounding a real phenomenon? Directly following from this question is the second key question: how can the existence (or absence) of dumbfounding inform theories of moral judgement? These questions are related, such that whether or not dumbfounding is real will have very different implications for the wider morality literature. Each of these primary questions may be broken up into component questions as follows:

1 Is moral dumbfounding a real phenomenon?

1.1 How should moral dumbfounding be measured?

1.2 Is it possible to elicit moral dumbfounding in a laboratory based task?

1.3 Is dumbfounded responding truly indicative of a state of dumbfoundedness or can it be attributed to features of experimental design (e.g., Royzman et al. 2015)

2 How can the existence (or absence) of dumbfounding inform theories of moral judgement?

2.1 Can the existence (or absence) of moral dumbfounding be adequately explained by the existing approaches to moral judgement?

2.1.1 Rationalism

2.1.2 Dual-Process approaches

2.1.3 Model theory

2.2 Can the existence (or absence) of moral dumbfounding be better explained by adopting an alternative theoretical position?

This thesis will attempt to provide answers to each of the questions above. Across Chapters 3 and 4, I attempt to establish whether or not moral dumbfounding is a real phenomenon (Question 1). In Chapter 3, I test the original paradigm and develop methods for measuring (1.1) and systematically eliciting dumbfounding (1.2). In Chapter 4, I assess the claim that dumbfounded responding occurs as a result of social pressure not to appear uncooperative and that participants' reasons are unfairly dismissed by the researcher (1.3). Question 2 is addressed across Chapters 4 through 8. Firstly a rationalist explanation (2.1.1) for moral dumbfounding is tested in Chapter 4. Chapter 5 assesses one prediction of a dual-process explanation of moral dumbfounding (2.1.2). Chapter 6 tests a second prediction of a dual-process explanation (2.1.2) of moral dumbfounding, and tests a prediction of a model theory explanation of moral dumbfounding (2.1.3). Chapter 7 provides a general discussion of the results of the studies conducted. Chapter 8 explores an alternative theoretical position that may provide a stronger explanation of moral dumbfounding (2.2).

2.5 Conclusion

This chapter has identified three primary areas of concern surrounding moral dumbfounding. Firstly, despite the influence dumbfounding has had on the morality literature, there are very few theories of moral judgement that both accept the existence of dumbfounding and offer testable explanations of dumbfounding. A second related concern is that the most detailed explanations of moral dumbfounding come from critics of the paradigm, who argue that dumbfounding is not a real phenomenon. Much of the uncertainty surrounding moral dumbfounding can be attributed to the third concern discussed above, the limited empirical evidence for dumbfounding.

From these areas of concern relating to moral dumbfounding, two broad research questions have been identified. A number related specific questions have been identified from these two broad questions. The remainder of this thesis attempts to address each of these in turn. Chapter 3 tests the original paradigm, and develops methods for measuring and testing moral dumbfounding. Chapter 4 uses the methods developed in Chapter 3 to address the rationalist challenge to dumbfounding from Royzman et al. (2015). Finally, Chapters 5 and 6 investigate dual-process and mental models explanations of dumbfounding respectively.

3 Chapter 3 – Searching for Moral Dumbfounding: Identifying Measurable Indicators of Moral Dumbfounding

Moral dumbfounding has been identified as playing an influential role in the development of theories of moral judgement. It provides strong evidence for theories that are grounded in some form of intuitionism (e.g., Cushman et al., 2010; Haidt, 2001; Prinz, 2005). The existence of moral dumbfounding is consistent with the fundamental claim of such theories, that our moral judgements are grounded in an emotional or intuitive automatic response rather than slow deliberate reasoning (Cameron et al., 2013; Crockett, 2013; Cushman, 2013; Cushman et al., 2010; Greene, 2008; Haidt, 2001; Prinz, 2005). In contrast, moral dumbfounding is incompatible with a rationalist approach that identifies reason or principles as the causes of moral judgements (e.g., Kohlberg, 1971; Narvaez, 2005; Royzman et al., 2015; Topolski, Weaver, Martin, & McCoy, 2013).

Despite the theoretical weight moral dumbfounding carries, there are limited theoretical explanations of moral dumbfounding, and the existence of dumbfounding has been challenged in recent years. This uncertainty and related controversy is unsurprising given the paucity of empirical evidence demonstrating moral dumbfounding, and the related uncertainty regarding how moral dumbfounding should be measured.

3.1 Searching for Moral Dumbfounding

In response to the limited number of demonstrations of, and related uncertainty surrounding moral dumbfounding, the primary aims of the current chapter are (a) to identify specific measurable indicators of moral dumbfounding; and (b) use these measures to examine the reliability with which dumbfounded responding can be evoked. These aims address two elements of primary Research

Question 1 identified in Chapter 2: Is moral dumbfounding a real phenomenon. Specifically, the aims of this chapter assess 1.1 "How should moral dumbfounding be measured?" and 1.2 "Is it possible to elicit moral dumbfounding in a laboratory based task?".

We conducted four studies, each of which is a modified replication attempt of the original moral dumbfounding study (Haidt et al., 2000). In these studies, dumbfounding is measured according to two sets of responses: (a) an admission of having no reasons for a judgement (a measure of self-reported dumbfounding) and, (b) use of unsupported declarations ("it's just wrong") or tautological reasons ("because it's incest") as a justification for a judgement (measures of a failure to provide reasons). Study 1 was designed to replicate Haidt et al.'s (2000) initial study using the original methods (face to face interview). In Study 2 we piloted alternative methods (a computer-based task) in an attempt to evoke moral dumbfounding in a systematic way with a larger sample. In Study 3a and 3b the materials that were piloted in Study 2 were refined and administered to a larger sample in an attempt to systematically evoke dumbfounded responding.

3.2 Study 1: Interview

The primary aim of Study 1 was to replicate the original dumbfounding study (Haidt et al., 2000). However, in response to some of the limitations of the original study identified in Chapter 2, a number of changes were made to the materials and methods used. Firstly, four moral judgement vignettes were used (Appendix A) instead of three. Three of these vignettes (*Heinz*, *Incest*, and *Cannibal*) were taken from Haidt et al. (2000). A fourth vignette (*Trolley*) was adapted from Greene et al. (2001). Haidt et al. (2000) contrasted *Heinz*, a so-called reasoning scenario, against *Cannibal* and *Incest*, so-called intuition scenarios. Modifications were made to the procedure in order to ensure that, in so far as possible, all participants were defending a judgement of “morally wrong”. The original study also included two tasks that did not have any moral content. For the purposes of consistency and balance, the non-moral tasks were omitted from the present study, and a second moral reasoning vignette was included in their stead, such that two reasoning vignettes (*Heinz* and *Trolley*) were contrasted against two intuition vignettes (*Incest* and *Cannibal*). It was hypothesised that dumbfounding would be elicited. It was also hypothesised that rates of dumbfounded responding would vary depending on the content of the dilemma, with the intuition scenarios eliciting more dumbfounded responses than the reasoning scenarios. Two measures of dumbfounding were taken reflecting the two distinct ways in which absence of reasons may present: admissions of not having reasons (self-reported dumbfounding), and the use of an unsupported declaration (it’s just wrong) as a justification for a judgement, with a failure to provide any alternative reason when the unsupported declaration was questioned (a failure to provide reasons). As in the original study (Haidt et al., 2000), various non-verbal measures were also recorded in order to address questions relating to participants' stubbornness and puzzlement. Given the changes to the original procedure, there was a brief piloting phase during which the materials and methods described below were finalised.

3.2.1 Method.

3.2.1.1 Participants and design. Study 1 was a frequency based attempted replication. The aim was to identify if dumbfounded responding could be evoked. All participants were presented with the same four moral vignettes. Results are primarily descriptive. Any further analysis tested for differences in rates of responding depending on the vignette, or type of vignette, presented.

A sample of thirty-one participants (16 female, 15 male) with a mean age of M_{age} = 28.83 (min = 19, max = 64, *SD* = 11.14) took part in this study. Participants were undergraduate students, postgraduate students, and alumni from Mary Immaculate College (MIC), and University of Limerick (UL). Participation was voluntary and participants were not reimbursed for their participation.

3.2.1.2 Procedure and materials. Four moral judgement vignettes were used (Appendix A). Three of the vignettes (*Heinz*, *Incest*, and *Cannibal*) were taken from Haidt et al. (2000). *Incest* was taken directly from the original study however *Cannibal* and *Heinz* were modified slightly, following piloting.

The original version of *Cannibal* stated that people had "donated their body to science for research"; participants during piloting were able to argue that eating does not constitute "research". In order to remove this as a possible argument, the modified version stated that bodies had been donated for "the general use of the researchers in the lab" and that the "bodies are normally cremated, however, severed cuts may be disposed of at the discretion of lab researchers".

Similarly, piloting suggested that participants agreed with the actions of Heinz and condemned the actions of the druggist. The original wording of *Heinz* suggested that any discussion related to Heinz as opposed to the druggist meaning that, for *Heinz*, participants would typically be defending an approval of the character's actions. However, for *Incest* and *Cannibal* participants generally condemn the actions of the principal character to the degree that in the majority of cases, participants are defending a judgement of "morally wrong". In order to ensure that participants were consistently defending a judgement of "morally wrong" across all scenarios, *Heinz* was modified to include "The druggist had Heinz arrested and charged". Any discussion on *Heinz* then related to the character whose behaviour participants thought was wrong.

In the original study by Haidt et al., (2000), *Incest* and *Cannibal* are presented as “intuition” stories, and contrasted against a single “reasoning” dilemma: *Heinz*. In order for a more balanced comparison, a bridge variant of the classic trolley dilemma (*Trolley*) was included as a second “reasoning” dilemma. In this vignette, participants judge the actions of Paul, who pushes a large man off a bridge to stop a trolley and save five lives. The inclusion of *Trolley* meant that there were two “reasoning” dilemmas to be contrasted with the two “intuition” stories.

Sample counter arguments were prepared for each scenario. To ensure that participants were only pushed to defend a judgement of “morally wrong” these counter arguments exclusively defended the potentially questionable behaviour of the characters. A list of prepared counter arguments can be seen in Appendix B. A post-discussion questionnaire, taken from Haidt et al. (2000) was administered after discussion of each scenario (Appendix C).

Two other measures were also taken for exploratory purposes. Firstly, in response to a possible link between meaning and morality (e.g., Bellin, 2012; Schnell, 2011), the Meaning in Life questionnaire (MLQ; Steger, Kashdan, Sullivan, & Lorentz, 2008) was included. This ten item scale, is made up of two five item sub scales: presence (e.g., “I understand my life’s meaning.”) and search (e.g., “I am looking for something that makes my life feel meaningful.”). Responses were recorded using a 7-point Likert scale ranging from 1 (*strongly disagree)* to 7 *(strongly agree)*. Secondly, in line with Haidt’s (2007; see also, Haidt & Hersh, 2001) work describing a link between religious conservatism and moral views, it was hypothesised that incidences of dumbfounding may be moderated by individual differences in religiosity. To assess this possibility, the seven item CRSi7 scale, taken from The Centrality of Religiosity Scale (Huber & Huber, 2012) was also included. Participants responded to questions relating to the frequency with which they engage in religious or spiritual activity (e.g., “How often do you think about religious issues?”). Responses were recorded using a 5-point Likert scale ranging from 1*(never)* to 5 *(very often)*.

The interviews took place in a designated psychology lab in MIC and were recorded on a digital video recording device. Participants were presented with an information sheet and a consent form. The consent form required two signatures: firstly, all participants consented to take part in the study (including consent to be video recorded); the second signature related to use of the video for any presentation of the research (with voice distorted and face pixelated). Only two participants opted not to sign the second part.

Participants read brief vignettes describing each scenario, and were subsequently interviewed regarding the protagonists. All four scenarios were discussed in a single interview session, with a brief pause between each discussion for the participant to complete a questionnaire about their judgements, and to read the next scenario. The conversation continued when they were happy to do so. Each of the four moral dilemmas *Heinz*, *Trolley*, *Cannibal* and *Incest* (Appendix A) were presented in this way and participants asked to judge the behaviour of the characters in the dilemmas. The order of presenting the scenarios was randomised. Judgements made by participants were challenged by the experimenter (“Nobody was harmed, how can there be anything wrong?”; “Do you still think it was wrong? Why?”; “Why do you think it is wrong?”; “Have you got a reason for your judgement?”). The resulting discussion continued until participants could not articulate any further arguments. Participants filled in a brief questionnaire after discussing each dilemma. In this they were asked to rate, on a 7-point Likert scale, how right/wrong they thought the behaviour was; how confident they were in their judgement, how confused they were; how irritated they were; how much their judgement had changed; how much their judgement was based on reason; and how much their judgement was based on "gut" feeling. Participants completed a longer questionnaire at the end of the interview. This contained the MLQ (Steger et al., 2008), the Centrality of Religiosity Scale (Huber & Huber, 2012), and some questions relating to demographics. The entire study lasted approximately 20 to 25 minutes. The videos were analysed using BORIS – Behavioural Observation Research Interactive Software (Friard & Gamba, 2015). All statistical analysis was conducted primarily using R (3.4.0, R Core Team, 2017b)[4]; SPSS (IBM Corp, 2015) was also used.

3.2.2 Results and discussion. The videos of the interviews were analysed and coded by the primary researcher.[5] Participants were identified as dumbfounded if they (a) admitted to not having reasons for their judgements; or (b) resorted to

[4] R (3.4.0, R Core Team, 2017b) and the R-packages *afex* (0.15.2, Singmann, Bolker, & Westfall, 2015), *car* (2.1.4, Fox & Weisberg, 2011), *citr* (0.2.0, Aust, 2016), *desnum* (0.1.1, McHugh, 2017), *devtools* (1.13.1, Wickham & Chang, 2017), *dplyr* (Wickham, Francois, Henry, & Müller, 2017), *estimability* (1.2, Lenth, 2016a), *extrafont* (0.17, Winston Chang, 2014), *foreign* (0.8.68, R Core Team, 2017a), *ggplot2* (2.2.1, Wickham, 2009), *lme4* (1.1.13, Bates, Mächler, Bolker, & Walker, 2015), *lsmeans* (2.26.3, Lenth, 2016b), *lsr* (Lenth, 2016b), *Matrix* (1.2.10, Bates & Maechler, 2017), metap (Dewey, 2017), *papaja* (0.1.0.9492, Aust & Barth, 2017), *plyr* (1.8.4, Wickham, 2011), *pwr* (Champely, 2018), *reshape2* (1.4.2, Wickham, 2007), *scales* (0.4.1, Wickham, 2016), *sjstats* (Lüdecke, 2018), and *wordcountaddin* (0.2.0, Marwick, n.d.).

[5] The coding was conducted according to a codebook which contained behaviours/responses that may be objectively verifiable. The development of this codebook was done in consultation with various other parties, some of whom were blind to the hypotheses. Advice was sought for utterances/behaviours that appeared ambiguous.

using unsupported declarations (“It’s just wrong!”) as justification for their judgements, and subsequently failed to provide reasons when questioned further. Table 3.1 shows the initial and revised ratings of the behaviours for each scenario.

Table 3.1: Study 1: Initial and revised ratings for each scenario

	Heinz		Trolley		Cannibal		Incest	
Judgement	Count	%	Count	%	Count	%	Count	%
Initial judgement wrong	27	87.1%	23	74.2%	25	80.6%	26	83.9%
Initial judgement neutral	0	-	0	-	0	-	0	-
Initial judgement ok	4	12.9%	8	25.8%	6	19.4%	5	16.1%
Revised judgement wrong	5	16.1%	22	71%	23	74.2%	20	64.5%
Revised judgement neutral	16	51.6%	1	3.2%	0	-	0	-
Revised judgement ok	10	32.3%	8	25.8%	8	25.8%	11	35.5%

Twenty-two of the 31 participants (71%) produced a dumbfounded response (admission of having no reasons; or the use of an unsupported declaration as a justification for a judgement, with a failure to provide any alternative reason when the unsupported declaration was questioned) at least once. Examples of such responses included “It just seems wrong and I cannot explain why, I don’t know”, “because I just think it’s wrong, oh God, I don’t know why, it’s just [pause] wrong”. Table 3.2 shows the number, and percentage, of participants who displayed dumbfounded responses and non-dumbfounded responses for each dilemma. The rates of each type of dumbfounded response are also displayed. Table 3.3 shows the responses to the questionnaires presented between dilemmas.

Table 3.2: Study 1: Observed frequency and percentage of each of the responses: dumbfounded, nothing wrong, and reasons provided.

	Heinz		Cannibal		Incest		Trolley	
Response	N	%	N	%	N	%	N	%
Nothing wrong	6	19.35%	8	25.81%	11	35.48%	8	25.8%
Dumbfounded:	0	-	11	35.48%	18	58.06%	3	9.67%
(Admissions)	0	-	8	25.81%	10	32.26%	3	9.67%
(Unsupported declarations)	0	-	3	9.67%	8	25.81%	0	-
Reasons	25	80.65%	12	38.71%	2	6.45%	20	64.52%

Table 3.3: Study 1: Responses to post-discussion questionnaire questions (7-point Likert scale, 1 = Not at all, 7 = Extremely).

	Heinz	Cannibal	Incest	Trolley
How much did you change your mind?	2.87	3.40	2.63	2.60
How confident were you?	5.30	4.77	5.40	5.07
How confused were you?	3.00	3.67	3.33	3.70
How irritated were you?	3.00	3.33	3.13	3.37
How much was your judgement based on "gut" feeling?	5.23	5.20	4.97	5.07
How much was your judgement based on reason?	4.83	4.40	4.43	4.77
Gut minus reason	.40	.80	.53	.30

In line with the original study (Haidt et al., 2000), the videos were also coded, by the primary researcher, across a range of measures (for a full behavioural profile for each participant for each scenario, see the content of the "Graphs – Interview behaviours" folder on this chapter's project page on the Open Science Framework at https://osf.io/wm6vc/ (McHugh, McGann, Igou, & Kinsella, 2017)). Haidt et al. (2000) report differences, between intuition and reasoning scenarios. They do not, however, report comparisons between participants identified as dumbfounded and participants not identified as dumbfounded. Two verbal responses were taken as indicators of dumbfounding (admissions of not having reasons, and unsupported declarations only). Given that each participant responded to four scenarios, the following analyses aggregated the responses for each individual to give 124 cases (each of 31 participants participants' responses to each of the four scenarios).

There were two stages in the following analyses. Firstly all cases of participants presenting as dumbfounded (by either measure, *N* = 32) were compared against cases of participants providing reasons (*N* = 59). Secondly, cases of participants identified as dumbfounded were grouped according to type of dumbfounded response (admissions of not having reasons, *N* = 21; unsupported declarations only, *N* = 11), and cases participants not rating the behaviour as wrong (*N* = 33) were also included in the analysis, along with cases of participants providing reasons (*N* = 59).

Judgement variables reported by Haidt et al. (2000) included the length of time until the first argument, the length of time until the first evaluation, the length of time between the first evaluation and the first argument. We measured and report the same judgement variables. A range of "argument variables" were also reported. Identifying specific objectively verifiable measurable indicators for some of the "argument variables" reported by Haidt et al. (2000) was problematic (e.g., "dead-ends", "argument kept", "argument dropped"). We coded each verbal utterance according to a relevance for forming an argument. As a result, some of the argument variables reported by Haidt et al. (2000) are not reported here in the same way, however, related measures are reported. Every sentence of speech was coded, during coding an additional argument variable was identified: "working towards a reason". This code was applied when participants made a statement that was relevant to the content of the scenario, but could not be coded as providing a judgement, or providing a reason.

Paralinguistic variables reported by Haidt et al. (2000) include frequency (per minute) of: “ums, uhs, hmms”, “turns with laughter”, “turns with face touch”, “doubt faces”, and “turns with pen fiddle”. As with the argument variables, the coding of the non-verbal/paralinguistic responses also varies slightly from what was reported by Haidt et al. (2000). In Study 1 we coded for both verbal hesitations (“um/em/uh”) and non-verbal hesitations/stuttering. “Turns” was coded independently of other behaviours as changing position. Laughter was coded for independently of changing position, The coding of hands touching the self was not limited to the face. Participants did not have pens to fiddle with, however generic fidgeting was coded for. The term “doubt faces” presented as problematic to code for rigorously across different individuals. Two objectively verifiable and opposing facial expressions were coded for: smiling and frowning.

3.2.2.1 Dumbfounded versus reasons. Fifty-nine cases of participants providing reasons, were compared with 32 cases of dumbfounded responding. There was no difference in time until first judgement between the dumbfounded group, (M = 14.60, SD = 20.14) and the group who provided reasons (M = 15.09, SD = 40.19), F(1, 89) = .004, p = .949, partial η^2 = .00005. Similarly, there was no difference in time until first argument between the dumbfounded group, (M = 35.71, SD = 27.70) and the group who provided reasons (M = 29.82, SD = 31.61), F(1, 89) = 0.782, p = .379, partial η^2 = .009. There was no difference in time from first judgement to time of first argument between the dumbfounded group, (M = 21.11, SD = 35.47) and the group who provided reasons (M = 14.73, SD = 44.97), F(1, 89) = .480, p = .490, partial η^2 = .005.

There was a significant difference in frequency (per minute) of utterances whereby participants were working towards a reason between the dumbfounded group, (*M* = 1.53, *SD* = 1.40) and the group who provided reasons (*M* = 2.73, *SD* = 1.47), *F*(1, 89) = 14.441, *p* < .001, partial η^2 = .140. There was no difference in frequency (per minute) of irrelevant arguments between the dumbfounded group, (*M* = 1.03, *SD* = .74) and the group who provided reasons (*M* = .86, *SD* = .77), *F*(1, 89) = 1.051, *p* = .308, partial η^2 = .012. There was a significant difference in frequency (per minute) of expressions of doubt between the dumbfounded group, (*M* = .63, *SD* = .65) and the group who provided reasons (*M* = .31, *SD* = .58), *F*(1, 89) = 5.868, *p* = .017, partial η^2 = .062.

A one-way ANOVA revealed a significant difference in number of times per minute participants laughed between the dumbfounded group, (*M* = 2.81, *SD* = 2.84) and the group who provided reasons (*M* = 1.18, *SD* = 1.25), *F*(1, 89) = 14.355, *p* < . 001, partial η^2 = .139. Similarly, A one-way ANOVA revealed a significant difference between groups in the relative amount of time spent smiling (as a proportion of the total time spent on the given scenario) between the dumbfounded group, (*M* = .32, *SD* = .15) and the group who provided reasons (*M* = .16, *SD* = .14), *F*(1, 89) = 25.243, *p* < .001, partial η^2 = .221. Consistent with the results reported by Haidt et al., a series of one-way ANOVAs revealed no differences in verbal hesitations, *F*(1, 89) = 2.348, *p* = .129, partial η^2 = .026, non-verbal hesitations, *F*(1, 89) = 3.264, *p* = .074, partial η^2 = .035, changing posture (*F*(1, 89) = .491, *p* = .485, partial η^2 = .005, hands on the self, *F*(1, 89) = .030, *p* = .864, partial η^2 = .0003, frowning, *F*(1, 89) = .003, *p* = .958, partial η^2 = .00003, and fidgeting, *F*(1, 89) = 1.660, *p* = .201, partial η^2 = .018. A one-way ANOVA revealed a significant difference between groups in relative amount of time spent in silence (as a proportion of the total time spent on the given scenario) between the dumbfounded group, (*M* = .14, *SD* = .08) and the group who provided reasons (*M* = .09, *SD* = .06), *F*(1, 89) = 9.721, *p* = .002, partial η^2 = .098.

From the above analysis, it appears that, working towards reasons, expressions of doubt, laughter, smiling, and silence were the only measures that varied significantly depending on whether a person was identified as dumbfounded or provided reasons. Having identified differences between dumbfounded participants and participants providing reasons, the following analysis investigates if there are differences depending the type of dumbfounded response provided. participants who did not rate the behaviour as wrong are also included in the following analysis.

3.2.2.2 Variation between different types of dumbfounded responses. Four groups of cases, based on overall reaction to scenarios, were identified: cases of participants who did not rate the behaviour as wrong (*N* = 33), cases of participants providing reasons (*N* = 59), cases of participants providing unsupported declarations (*N* = 11), and cases of participants admitting to not having reasons (*N* = 21).

A one-way ANOVA revealed a significant difference in relative frequency of utterances whereby participants were working towards a reason depending on overall reaction to scenarios, *F*(3, 120) = 7.459, *p* < .001, partial η^2 = .157. Tukey's post-hoc pairwise comparison revealed that participants who provided reasons were identified as working towards a reason significantly more frequently (*M* = 2.74, *SD* = 1.47) than participants who did not rate the behaviour as wrong (*M* = 1.85, *SD* = 1.40), *p* = .023, and more frequently than participants who provided unsupported declarations as justifications (*M* = .79, *SD* = .59), *p* < .001. There was no difference between participants who admitted to not having reasons (*M* = 1.90, *SD* = 1.56) and any of the other groups. A one-way ANOVA revealed no significant difference in relative frequency of expressions of doubt depending on overall reaction to scenarios, $F(3, 120) = 2.166$, $p = .096$, partial $\eta^2 = .051$.

A one-way ANOVA revealed a significant difference in relative frequency of laughter depending on overall reaction to scenarios, $F(3, 120) = 8.269$, $p < .001$, partial $\eta^2 = .171$. Tukey's post-hoc pairwise comparison revealed that participants who admitted to not having reasons laughed significantly more frequently ($M = 2.41$, $SD = 2.00$), than participants who provided reasons ($M = 1.18$, $SD = 1.25$), $p = .039$, and more frequently than participants who provided did not rate the behaviour as wrong ($M = .97$, $SD = 1.29$), $p = .025$. Similarly participants who provided unsupported declarations laughed significantly more frequently ($M = 3.57$, $SD = 4.00$), than participants who provided reasons ($M = 1.18$, $SD = 1.25$), $p < .001$, and more frequently than participants who did not rate the behaviour as wrong ($M = .97$, $SD = 1.29$), $p < .001$. There was no difference between participants who provided reasons ($M = 1.18$, $SD = 1.25$), and participants who did not rate the behaviour as wrong ($M = .97$, $SD = 1.29$), $p = .951$. Interestingly, there was no difference between participants who admitted to not having reasons ($M = 2.41$, $SD = 2.00$) and participants who provided unsupported declarations ($M = 3.57$, $SD = 4.00$), $p = .305$.

A similar pattern of results was found for time spent smiling. A one-way ANOVA revealed a significant difference in relative time spent smiling depending on overall reaction to scenarios, $F(3, 120) = 9.975$, $p < .001$, partial $\eta^2 = .200$. Tukey's post-hoc pairwise comparison revealed that participants who admitted to not having reasons spent significantly more time smiling ($M = .33$, $SD = .14$), than participants who provided reasons ($M = .16$, $SD = .14$), $p < .001$, and more time smiling than participants who provided did not rate the behaviour as wrong ($M = .16$, $SD = .13$), p

$< .001$. Participants who provided unsupported declarations spent significantly more time smiling ($M = .31$, $SD = .17$), than participants who provided reasons ($M = .16$, $SD = .14$), $p = .008$, and more time than participants who did not rate the behaviour as wrong ($M = .16$, $SD = .13$), $p = .014$. There was no difference between participants who provided reasons, and participants who did not rate the behaviour as wrong, $p > .999$. Again, there was no difference between participants who admitted to not having reasons ($M = .33$, $SD = .14$) and participants who provided unsupported declarations ($M = .31$, $SD = .17$), $p = .996$.

A one-way ANOVA revealed a significant difference in relative amount of time spent in silence depending on overall reaction to scenarios, $F(3, 120) = 3.305$, $p = .023$, partial $\eta^2 = .076$. Mean proportion of interview time spent in silence are as follows: participants providing reasons, $M = .09$, $SD = .06$; participants not rating the behaviour as wrong, $M = .12$, $SD = .07$; participants admitting to not having reasons, $M = .14$, $SD = .09$; and participants providing unsupported declarations, $M = .14$, $SD = .05$. Tukey's post-hoc pairwise comparison did not reveal any significant differences between specific groups.

3.2.2.3 Further analyses. An exploratory analysis revealed no association between number of times a participant was dumbfounded and their score on either measures from the MLQ: Presence, $r(31) = .14$, $p = .466$, or Search, $r(31) = .25$, $p = .179$, or the Centrality of Religiosity Scale $r(31) = .07$, $p = .726$. There was no difference in observed rates of dumbfounded responses depending on the order of scenario presentation, $\chi^2(6, N = 124) = 4.01$, $p = .676$. Rates of dumbfounded responses varied depending on which moral dilemma was being discussed, $\chi^2(6, N = 124) = 46.82$, $p < .001$. The highest rate of dumbfounding was recorded for *Incest*, with 18 of the 31 (58.06%) participants displaying dumbfounded responses. Eleven participants (35.48%) displayed dumbfounded responses for *Cannibal* and three participants (9.67%) displayed dumbfounded responses for *Trolley*. The lowest recorded rate of dumbfounded response was for the *Heinz* dilemma, with no participants resorting to unsupported declarations as justification or admitting to not having reasons for their judgement. This trend is generally consistent with that which emerged in the original study (with the exception of *Trolley,* which was not used in the original study). Furthermore, rates of dumbfounded responding varied depending on which type of moral scenario was being discussed. *Heinz* and *Trolley,* identified as reasoning scenarios, were contrasted against the intuition scenarios *Incest* and *Cannibal*. There was significantly more dumbfounded responding for the intuition scenarios (29 instances) than for the reasoning scenarios (3 instances), $\chi^2(1, N = 124) = 38.17, p < .001$. An alternative explanation for this variation between the scenarios may be found by classing the scenarios according to the level of moral disgust (see Russell & Giner-Sorolla, 2013) that may be elicited by each scenario. According to this approach, moral disgust is elicited by the violation of “bodily norms”, e.g., norms involving immoral eating or sex acts (Russell & Giner-Sorolla, 2013, p. 337). Both the “intuition” (Haidt et al., 2000) scenarios contain such acts and as such likely elicit moral disgust to a greater degree than the “reasoning” scenarios. Further research is required to test this possibility.

The aim of Study 1 was to examine the replicability of moral dumbfounding as identified by Haidt et al. (2000), and identify specific measurable responses that may be indicative of dumbfounding. The overall pattern of responses, and pattern of inter-scenario variability in responding resembled that observed in the original study. Study 1 successfully replicated the findings of the original moral dumbfounding study (Haidt et al., 2000). Participants were identified as dumbfounded according to two specific measures, admissions of having no reasons, and unsupported declarations followed by a failure to provide reasons when questioned further. Both of these responses were accompanied by similar increases in incidences of laughter, and time spent smiling, when compared to participants providing reasons, and participants not rating the behaviour as wrong. When taken together, these responses were also accompanied by more silence during the interview, when compared with participants who provided reasons. It appears that identifying incidences of dumbfounding according to unsupported declarations or admissions of not having reasons largely captures dumbfounding as described by Haidt et al. (2000).

Study 1 provides evidence supporting the view that moral dumbfounding is a genuine phenomenon and can be elicited in an interview setting when participants are pressed to justify their judgements of particular moral scenarios. Two key limitations have been identified as a result of conducting studies in an interview setting. Firstly, conducting video-recorded interviews, and the accompanying analyses, is particularly labour intensive, which leads to a smaller sample size – indeed the limited sample size previously identified as a limitation of the original study has not been improved upon in Study 1. The aims of the present study were to examine the replicability of dumbfounding, and to identify specific measurable indicators of dumbfounding. A sample size of thirty-one is not sufficient in fulfilling the first aim. Secondly, an interview setting introduces a social context that may influence the responses of participants, in that, participants may feel a social pressure to behave in a particular way (see Royzman et al., 2015). Alternative methods are required to examine dumbfounding with a larger sample, and whether it still occurs in the absence of the social pressure that is present in an interview setting. Two responses have been identified as indicators of dumbfounding. The degree to which each of these responses can be elicited in a setting other than an interview is investigated in Studies 2 and 3.

3.3 Study 2: Initial Computerised Task

Having successfully elicited dumbfounded responses in a video recorded interview with a small sample, the aim of Study 2 was to devise methods that might elicit dumbfounding in a systematic way, using standardised materials and procedure that can be administered without the need for an interviewer. It has been argued that dumbfounded responding emerges as a result of social pressure (e.g., Royzman et al., 2015). Developing means for studying moral dumbfounding without the need for an interviewer will eliminate social pressure as a potential confound. That is, it will eliminate participant-interviewer interaction as a source of possible variability, remove the social pressure associated with an interview setting, while also enabling the study to be conducted with a larger sample. It was hypothesised that presenting participants with the same dilemmas and counter-arguments as in Study 1 as part of a computer task, as opposed to in an interview, would lead to a similar state of dumbfoundedness as found in Study 1. However, a major challenge to this alternative medium of conducting the study is identifying specific behavioural responses that are indicative of a state of dumbfoundedness that can be elicited and recorded. Without the benefit of an experimenter to guide the discussion, and a video recording that can be analysed, this challenge was addressed by developing a *critical slide* (described below). Scenarios and counter-arguments to commonly made judgements were presented on a sequence of slides before participants were asked to describe their judgement on a forced choice critical slide. Participants were identified as dumbfounded if they selected an unsupported declaration from a selection of three possible responses present on the critical slide, or if they provided an unsupported declaration as a reason.

3.3.1 Method.

3.3.1.1 Participants and design. Study 2 was a frequency-based, conceptual replication of Study 1. The aim was to identify if dumbfounded responding could be evoked via a computer-based task. All participants were presented with the same four moral vignettes. Results are primarily descriptive. Further analysis tested for differences in responding depending on the vignette, or type of vignette, presented.

A sample of 72 participants (52 female, 20 male) with a mean age of M_{age} = 21.18 (min = 18, max = 50, *SD* = 5.21) took part in this study. Participants were undergraduate students and postgraduate students from MIC. Participation was voluntary and participants were not reimbursed for their participation.

3.3.1.2 Procedure and materials. This study used largely the same materials as in Study 1. The four vignettes from Study 1 *Heinz*, *Incest*, *Cannibal*, and *Trolley* (Appendix A) along with the same prepared counter arguments (Appendix B) were used. Dumbfounding was measured using responses to the critical slide. The critical slide contained a statement defending the behaviour and a question as to how the behaviour could be wrong (e.g., “Julie and Mark’s behaviour did not harm anyone, how can there be anything wrong with what they did?”). There were three possible answer options: (a) “There is nothing wrong”; (b) an unsupported declaration, naming the specific behaviour described in the scenario (e.g., “Incest is just wrong”); and finally a judgement with accompanying justification (c) “It’s wrong and I can provide a valid reason”. The order of these response options was randomised. Participants who selected (c) were then prompted on a following slide to type a reason. The selecting of option (b), the unsupported declaration, was taken to be a dumbfounded response, as was the use of an unsupported declaration as a justification for option (c).

This study made use of the same post-discussion questionnaire as in Study 1 (Appendix C). This was administered after the critical slide for each scenario. There was a change to one of the questions on this post-discussion questionnaire: the question asking if participants had changed their judgements was changed from "how much did your judgement change?" with a 7-point Likert scale response to "did your judgement change?" with a binary "yes/no" response option. Both MLQ (Steger et al., 2008) and CRSi7 taken from The Centrality of Religiosity Scale (Huber & Huber, 2012) were also used.

OpenSesame was used to present the vignettes and collect responses (Mathôt, Schreij, & Theeuwes, 2012). The same four moral dilemmas (Appendix A) as in Study 1 were presented to participants (in randomised order). Following the presentation of each dilemma, participants were asked to judge, on a 7-point Likert scale how right or wrong they would rate the behaviour of the characters in the given scenario. After making a judgement participants were then presented with a series of counter-arguments (Appendix B). Following these counter-arguments, participants were presented with the critical slide. Following the critical slide participants completed the same brief questionnaire as in Study 1 (between scenarios) in which they were asked to rate, on a 7-point Likert scale, how right/wrong they thought the behaviour was; how confused they were; how irritated they were; how much their judgement had changed; how much their judgement was based on reason; and how much their judgement was based on "gut" feeling. When participants had completed all questions relating to all four dilemmas they completed the same longer questionnaire as in Study 1 containing the MLQ (Steger et al., 2008), the Centrality of Religiosity Scale (Huber & Huber, 2012), and some questions relating to demographics. The entire study lasted approximately fifteen to twenty minutes.

3.3.2 Results and discussion. Participants who selected the unsupported declaration on the critical slide were identified as dumbfounded. Table 3.4 shows the ratings of the behaviours across each scenario. Table 3.5 shows the number, and percentage, of participants who displayed "dumbfounded" responses (identified as the selecting of an unsupported declaration) and non-dumbfounded responses for each dilemma. Figure 3.1 shows the percentage of participants displaying dumbfounded responses for each dilemma. Table 3.6 shows the responses to the questionnaires presented between dilemmas. The open-ended responses provided by participants who selected option (c) "It's wrong and I can provide a valid reason" were analysed and coded, by the primary researcher, and unsupported declarations provided here were also identified as dumbfounded responses. Following this coding, one additional participant was identified as dumbfounded for *Trolley*. Sixty-eight of the 72 participants (94%) selected the unsupported declaration at least once. There was no statistically significant difference in responses to the critical slide depending on the order of scenario presentation, $\chi^2(6, N = 288) = 4.50, p = .610$. There was no statistically significant difference in responses to the critical slide depending on scenario presented, $\chi^2(6, N = 288) = 9.13, p = .167$. Rates of dumbfounded responding did not vary with type of moral scenario (100 instances for intuition scenarios, 91 instances for reasoning scenarios) being discussed, $\chi^2(1, N = 288) = 1.26, p = .262$. Forty-five participants (62.50%) selected the unsupported for *Heinz*. Forty-six participants (63.89%) selected (or provided) the unsupported declaration for *Cannibal* and *Trolley*. Fifty-four participants (75%) selected the unsupported declaration for *Incest*. There was no association between number of times dumbfounded and score on either measure on the Meaning and Life questionnaire; Presence, *r*(72) = -.05, *p* = .662, or Search, *r*(72) = .13, *p* = .268, or the Centrality of Religiosity Scale *r*(72) = .17, *p* = .146.

Table 3.4: Study 2: Initial and revised ratings for each scenario

	Heinz		Trolley		Cannibal		Incest	
Judgement	Count	%	Count	%	Count	%	Count	%
Initial judgement wrong	53	73.6%	50	64.9%	68	94.4%	63	87.5%
Initial judgement neutral	9	12.5%	6	8.3%	3	4.2%	3	4.2%
Initial judgement ok	10	13.9%	16	22.2%	1	1.4%	6	8.3%
Revised judgement wrong	51	70.8%	48	66.7%	67	93.1%	66	91.6%
Revised judgement neutral	7	9.7%	9	12.5%	3	4.2%	3	4.2%
Revised judgement ok	14	19.4%	15	20.8%	2	2.8%	3	4.2%

Table 3.5: Study 2: Observed frequency and percentage of each of the responses: dumbfounded, nothing wrong, and reasons provided.

	Heinz		Cannibal		Incest		Trolley	
Response	N	%	N	%	N	%	N	%
Nothing wrong	8	11.11%	4	5.56%	2	2.78%	10	13.89%
Dumbfounded	45	62.50%	46	63.89%	54	75.00%	46	63.89%
Reasons	19	26.39%	22	30.56%	16	22.22%	17	23.61%

Table 3.6: Study 2: Responses to post-discussion questionnaire questions (7-point Likert scale, 1 = Not at all, 7 = Extremely).

	Heinz	Cannibal	Incest	Trolley
How confident were you?	-	5.86	5.63	5.26
How confused were you?	2.40	3.08	4.14	3.17
How irritated were you?	4.58	4.68	4.32	4.28
How much was your judgement based on "gut" feeling?	5.29	5.54	5.82	4.96
How much was your judgement based on reason?	4.89	5.19	4.89	4.93
Gut minus reason	.40	.35	.93	.03

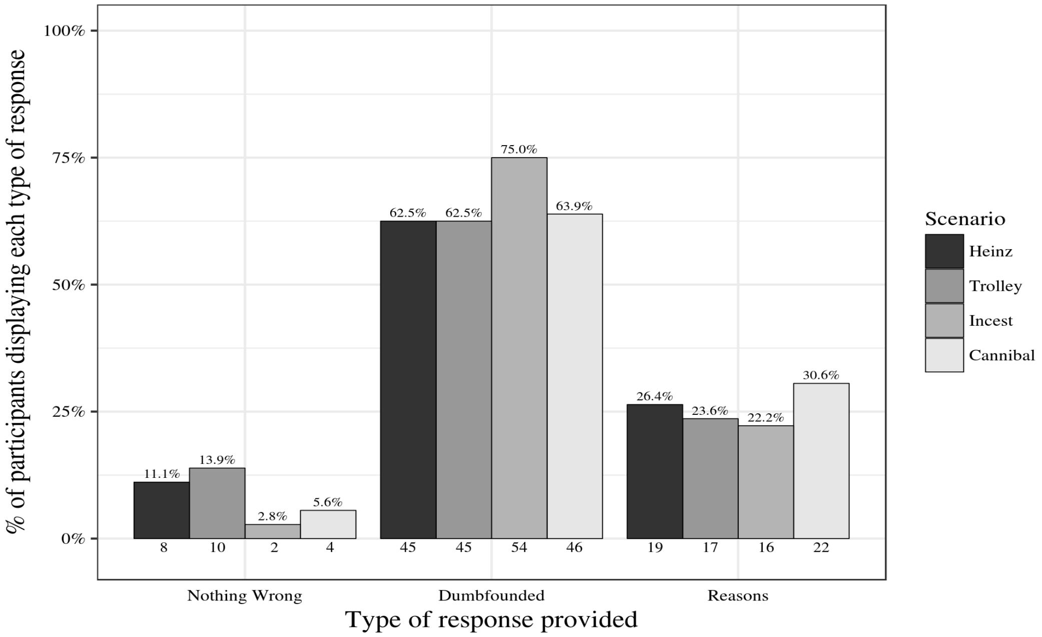

Figure 3.1: Study 2: Percentage of participants selecting each type of response on the critical slide

The most striking result from this study was the willingness of participants to select the unsupported declaration in response to a challenge to their judgement. This is inconsistent with what was found in both Study 1 and in the original study by Haidt et al. (2000). In these studies, participants did not readily offer an unsupported declaration as justification for their judgement, rather it was a last resort following extensive cross-examining. The exceptionally high rates of dumbfounding observed in Study 2 do not appear to be representative of the phenomenon more generally. There is, therefore, clearly a difference between offering an unsupported declaration as a justification for a judgement during an interview and selecting an unsupported declaration from a list of possible response options during a computerised task.

There are two possible explanations for this. Firstly, it is possible that, during the interview, participants experienced a social pressure to successfully justify their judgement. This social pressure may also have made participants were more aware of the illegitimacy of using an unsupported declaration as a justification for their judgement. Secondly is also possible that the measure used (the selecting of an unsupported declaration) contributed to the high rates of seemingly dumbfounded responding. Seeing an unsupported declaration written down as a possible answer legitimises selecting it as a justification for the judgement. The unsupported declaration does not provide an acceptable answer to the question on the critical slide, however, its presence in the list of possible response options may imply to participants that it is an acceptable answer, particularly if they do not put too much thought into it. By selecting the unsupported declaration participants can move quickly along to the next stage in the study without necessarily acknowledging any inconsistency in their reasoning, avoiding potentially dissonant cognitions (e.g., Case, Andrews, Johnson, & Allard, 2005; Harmon-Jones & Harmon-Jones, 2007; see also Heine, Proulx, & Vohs, 2006). Selecting the unsupported declaration may also allow the participant to proceed without expending effort trying to think of reasons for their judgement beyond the intuitive justifications that had already been de-bunked.

Rates of dumbfounded responding in Study 2 were higher than expected. Possible reasons for this are (a) reduced social pressure to appear to have reasons for judgements; (b) a failure of participants to comprehend that the unsupported declaration does not provide a logically justifiable response to the question asked in the critical slide; (c) the apparent legitimising of the unsupported declaration by its inclusion in the list of possible response options; or (d) the selecting by participants of an "easy way out" option without thinking about it fully (through carelessness/laziness/eagerness to move on to a less taxing task). A follow-up study may be useful in identifying which of these reasons (a) to (d) led to the unusually high rates of dumbfounded responding. Devising an alternative measure that addresses (b), (c), and (d) as possible reasons for the observed results in Study 2 would provide a more robust measure of moral dumbfounding. This could then be used to investigate the degree to which (a) influenced dumbfounded responding.

It appears likely that the selecting of unsupported declarations in this instance is not an accurate measure of dumbfounding. In Study 1, participants were only identified as dumbfounded based on the providing of an unsupported declaration if they subsequently failed to provide further reasons when the unsupported declaration was questioned. However, in some cases, participants who provided unsupported declarations were not identified as dumbfounded, based on subsequent responses. A follow up analysis of the interview data revealed that 23 of the 31 participants provided an unsupported declaration and proceeded to provide reasons for at least one of their judgements; a further six participants provided an unsupported declaration and proceeded to revise their judgement at least once. A stricter measure of dumbfounding, one by which participants are required to explicitly acknowledge a state of dumbfoundedness is necessary to address the issues with the selecting of an unsupported declaration that may have led to the unusually high rates of dumbfounding observed in Study 2.

3.4 Study 3a: Revised Computerised Task – College sample

Study 3a was designed in response to the unexpectedly high rates of observed dumbfounding in Study 2. Four limitations of the use of the unsupported declaration selection as a measure of dumbfounding were identified. It was hypothesised that replacing the unsupported declaration with an explicit admission of not having reasons would address each of these limitations, and bring the option selection more in line with conversational logic, making participants less willing to casually select the dumbfounded response. Making participants explicitly acknowledge the absence of reasons for their judgement means that their selecting of a dumbfounded response cannot be attributed to a mere misunderstanding and thus, might provide a truer measure of dumbfounding.

3.4.1 Method.

3.4.1.1 Participants and design. Study 3a was a frequency based, modified replication. The aim was to identify if dumbfounded responding could be evoked. All participants were presented with the same four moral vignettes. Results are primarily descriptive. Further analyses tested for differences in responding depending on the vignette, or type of vignette, presented.

A sample of seventy-two participants (46 female, 26 male) with a mean age of M_{age} = 21.80 (min = 18, max = 46, *SD* = 3.92) took part in this study. Participants were undergraduate students and postgraduate students from MIC. Participation was voluntary and participants were not reimbursed for their participation.

3.4.1.2 Procedure and materials. The materials in this study were almost the same as in Study 2 with a change to the "dumbfounded" response option on the critical slide. Extra questions were included following each of the counter-arguments. On the critical slide, the unsupported declaration option was replaced with an admission of not having reasons ("It's wrong but I can't think of a reason"). Following each counter-argument, participants were asked if they (still) thought the behaviour was wrong, and if they had a reason for their judgement. There was also a revision to the question on the post-discussion questionnaire asking if participants had changed their judgements: "did your judgement change?" with a binary "yes/no" response option reverted back to "how much did your judgement change?" with a 7-point Likert scale response (1 = *Not at all*, 7 = *Extremely*) as in Study 1. The same four dilemmas *Heinz*, *Incest*, *Cannibal* and *Trolley* (Appendix A) along with the same prepared counter arguments (Appendix B) as in Study 2 were used in Study 3a. Both the MLQ (Steger et al., 2008); and CRSi7 (Huber & Huber, 2012) were also used. This study was conducted in a designated psychology computer lab in MIC and was administered entirely on individual computers using OpenSesame (Mathôt et al., 2012).

Participants were seated, given instructions, and allowed to begin the computer task. The four vignettes from Study 1, *Heinz*, *Incest*, *Cannibal* and *Trolley* (Appendix A) along with the same pre-prepared counter arguments (Appendix B) were used. Dumbfounding was measured using the critical slide. The updated critical slide contained a statement defending the behaviour and a question as to how the behaviour could be wrong (e.g., “Julie and Mark’s behaviour did not harm anyone, how can there be anything wrong with what they did?”) with three possible response options: (a) “There is nothing wrong”; (b) “It’s wrong, but I can’t think of a reason”; (c) “It’s wrong and I can provide a valid reason”. The order of these response options was randomised. Participants who selected (c) were required to provide a reason. The selecting of option (b), the admission of not having reasons, was taken to be a dumbfounded response. When participants had completed all questions relating to all four dilemmas they completed the same longer questionnaire as in Studies 1 and 2 containing the MLQ (Steger et al., 2008), the Centrality of Religiosity Scale (Huber & Huber, 2012), and some questions relating to demographics. The entire study lasted approximately fifteen to twenty minutes.

3.4.2 Results and discussion. Participants who selected the admission of not having reasons on the critical slide (option b) were identified as dumbfounded. Forty of the 72 participants (55%) selected the admission of not having reasons at least once. Table 3.7 shows the ratings of the behaviours across each scenario. Table 3.8 and Figure 3.2 show the percentage of participants displaying dumbfounded responses for each dilemma (the percentage of participants providing each response when the coded string responses are included in the analysis are shown in Figure 3.3). Table 3.9 shows the responses to the questionnaires presented between dilemmas. Again there was no statistically significant difference in responses to the critical slide depending on the order of scenario presentation, $\chi^2(6, N = 288) = .61$, $p = .996$. There was no difference in responses to the critical slide depending on scenario, $\chi^2(6, N = 288) = 9.60$, $p = .143$, or, type of scenario (32 instances for intuition scenarios, 27 instances for reasoning scenarios), $\chi^2(1, N = 288) = .53$, $p = .465$. Thirteen participants (18.06%) selected the admission of having no reasons for *Heinz*. Fourteen participants (19.44%) selected the admission of not having reasons for *Cannibal* and *Trolley* and eighteen participants (25%) selected the admission of not having reasons for *Incest*. This lack of variation between scenarios is inconsistent with the results of the interview in Study 1. Given that variability was identified as a hallmark of intuitionist approaches in Chapter 1, this finding may pose a challenge to an intuitionist explanation of moral dumbfounding. Indeed, it is likely that in an interview, with real time feedback encouraging participants to engage in deliberation, it is likely that people engage in a greater level of deliberation. That this deliberation may provide the source of variability depending on scenario is not predicted by an intuitionist (e.g., Haidt, 2001) perspective.

Table 3.7: Study 3a: Initial and revised ratings for each scenario

	Heinz		Trolley		Cannibal		Incest	
Judgement	Count	%	Count	%	Count	%	Count	%
Initial judgement wrong	54	75%	48	66.7%	67	93.1%	61	84.7%
Initial judgement neutral	6	8.3%	10	13.9%	3	4.2%	7	9.7%
Initial judgement ok	12	16.7%	14	19.4%	2	2.8%	4	5.6%
Revised judgement wrong	53	73.6%	43	59.7%	67	93.1%	57	79.2%
Revised judgement neutral	11	15.3%	15	20.8%	4	5.6%	12	16.7%
Revised judgement ok	8	11.1%	14	19.4%	1	1.4%	3	4.2%

Table 3.8: Study 3a: Observed frequency and percentage of each of the responses: dumbfounded, nothing wrong, and reasons provided.

		Heinz		Cannibal		Incest		Trolley	
Response		Count	%	Count	%	Count	%	Count	%
Nothing wrong		14	19.44 %	4	5.56%	12	16.67 %	15	20.83%
Dumbfounded	(just critical slide)	13	18.06 %	14	19.44%	18	25.00 %	14	19.44%
Dumbfounded	(including coded responses)	19	26.39 %	21	29.17%	31	43.06 %	22	30.56%
Reasons		45	62.50 %	54	75.00%	42	58.33 %	43	59.72%
Reasons	(after coding)	39	51.17 %	47	65.28%	29	40.28 %	35	48.61%

Table 3.9: Study 3a: Responses to post-discussion questionnaire questions (7-point Likert scale, 1 = Not at all, 7 = Extremely).

	Heinz	Cannibal	Incest	Trolley
How much did you change your mind?	2.38	1.67	2.00	2.00
How confident were you?	5.22	5.50	5.38	4.81
How confused were you?	2.75	2.96	3.25	2.89
How irritated were you?	3.94	4.64	4.07	3.60
How much was your judgement based on “gut” feeling?	4.78	5.44	5.44	4.92
How much was your judgement based on reason?	5.07	5.26	5.11	5.06
Gut minus reason	-.29	.18	.33	-.14

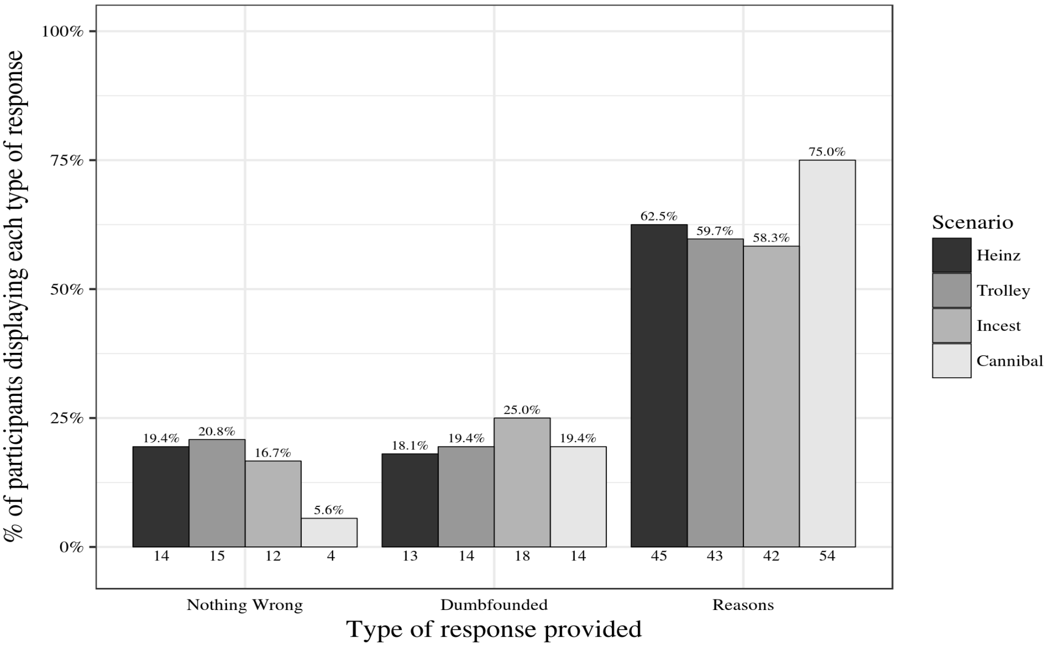

Figure 3.2: Study 3a: Percentage of participants selecting each type of response on the critical slide.

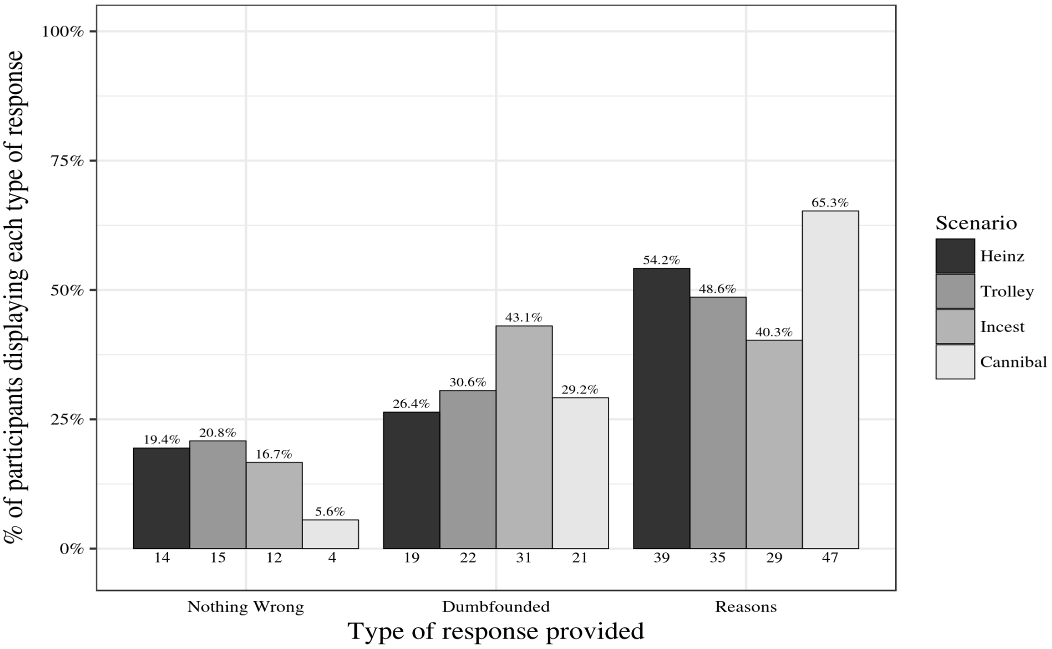

Figure 3.3: Study 3a: Percentage of participants providing each type of response when the coded string responses are included

The replacing of an unsupported declaration with an admission of having no reasons led to substantially lower rates of dumbfounding than observed in Study 2. It appears that the issues associated with the selecting of an unsupported declaration have been addressed in Study 3a. However, the rates of dumbfounding observed for *Incest* and *Cannibal* in Study 3a were considerably lower than those observed in Study 1. This suggests the revised measure may be too strict, measuring only open admissions of not having reasons, without measuring a failure to provide reasons. As in the first computerised task, participants who selected “It’s wrong and I can provide a valid reason” were then required to provide a reason. In order to provide a measure of a failure to provide reasons, these responses were analysed and coded, by the primary researcher. Those containing unsupported declarations were taken as evidence for a failure to provide a reason and identified as dumbfounded responses.

During the coding, another class of dumbfounded response was identified. Participants occasionally provided undefended tautological responses as justification for their judgements, whereby they simply named or described the behaviour in the scenario as justification for their judgement (e.g., “They are related”, “Because it is canibalism[sic]”). These responses may be viewed as largely equivalent to unsupported declarations (e.g., Mallon & Nichols, 2011). In Study 1, they were not identified as dumbfounded responses, because when provided in an interview setting, they were always followed by further questioning. This further questioning could lead to two possible responses: (a) a dumbfounded response (unsupported declaration or an admission of not having reasons) or (b) an alternative reason. A computerised task does not allow for a follow-up probe to encourage participants to elaborate on such responses. Participants were not placed under time pressure and could articulate and review their typed reason at their own pace. It is reasonable to expect then, that, if participants did have a valid reason for their judgement, they would have provided it along with, or instead of, the undefended tautological response. As such, an undefended tautological reason appears to be evidence of a failure to identify reasons. For this reason, these undefended tautological reasons were also coded as dumbfounded responses, along with the unsupported declarations. Analysis of the reasons provided indicated that participants did not provide reasons that were directly contrary to the facts presented in the scenarios. In some cases participants challenged the facts presented (e.g., for *Incest*, “Contraception does not always work”).

Table 3.8 and Figure 3.3 show the number and percentage of dumbfounded responses when the coded string responses are included in the analysis. When the coded string responses are included in the analysis, the number of participants displaying a dumbfounded response at least once increased from 40 (55%) to 57 (79%). Observed rates of dumbfounding increased for each scenario when the coded open-ended responses were included, with 19 participants (26.39%) appearing to be dumbfounded by *Heinz*, 21 (29.16%) by *Cannibal*, 31 (43.06%) by *Incest,* and 22 (30.56%) apparently dumbfounded by *Trolley*. Still, rates of dumbfounded responding did not vary with type of moral scenario (52 instances for intuition scenarios, 41 instances for reasoning scenarios) being discussed, $\chi^2(1, N = 288) = 1.92, p = .166$. There was no association between number of times dumbfounded and score on either measure on the Meaning and Life questionnaire; Presence, $r(72) = .10, p = .413$, Search $r(71) = .01, p = .945$, or the Centrality of Religiosity Scale, $r(72) = .16, p = .184$.

When the coded open-ended responses were included in the analysis, the proportion of participants displaying a dumbfounded response at least once in Study 3a (79%) was much closer to that observed in the interview in Study 1 (74%) than before the open-ended responses were included (55%). The variation in observed rates of dumbfounded responding between dilemmas that was observed in the interview was not present in the computerised task. This finding, though interesting, is as yet unexplained and may have significant theoretical implications. Further study is required to attempt to explain why there remains a difference between the dumbfounding elicited during an interview and that elicited as part of a computerised task. However, it is clear that dumbfounded responses can be elicited as part of a computerised task. The participants in Studies 1, 2, and 3a were all college students (largely from the same institution). To address this limitation, the following study investigated the phenomenon in a more diverse sample.

3.5 Study 3b: Revised Computerised Task – MTurk

Having successfully elicited dumbfounded responses in a college sample using a computerised task in Study 3a, Study 3b was conducted in an attempt to replicate Study 3a using more diverse sample using online recruiting through MTurk (Amazon Web Services Inc., 2016).

3.5.1 Method.

3.5.1.1 Participants and design. Study 3b was a frequency based, modified replication. The aim was to identify if dumbfounded responding could be evoked. All participants were presented with the same four moral vignettes. Results are primarily descriptive. Further analysis tested for differences in responding depending on the vignette, or type of vignette, presented.

A sample of one hundred and one participants (53 female, 47 male, 1 other; M_{age} = 36.58, min = 18, max = 69, *SD* = 12.50) took part in this study. Participants were recruited online through MTurk (Amazon Web Services Inc., 2016). Participation was voluntary and participants were paid US $0.70 for their participation. Participants were recruited from English speaking countries or from countries where residents generally have a high level of English (e.g., The Netherlands, Denmark, Sweden). Location data for individual participants was not recorded, however, based on other studies, using the same selection criteria, it is likely that 90% of the sample was from the United States.

3.5.1.2 Procedure and materials. The materials in this study were almost the same as in Study 3a, however, a different software package was used to present the materials and collect the responses. OpenSesame (Mathôt et al., 2012) was replaced with Questback (Unipark, 2013) in order to facilitate online data collection. This meant that the recording of responses changed from keyboard input to mouse input.

It also allowed for multiple questions to be displayed on the screen at the same time. Other than these changes, the materials were the same as in Study 3a.

The computer task in Study 3b was much the same as Study 3a. The four vignettes from Study 1: *Heinz*, *Incest*, *Cannibal*, and *Trolley* (Appendix A) along with the same pre-prepared counter arguments (Appendix B). Dumbfounding was measured using the critical slide.

The critical slide contained a statement defending the behaviour and a question as to how the behaviour could be wrong, with three possible response options: (a) “There is nothing wrong”; (b) “It’s wrong but I can’t think of a reason”; (c) “It’s wrong and I can provide a valid reason”. Participants who selected (c) were required to provide a reason. The order of these response options was randomised. When participants had completed all questions relating to all four dilemmas they completed the same longer questionnaire as in Studies 1 and 2 containing the Meaning and Life questionnaire (Steger et al., 2008), the Centrality of Religiosity Scale (Huber & Huber, 2012), and some questions relating to demographics. The entire study lasted approximately fifteen to twenty minutes.

3.5.2 Results and discussion. Participants who selected the admission of not having reasons on the critical slide (option b) were identified as dumbfounded. Table 3.10 shows the ratings of the behaviours across each scenario. Table 3.11 and Figure 3.4 show the percentage of participants displaying dumbfounded responses for each scenario (open ended responses included in Figure 3.5). Table 3.12 shows the responses to the questionnaires presented between scenario.

Table 3.10: Study 3b: Initial and revised ratings for each scenario

	Heinz		Trolley		Cannibal		Incest	
Judgement	Count	%	Count	%	Count	%	Count	%
Initial judgement wrong	81	80.2%	66	65.4%	85	84.2%	71	70.3%
Initial judgement neutral	9	8.9%	14	13.9%	13	12.9%	20	19.8%
Initial judgement ok	11	10.9%	21	20.8%	3	3.0%	10	9.9%
Revised judgement wrong	87	86.1%	59	58.4%	82	81.2	73	72.3%
Revised judgement neutral	10	9.9%	17	16.8%	15	14.9%	19	18.8%
Revised judgement ok	4	4.0%	25	24.8%	4	4.0%	9	8.9%

Table 3.11: Study 3b: Observed frequency and percentage of each of the responses: dumbfounded, nothing wrong, and reasons provided.

		Heinz		Cannibal		Incest		Trolley	
Response		Count	%	Count	%	Count	%	Count	%
Nothing wrong		21	20.79%	10	9.90%	31	30.69%	24	23.76%
Dumbfounded	(just critical slide)	12	11.88%	19	18.81%	16	15.84%	16	15.84%
Dumbfounded	(including coded responses)	16	15.84%	30	29.70%	28	27.72%	22	21.78%
Reasons		68	67.33%	72	71.29%	54	53.47	61	60.40%
Reasons	(after coding)	64	63.37%	61	60.40%	42	41.58%	55	54.46%

Table 3.12: Study 3b: Responses to post-discussion questionnaire questions (7-point Likert scale, 1 = Not at all, 7 = Extremely).

	Heinz	Cannibal	Incest	Trolley
How much did you change your mind?	1.74	1.60	1.57	1.83
How confident were you?	5.78	6.16	5.81	5.36
How confused were you?	2.06	2.07	2.12	2.22
How irritated were you?	4.42	4.01	3.56	3.39
How much was your judgement based on "gut" feeling?	4.42	4.43	4.47	4.01
How much was your judgement based on reason?	5.46	5.69	5.26	5.58
Gut minus reason	-1.04	-1.27	-.79	-1.57

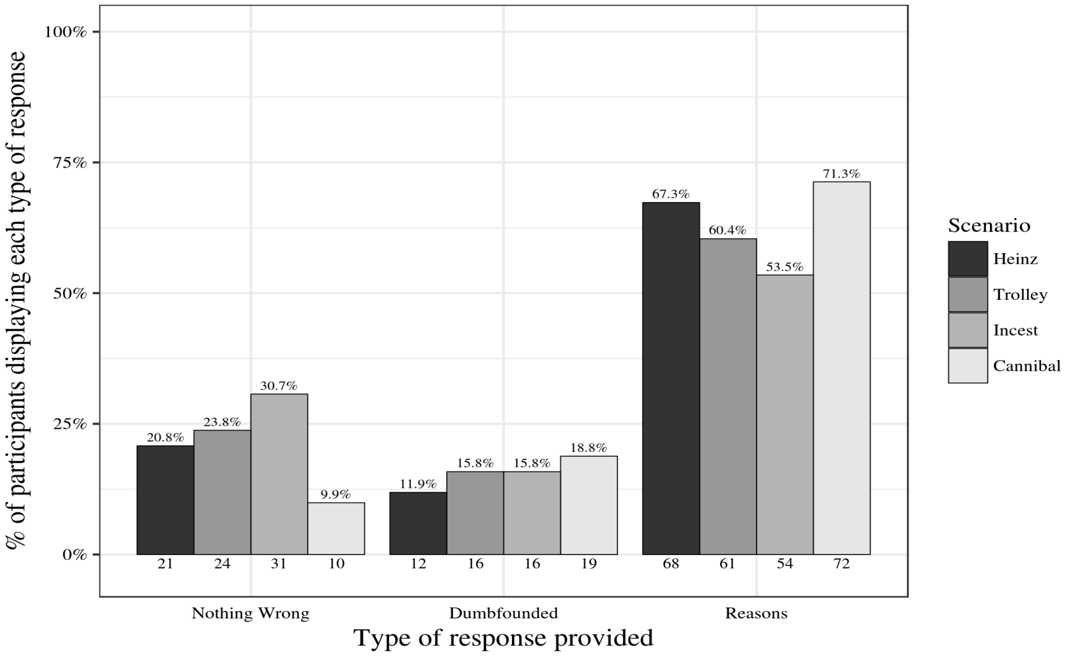

Figure 3.4: Study 3b: Percentage of participants selecting each type of response on the critical slide.

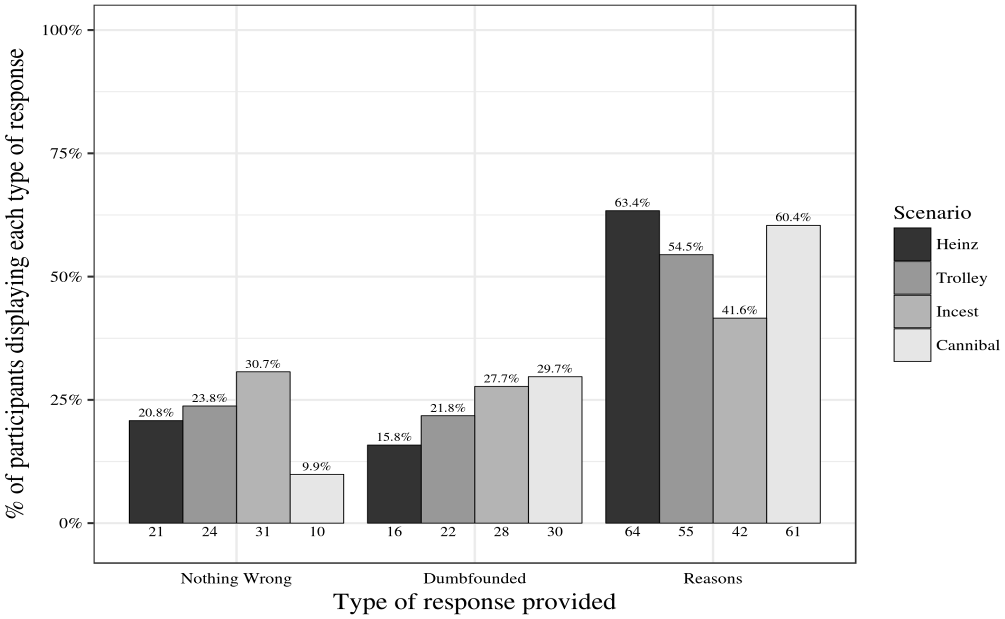

Figure 3.5: Study 3b: Percentage of participants providing each type of response when the coded string responses are included.

On this occasion there was a statistically significant difference in responses to the critical slide depending on the order of scenario presentation, $\chi^2(6, N = 404) = 14.77, p = .022$. The observed rates of dumbfounded responses were higher for the third scenario, however they went down again for the fourth scenario along with rates of selecting “nothing wrong”, meaning that the rates of participants providing reasons went up again for the fourth scenario. The higher rates of providing reasons observed for the fourth scenario presented means that this fluctuation is unlikely to be due to experimental fatigue, which was the primary reason for testing for order effects. There was also a difference in responses to the critical slide depending on scenario, $\chi^2(6, N = 404) = 15.18, p = .019$. This appeared to be due to a lower number of participants selecting “There is nothing wrong” for *Cannibal* (10), and a higher number of participants selecting “There is nothing wrong for *Incest* (31). The numbers of participants selecting “There is nothing wrong” for *Trolley* and *Heinz* were 24 and 21 respectively. The response of interest was the dumbfounded response. Dumbfounded responses were isolated and compared against the combined selecting “There is nothing wrong” and providing reasons. This analysis found no difference in dumbfounded responding depending on scenario, $\chi^2(3, N = 404) = 1.86, p = .602$. Forty-four participants (44%) selected the admission of not having reasons at least once. Twelve participants (11.88%) selected the admission of not having reasons for *Heinz*. Sixteen participants (15.84%) selected the admission of not having reasons for *Incest* and *Trolley* and 19 participants (18.81%) selected the admission of not having reasons for *Cannibal*. Again, this highlights a difference between the interview in Study 1 and the computerised task and it appears that increased deliberation in an interview leads to greater variability.

As in Study 3a, participants who selected option (c) "It's wrong and I can provide a valid reason", were there then required to provide a reason through open-ended response. These open-ended responses were coded, by the primary researcher, for dumbfounded responses, again, identified as unsupported declarations or as undefended tautological responses. Table 3.11 and Figure 3.7 show the rates of observed dumbfounding when the coded open-ended responses were included in the analysis. As expected, the number of participants displaying a dumbfounded response at least once increased, from 44 (44%) to 57 (56%). Observed rates of dumbfounding increased for each scenario when the coded reasons were included with 16 participants (15.84%) appearing to be dumbfounded by *Heinz*, 30 (29.70%) by *Cannibal*, 28 (27.72%) by *Incest* and 22 (21.78%) apparently dumbfounded by *Trolley*. Taking these revised rates of dumbfounding there was a no significant difference in rates of dumbfounded responding depending on scenario, $\chi^2(3, N = 404) = 6.56, p = .087$. There was however, significantly more dumbfounded responding for the intuition scenarios (58 instances) than for the reasoning scenarios (38 instances), $\chi^2(1, N = 404) = 4.93, p = .026$. Again, analysis of the reasons provided indicated that participants did not provide reasons that were directly contrary to the facts presented in the scenarios, instead challenged the validity of the facts presented (e.g., for *Incest*, "The condom could have broke, and no birth control pill is exactly 100% effective").

There was no association between number of times dumbfounded and score on either measure on the Meaning and Life questionnaire; Presence $r(101) = -.08, p = .436$, or Search $r(101) = .06, p = .532$, or the Centrality of Religiosity Scale $r(101) = .04, p = .662$. This is consistent with Studies 1, 2, and 3a. It appears that susceptibility to dumbfounding is not related to either measure.

3.6 Combined Results and Discussion

3.6.1 Evaluating each measure of dumbfounding. The studies reported in this chapter identify moral dumbfounding as a rare demonstration of a separation between intuitions and reasons for these intuitions (e.g., Barsalou, 2003, 2008, 2009; Crockett, 2013; Cushman, 2013). Two ways in which this separation may manifest were identified. Firstly participants may acknowledge that they do not have reasons for their judgements, admitting to not having reasons. Secondly, participants may fail to provide reasons when asked, providing responses that fail to answer the question they were asked. Two such responses were identified, unsupported declarations and tautological responses.

Measuring dumbfounding as an admission of not having reasons only provides a stricter measure of dumbfounding. Across Studies 1, 3a, and 3b (*N* = 204), 100 participants (49%) admitted to not having reasons for their judgements at least once. When a failure to provide reasons (taken as the providing of unsupported declarations in Study 1; and, unsupported declarations and tautological responses in Study 3) was included as a dumbfounded response, 136 participants (67%) were identified as dumbfounded at least once across Studies 1, 3a, and 3b. When the selecting of an unsupported declaration (Study 2, *N* = 72) was included (*N* = 276), 204 participants, (74%) were identified as dumbfounded at least once.

The disparity in results between Study 2 and the other studies suggests that the selection of an unsupported declaration does not provide a good measure of moral dumbfounding. Participants in Studies 1, 3a, and 3b, recognised the illegitimacy unsupported declarations as justifications for their judgement, with the majority of participants avoided resorting to this type of response at all. The vast majority of participants appeared to be willing to ignore the illegitimacy of the response, with large numbers of participants selecting the unsupported declaration. While Study 2 did not identify a means to measure dumbfounding, these results are interesting, and may provide an insight into the cognitive processes that lead to dumbfounding.

Providing an unsupported declaration is clearly different to selecting one from a list of possible responses. One possible explanation, is that dumbfounding is an aversive state, similar to experiencing a threat to meaning (Heine et al., 2006; Proulx & Inzlicht, 2012), or cognitive dissonance (Cooper, 2007; Festinger, 1957; Harmon-Jones & Harmon-Jones, 2007). The selecting of an unsupported declaration without deliberation allows participants to avoid or minimise the impact of this aversive state and move on. Providing an unsupported declaration in an interview involves more deliberation, making the illegitimacy of it more salient, reducing its effectiveness in avoiding the aversive state of dumbfoundedness. Furthermore, the relative attractiveness of these different responses to participants may be linked to social desirability (e.g., Chung & Monroe, 2003; Latif, 2000; Morris & McDonald, 2013). Follow-up work could investigate these questions directly.

The explicit acknowledgement of an absence of reasons can be measured systematically by the selection of an admission of having no reasons. This is an unambiguous measure of moral dumbfounding, does not account for participants who fail to provide reasons. Measuring a failure to provide reasons, however, is more problematic. What is termed as a valid reason is somewhat subjective. The providing of unsupported declarations and tautological responses has been identified here as an indicator of a failure to provide reasons. This is grounded in discussions of dumbfounding in the wider literature (Haidt, 2001; Mallon & Nichols, 2011; Prinz, 2005), and the theoretical framework adopted here. Evidence for equivalence of unsupported declarations and admissions of not having reasons was also found in Study 1 whereby both measures displayed similar variability in non-verbal behaviours when contrasted against participants who provided reasons, and participants who did not rate the behaviour as wrong. However, caution is advised in taking unsupported declarations as evidence for dumbfounding, particularly given the pattern of responses in Study 2, and that a number of participants in Study 1 who provided an unsupported declaration proceeded to provide reasons, or a revised judgement.

The two measures of dumbfounding were identified in this chapter. Limitations are associated with each. Relying on admissions of having no reasons only, provides an overly strict measure whereby a failure to provide reasons is not measured. Taking unsupported declarations (and tautological reasons) as a measure of dumbfounding may provide too broad a measure, risks identifying lazy or inattentive participants as dumbfounded. The providing of a type-written response as part of a computerised task requires effort, and the majority of participants avoid the use of unsupported declarations as justifications for their judgements. This suggests that those who provided unsupported declarations did so because they failed to identify alternative reason. It appears that the most practicable means to measure dumbfounding accurately requires each of the responses: providing/selecting admissions of not having reasons, and the providing of an unsupported declaration, to be accounted for. Participants providing either of these responses may be identified as dumbfounded.

3.6.2 Differences between scenarios. In Study 1 we found that rates of dumbfounded responding varied depending on the scenario presented. Study 2 recorded high rates of dumbfounded responses for all scenarios. In Studies 3a and

3b, we observed low rates of dumbfounded responding for all scenarios. In Study 1 and Study 3b, we observed varying rates of dumbfounded responses depending on scenario type. When Studies 3a and 3b are analysed together this variation is still observed, with significantly more dumbfounded responses recorded for the intuition scenarios (110 instances) than for the reasoning scenarios (79 instances), $\chi^2(1, N = 692) = 6.55$, $p = .010$. However, this combined analysis may be skewed in favour of Study 3b, due to the larger sample size, 101 participants; Study 3a had only 72 participants. Further research and continued replication is needed to confirm the reliability of this finding. When the open-ended responses coded as tautological were included in the analysis of Studies 3a and 3b, the rates of dumbfounding appeared to be closer to those observed in Study 1.

Table 3.13 and Figure 3.6 show the initial observed rates of dumbfound responding for each study. Table 3.13 and Figure 3.7 show the revised rates of observed dumbfound responding in each study once the open-ended coded responses from Studies 3a and 3b are included. Rates of dumbfounding reported by Haidt et al. (2000) are also included for comparison. Study 2 was a primarily a pilot study, and, as discussed, the observed rates of dumbfounding do not appear to be representative of the phenomenon being studied, as such Study 2 is not included in Figure 3.7.

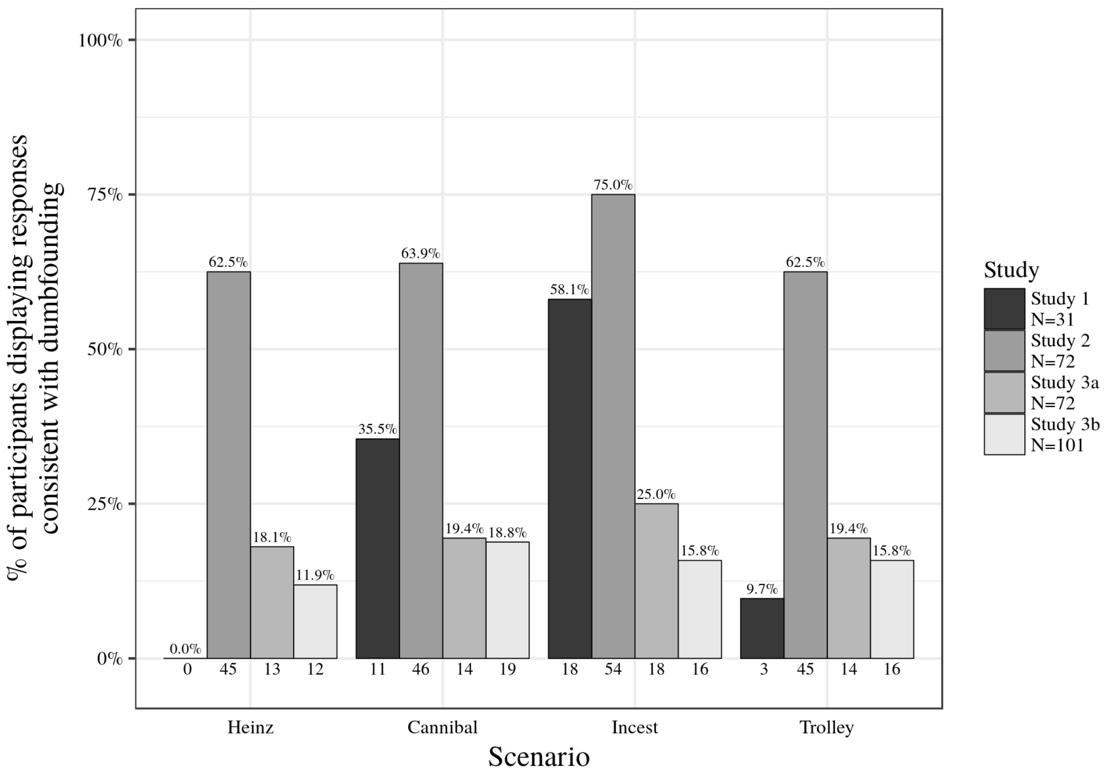

Figure 3.6: Percentage of participants providing/selecting responses consistent with dumbfounding across all four studies.

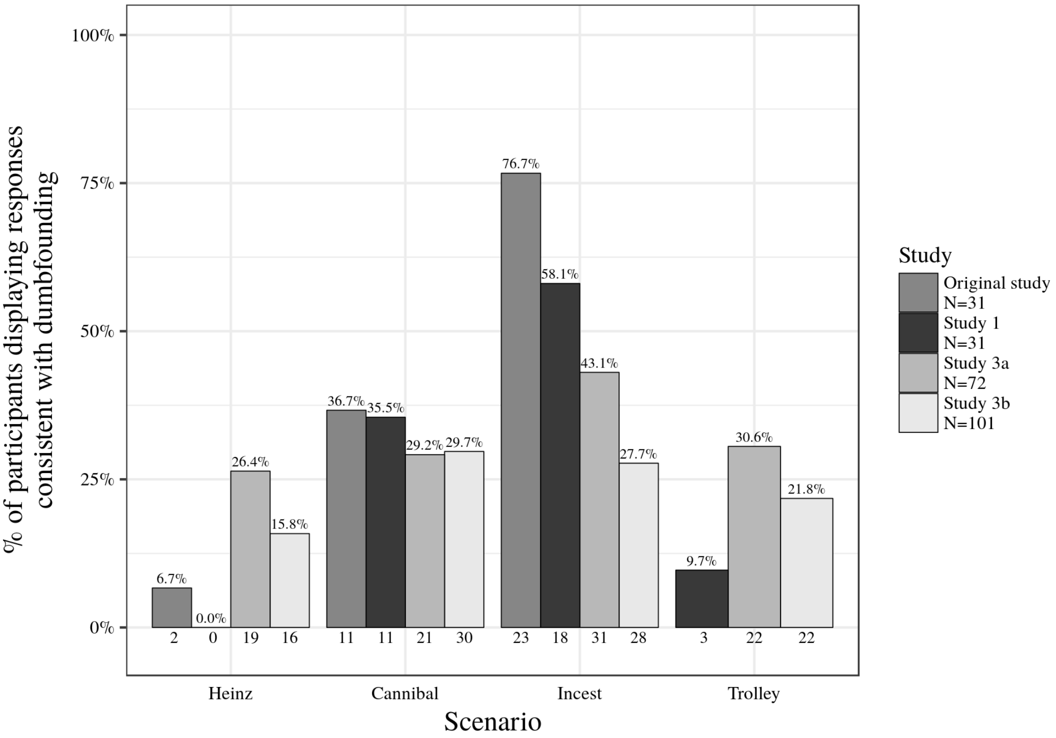

Figure 3.7: Rates of dumbfounded responding across 1, 3a, 3b, together with the original Haidt et al., study (2000) – including coded string responses

Table 3.13: Studies 1-3: Observed frequency and percentage of each of the responses: dumbfounded, nothing wrong, and reasons provided.

		Heinz		Cannibal		Incest		Trolley	
		N	%	N	%	N	%	N	%
Study 1	Nothing wrong	6	19.35%	8	25.81%	11	35.48%	8	25.8%
	Dumbfounded:	0	-	11	35.48%	18	58.06%	3	9.67%
	(Admissions)	0	-	8	25.81%	10	32.26%	3	9.67%
	(Unsupported declarations)	0	-	3	9.67%	8	25.81%	0	-
	reasons	25	80.65%	12	38.71%	2	6.45%	20	64.52%
Study 2	Nothing wrong	8	11.11%	4	5.56%	2	2.78%	10	13.89%
	dumbfounded	45	62.50%	46	63.89%	54	75.00%	46	63.89%
	reasons	19	26.39%	22	30.56%	16	22.22%	17	23.61%
Study 3a (just critical slide)	Nothing wrong	14	19.44%	4	5.56%	12	16.67%	15	20.83%
	dumbfounded	13	18.06%	14	19.44%	18	25.00%	14	19.44%
	reasons	45	62.50%	54	75.00%	42	58.33%	43	59.72%
Study 3a (coded responses)	Nothing wrong	14	19.44%	4	5.56%	12	16.67%	15	20.83%
	dumbfounded	19	26.39%	21	29.17%	31	43.06%	22	30.56%
	reasons	39	51.17%	47	65.28%	29	40.28%	35	48.61%
Study 3b (just critical slide)	Nothing wrong	21	20.79%	10	9.90%	31	30.69%	24	23.76%
	dumbfounded	12	11.88%	19	18.81%	16	15.84%	16	15.84%
	reasons	68	67.33%	72	71.29%	54	53.47	61	60.40%
Study 3b (coded responses)	Nothing wrong	21	20.79%	10	9.90%	31	30.69%	24	23.76%
	dumbfounded	16	15.84%	30	29.70%	28	27.72%	22	21.78%
	reasons	64	63.37%	61	60.40%	42	41.58%	55	54.46%

3.6.3 Differences between the samples. The trend in observed rates of dumbfounded responses, across the dilemmas, identified by Haidt et al. (2000) appears to also be present in Study 1 (Interview). There does not appear to be a difference between scenarios in the computerised tasks. When the open-ended responses are included, the rates of observed dumbfounding for *Cannibal* appear to be similar across all the studies included in Figure 3.7 (two interviews and two computerised tasks). The computerised tasks appear to have higher rates of dumbfounding for both *Heinz* and *Trolley* than the interviews. There is a large degree of variation in the observed rate of dumbfounding for *Incest* between the four studies.

Incest recorded higher rates of dumbfounding than the other scenarios in both interview studies (Study 1 and Haidt et al. (2000)) and, to some degree, in Study 3a, the computer task with a college sample. The rate of dumbfounding observed for *Incest* with the online sample, in Study 3b, is lower than that observed with the college sample in Study 3a and is also slightly lower than that observed for *Cannibal* in the online sample. This is surprising, in that, the *Incest* dilemma is the most commonly cited example (e.g., Haidt, 2001; Prinz, 2005; Royzman et al., 2015), and, in Studies 1, 2, and 3a, is the most reliable for eliciting dumbfounding, consistently eliciting higher rates than the other dilemmas. Looking at the ratings of the behaviours in each dilemma for each study may provide some clue as to where this variation comes from. The online sample were less inclined to rate the behaviour in *Incest* as wrong relative to the participants in the other studies. The percentage of participants initially rating *Incest* as wrong for each study are as follows: Study 1: 83.9%; Study 2: 87.5%; Study 3a: 84.7%; Study 3b: 70.3%. Furthermore, on the critical slide, the proportion of participants who selected “nothing wrong” for *Incest* for Study 3b (30.69%; 31 participants) was nearly double the proportion that selected “nothing wrong” for *Incest* for Study 3a (16.67%; 12 participants). When these participants are excluded from the analysis of Study 3b (see Table 3.14 and Figure 3.8), the percentage of participants appearing to be dumbfounded by *Incest* (22.86%; 16 participants; or 40%; 28 participants when open-ended responses are included; N = 70) exceeds the percentage of participants appearing to be dumbfounded by *Cannibal* (20.88%; 19 participants; or 33%; 30 participants when open-ended responses are included; N = 91). It appears that the apparent uncharacteristically low rates of observed dumbfounding for *Incest* in Study 3b, when compared to *Cannibal*, may be due to the online sample being less inclined to rate the behaviour as morally wrong rather than a difference in this sample's ability to provide justifications for their judgements to the two scenarios.

Table 3.14: Percentage of participants dumbfounded excluding "nothing wrong"

	Heinz		Cannibal		Incest		Trolley	
	N	%	N	%	N	%	N	%
Study 1: Interview (N=31)	0/25	0.00%	11/23	47.83 %	18/20	90.00 %	3/23	13.04 %
Study 2: Pilot Computer task (N=72)	45/64	70.31 %	46/68	67.65 %	54/70	77.14 %	46/62	74.19 %
Study 3a: Revised Computer task (N=72)	19/58	32.76 %	21/68	30.88 %	31/60	51.67 %	22/67	32.84 %
Study 3b: Revised computer task Online	16/20	20.00 %	30/91	32.96 %	28/70	40.00 %	22/77	28.57 %

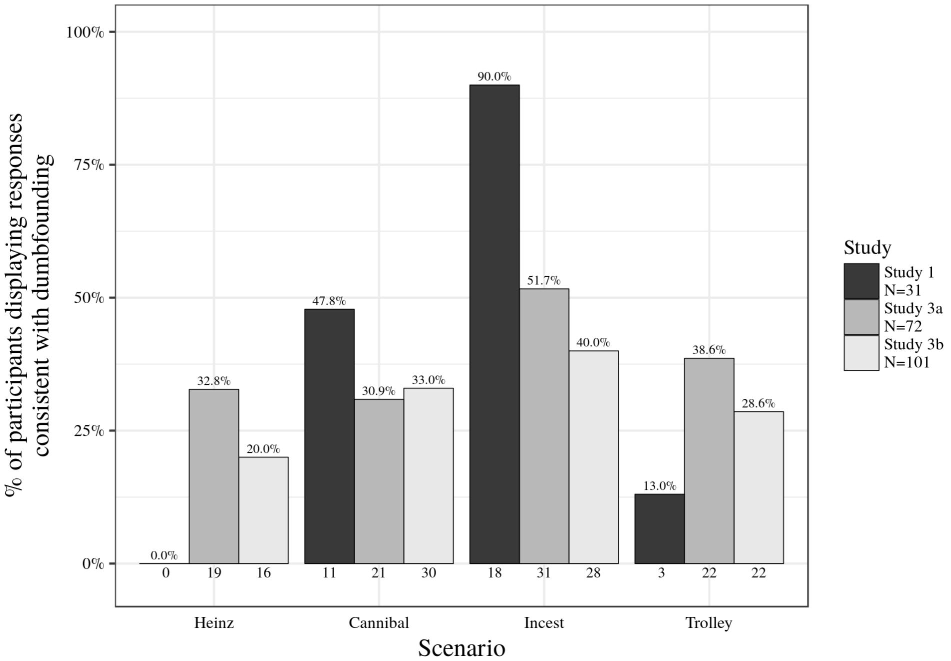

Figure 3.8: Percentage of dumbfounded responses when "nothing wrong" is excluded

It has been argued that moral dumbfounding occurs as a result of social pressure to conform to conversational norms (Royzman et al., 2015). The findings presented by Royzman et al. (2015) do not fully support this claim, however, they demonstrate that incidences of moral dumbfounding are sensitive to social pressure. Studies 2 and 3, aimed to reduce the influence of social pressure by testing dumbfounding as part of a computerised task, as opposed to in an interview setting. The varying rates of dumbfounding depending on task type indicate that the computerised task is different from the interview.

Evidence that social pressure is reduced in the computerised task can be found by examining the degree to which participants changed their minds, as measured in the self-report response, and by comparing the initial judgements and revised judgements. The self-report responses for Study 2 were of a binary yes/no form, whereas the responses in the other studies were provided on a 1-7 Likert scale. The self-report data from Study 2 is therefore not included in the analysis that follows.

The mean responses for the self-report question “How much did you change your mind?” are as follows: Study 1, $M = 2.88$, $SD = 1.59$; Study 3a, $M = 2.01$, $SD = 1.46$; Study 3b, $M = 1.69$, $SD = 1.27$. A one way ANOVA revealed significant differences in responses to this question between the different Studies $F(2, 809) = 33.811$, $p < .001$, partial $\eta^2 = .077$. Tukey's post-hoc pairwise comparison revealed that responses in Study 1 were significantly higher than both Study 3a, $p < .001$, and Study 3b, $p < .001$. The responses in Study 3a were also significantly higher than the responses in Study 3b, $p = .008$.

The initial judgements and revised judgements in the computer tasks were binned for comparison with the interview. “Wrong” judgements were assigned a value of “-1”, “Right” judgements were assigned a value of “+1”, “neutral” judgements were assigned a value of 0. The values for the revised judgements were subtracted from values for the initial judgements to create a new variable containing positive values ranging from -2 to +2. Negative values represent a change in judgement towards a more favourable judgement, and positive values represent a change in judgement towards condemning the actions. Higher values represent a greater swing in judgement. In the interview, there was only one incidence of a participant changing their judgement from favourable to condemnation, whereas 11 participants changed their judgement towards a more favourable judgement. In the computerised tasks, the numbers of participants changing their judgement in each direction is more balanced. There was a significant association between type of study and whether or not participants changed their mind in a given direction, $\chi^2(12, N = 1104) = 37.179, p < .001$. When Study 1 was removed this association disappeared, $\chi^2(8, N = 980) = 10.106, p = .258$. These pattern of results suggests that participants reacted differently in the interview than in the computerised tasks.

3.7 General Discussion

The goal of the studies conducted in this chapter was to address the Questions: 1. “Is moral dumbfounding a real phenomenon?”; 1.1 “How should moral dumbfounding be measured?”; and 1.2 “Is it possible to elicit moral dumbfounding in a laboratory based task?” The studies conducted were designed to examine the replicability of dumbfounded responding following a moral judgement task, and identify specific measurable responses that may be viewed as indicators of moral dumbfounding. Four studies, with a combined total sample of *N* = 276, were conducted in an attempt to replicate and extend the original demonstration (*N* = 30) of moral dumbfounding by Haidt et al. (2000). We predicted that dumbfounded responses would be evoked when participants were required to provide justification for their moral judgements, when their basic intuitive justifications had been refuted. Two measures of moral dumbfounding were taken, an explicit acknowledgement of the absence of reasons, and a failure to provide reasons when pushed. Rates of observed dumbfounding vary depending on which measure is being employed.

3.7.1 Intuition versus reasoning. Haidt et al. (2000) attribute the observed trend in dumbfounded responding to differences in type of scenario. They argue that *Heinz* is a "reasoning" scenario while *Cannibal* and *Incest* are "intuition" scenarios. Prinz (2005) suggests that these "intuition" scenarios have an emotional component, specifically that they elicit disgust, which leads to the judgement. Prinz argues that judgements grounded in disgust are more difficult to justify because they are grounded in emotion rather than reason. The variability between scenarios may be evidence for Haidt et al. prediction that judgements on the "intuition" scenarios would be more difficult to justify than the "reasoning" scenarios.

Study 1, the interview, was the only study to produce robust differences between the scenarios.[6] The results of the computerised tasks may indicate that there is no difference between the reasoning scenarios and the intuition scenarios. Alternatively, this may have highlighted a difference between an interview and a computerised task that influences the way people make moral judgements.

It is possible that there exists a social influence in an interview setting that changes the way participants respond (e.g., Asch, 1956; Sabini, 1995; Staub, 2013) and, that the interviewer may be seen as a person in authority, demanding justifications for judgements made (e.g., Milgram, 1974). This may motivate participants to identify reasons to justify their judgements, leading to the suppression of dumbfounded responses. On the other hand, it may also motivate participants to

[6] Some differences were observed in Study 3b, however these existed only when scenarios were grouped by type, this inter-scenario variation in rates of dumbfounding is not equivalent to that observed in Study 1.

heed the counter-arguments offered by the experimenter. This may lead to an interaction between scenario difficulty and social pressure to emerge, with the social pressure leading to fewer dumbfounded responses to the easier "reasoning" scenarios, but leading to more dumbfounded responses to the more difficult "intuition" (or bodily norm, Russell & Giner-Sorolla, 2013) scenarios. It may be the case that the rates of dumbfounding found in the computer tasks provide something of a crude baseline measure of participants' initial perception of their own ability to justify their judgement of the scenario, having read the scenario and a number of counter-arguments. In the interview, these initial responses to the scenarios are distilled by the discussion with the experimenter to reflect the variation in difficulty between the scenarios.

3.7.2 Implications. The existence of moral dumbfounding has informed various theories of moral judgement either directly (e.g., Haidt, 2001; Prinz, 2005; Cushman et al., 2010) or indirectly (e.g., Crockett, 2013; Cushman, 2013; Greene, 2008, 2013). The original demonstration of moral dumbfounding remains unpublished in peer reviewed form (Haidt et al., 2000) and has not been directly replicated before now. The studies presented here aimed to replicate and extend this original moral dumbfounding study (Haidt et al., 2000) and thus, assess the notion that moral dumbfounding is in fact a psychological phenomenon that can be consistently observed. Study 1 successfully replicated the original study. Study 2 piloted the use of a computer task and recorded unexpectedly high rates of dumbfounded responding. Possible reasons for this were identified and addressed in Studies 3a and 3b. Study 3a and 3b recorded more moderate rates of dumbfounding with two different samples. All three studies successfully elicited dumbfounded responding identified as (a) admissions of not having reasons; (b) use of unsupported declarations as justification of a judgement; or (c) use of undefended tautological response as justification for a judgement; however, differences remain between the interview in Study 1 and the computerised task in Studies 3a and 3b. Taking these responses to be indicators of a state of dumbfoundedness, it appears that moral dumbfounding can be evoked in face-to-face and online contexts. As such, the research presented here may be seen as more support for the existence of intuitionist theories of moral judgement (e.g., Cushman et al., 2010; Greene, 2008; Haidt, 2001; Haidt & Björklund, 2008; Hauser et al., 2008; Prinz, 2005) over rationalist theories (e.g., Kohlberg, 1971; Topolski et al., 2013).

3.7.3 Limitations and future directions. The studies reported in this chapter are exploratory in design. The aim was to identify whether or not the phenomenon of moral dumbfounding could be elicited in a robust fashion. There was no experimental manipulation and analyses were primarily descriptive. These studies raise significant questions about the mechanisms underlying dumbfounded responses to moral judgement tasks, but clearly indicate that such dumbfounded responses can be reliably elicited, and demonstrate interesting variability.

One hypothesised explanation of dumbfounded responding has been proposed by Royzman et al. (2015). According to this explanation dumbfounded responding does not provide evidence that people do not have reasons for their judgement. They suggest that people do have reasons for their judgements, however as part of the dumbfounding paradigm these reasons are dismissed by the experimenter as invalid. A participant may not accept that these reasons are invalid, however to object to their dismissal would risk presenting as stubborn. As such, in order to avoid appearing stubborn, participants may feign agreement with the experimenter regarding the validity of the reasons. In feigning agreement on the validity of the reasons, participants can no longer use these reasons as justification for their judgements. This forces participants to either revise their judgement or present as dumbfounded. The studies described in this chapter elicited dumbfounded responding, however the specific mechanisms that lead to this type of responding remain unknown. The possibility that dumbfounded responding emerges as a result of social pressure to accept the counter-arguments of the experimenter was not tested. This means that the studies described in this chapter do not necessarily provide evidence that people's judgements are intuitive (as opposed to rational, and grounded in reason). The possibility that participants providing dumbfounded responses may have reasons for their judgements is investigated in more detail in the next chapter.

3.8 Conclusion

The primary aim of the current studies was to examine the reliability of dumbfounded responding in moral judgements, and identify specific measurable indicators of moral dumbfounding. This is of particular interest considering the extent to which moral dumbfounding exists as a known phenomenon in the morality literature and its existence appears to inform theories of moral judgement. The studies presented in this chapter investigated two tenets of the research question “Is moral dumbfounding a real phenomenon?”, namely, “how should moral dumbfounding be measured?” and “Is it possible to elicit moral dumbfounding in a laboratory based task?”

Two indicators of dumbfounding were taken: an admission of not having reasons and a failure to provide reasons when requested (measured by the providing of unsupported declarations/tautological responses). Four studies revealed varying rates of moral dumbfounding as recorded by these indicators depending on the type of task and on which indicator is being used. While further work is necessary to identify the specific variables that may moderate this variability, the research presented here demonstrated that two types of dumbfounded responding can be reliably elicited. In other words, we found that people are not always able to justify their moral judgements; they maintain their judgements in the absence of supporting reasons, in some cases they resort to unsupported declarations as justifications for judgements, in others admit that they do not have reasons for their judgement. Further research is required to establish why this occurs.

The studies in this chapter did not directly address the questions raised by Royzman et al. (2015). However, dumbfounded responding was observed in the computerised tasks, along with some evidence that social pressure in the computerised task was reduced. This finding does not support the claim by Royzman et al. (2015) that dumbfounded responding can be attributed to social pressure to avoid appearing "uncooperative" (Royzman et al., 2015, p. 299), "inattentive" or "stubborn" (Royzman et al., 2015, p. 310). Their claim that the judgements of dumbfounded participants can be attributed to either norm-based reasons or reasons of potential harm will be tested in the next chapter.

4 Chapter 4 – Reasons or Rationalisations: Inconsistency in Articulating, Endorsing, and Applying Moral Principles

The studies described in the previous chapter demonstrated that dumbfounded responding could be elicited as part of an interview and a computer-based task. Two types of responses were taken as evidence of dumbfounding, admissions of not having reasons, and unsupported declarations. The degree to which these responses can be taken as evidence of a state of dumbfoundedness may be subject to challenge. It is not clear whether participants providing these responses truly accept that they have no reasons for their judgements, or if they believe they do have reasons and provide a dumbfounded response in order to be seen to be responsive to the counter-arguments (even if they disagree with them). In this second case, dumbfounded responding may be attributed to the social pressure inherent in the experimental paradigm as opposed to emerging as a result of the intuitive nature of the making of moral judgements. This explanation therefore undermines traditional explanations of moral dumbfounding, and, if true, would mean that using moral dumbfounding as evidence for intuitionist theories of moral judgement (e.g., Haidt, 2001; Prinz, 2005) would be problematic.

Explanations of moral dumbfounding that allow for judgements to be based on reasons have been proposed before (e.g., Gray et al., 2014; Jacobson, 2012; Royzman et al., 2015; Sneddon, 2007; Wielenberg, 2014). Some of the possible reasons that may underlie the judgements of dumbfounded participants have been suggested by various authors. Gray et al. (2014) suggest that when judging moral scenarios, people implicitly perceive harm even in scenarios that are construed as objectively harmless. If people perceive harm in the scenarios, then, even when the experimenter claims that they are harm free, this perception of harm still serves as a reason to condemn the behaviour. They conducted a series of experiments demonstrating that people do implicitly perceive harm in supposedly victim-less scenarios; e.g., “masturbating to a picture of one’s dead sister, watching animals have sex to become sexually aroused, having sex with a corpse, covering a Bible with feces” (Gray et al., 2014, p. 1063). Similarly Jacobson (2012) argues that, moral judgements are grounded in reasons and presents a number of plausible reasons why a person may condemn the actions of the characters in each of the dumbfounding scenarios. In the case of *Cannibal* he suggests that if Jennifer’s behaviour became known, people would be less willing to donate their bodies to the lab; in the case of *Incest* he suggests that the behaviour of Julie and Mark was risky, “reckless and licentious” (Jacobson, 2012, p. 25). He also suggests that when participants appear to be dumbfounded they have simply given up on the argument and conceded to the experimenter who is in a position of authority. Gray et al. (2014) used a different set of scenarios from the original dumbfounding study (and did not test for dumbfounding). Jacobson’s (2012) paper is speculative, he does not provide any evidence to suggest that dumbfounded participants judgements are grounded in the reasons he identified. However, the work of Royzman et al. (2015) appears to empirically test Jacobson’s claims.

4.1 Evidence for Principles Guiding Judgements

A recent series of studies by Royzman et al. (2015) investigating the classic *Incest* scenario from the original dumbfounding study (Haidt et al., 2000) aimed to identify if participants presenting as dumbfounded genuinely had no reasons to support their judgements. In line with Jacobson (2012), they claim that dumbfounding occurs as a result of social pressure to adhere to conversational norms, arguing that dumbfounded participants do have reasons for their judgements and that these reasons are incorrectly dismissed as invalid by the experimenter. They argue that dumbfounded responding occurs as a result of social pressure to avoid appearing “uncooperative” (Royzman et al., 2015, p. 299), “inattentive” or “stubborn” (Royzman et al., 2015, p. 310). However, recall that the original definition of dumbfounding, which Royzman et al., employ, refers to the “stubborn” maintenance of a judgement. This creates a paradoxical situation whereby presenting as stubborn (as part of a dumbfounded response) occurs as a result of an attempt to avoid appearing stubborn.

Royzman et al. (2015) identify two principles that may be guiding participants' judgements: the harm principle, and the norm principle. They claim that if participants endorse either of these principles, they do have legitimate reasons for their judgements. They argue that by excluding from analysis participants who endorse either of these principles, incidences of dumbfounding are negligible.

In identifying the harm principle, Royzman et al. (2015) draw on the work of Gray et al. (2014). They hypothesised that participants may not believe the scenario to be harm free even in the face of repeated assurances from the experimenter that it is harm free. If a participant does not believe that an act is truly harm free then this provides them with a perfectly valid reason to judge it as morally wrong (Gray et al., 2014; Royzman et al., 2015). They devised two questions which served as a “credulity check” (Royzman et al., 2015, p. 309), to assess whether or not participants believed that the *Incest* scenario was harm-free. The questions read as follows: (i) “Having read the story and considering the arguments presented, are you able to believe that Julie and Mark’s having sex with each other will not negatively affect the quality of their relationship or how they feel about each other later on?”; (ii) “Having read the story and considering the arguments presented, are you able to believe that Julie and Mark's having sex with each other will have no bad consequences for them personally and/or for those close to them?" (Royzman et al., 2015, pp. 302–303). If participants responded "No" to either of these questions, their judgements were attributed to harm-based reasons, and therefore could not be identified as dumbfounded.

The second principle identified by Royzman et al. (2015) is the norm principle. They argue that if people believe that committing a particular act is wrong, regardless of the circumstances, then, for these people, this belief may be sufficient to serve as a reason to condemn the behaviour of the characters in the scenario. Royzman et al. (2015) presented participants with two statements: (a) "violating an established moral norm just for fun or personal enjoyment is wrong only in situations where someone is harmed as a result, but is acceptable otherwise"; (b) "violating an established moral norm just for fun or personal enjoyment is inherently wrong even in situations where no one is harmed as a result" (Royzman et al., 2015, p. 305). If participants endorsed (b) over (a) they reasoned that a judgement could be legitimately defended using a normative statement. They suggest that the "unsupported declarations" (Haidt et al., 2000, p. 12) identified by Haidt et al. (2000) are statements of a normative position, and that, rather than being a viewed as a dumbfounded response, they are a legitimate reason for judgements.

According to Royzman et al. (2015), participants whose judgements could not be attributed to either the harm principle (based on responses to the credulity check) or the norm principle (based on endorsement) were classified as "fully convergent" (Royzman et al., 2015, p. 306). Participants' eligibility for analysis was based on these convergent criteria, such that only participants identified as "fully convergent" were deemed eligible for analysis. It should be noted that these criteria for convergence were grounded in the researchers' own reasoning regarding the premises of the *Incest* vignette. This means that eligibility for analysis is based on level convergence with the reasoning of the researchers. Using these stricter criteria for dumbfounding, Royzman et al. (2015) initially identified 4 participants, from a sample of 53, who presented as dumbfounded. Each of these participants was then interviewed and the inconsistencies in their responses pointed out to them. During these interviews 2 participants changed their judgement of the behaviour and 1 participant changed her position on the normative statements. This left just 1 fully convergent, dumbfounded participant. This participant did not resolve the inconsistency in his responses to the questions, and, following post-experiment interviews, Royzman and colleagues found dumbfounding to occur once in a sample of 53. This was found to be not significantly greater than 0 and therefore Royzman et al. (2015) concluded that moral dumbfounding does not exist.

4.2 Limitations of the Rationalist Explanation

There are three main issues with the Royzman et al. (2015) arrive at their conclusion that moral dumbfounding does not exist. Firstly, the initial estimate of incidences of dumbfounding was 4/53 (7.56%). Based on the same calculations used by Royzman et al. (2015), this estimate of 4/53 is significantly greater than 0 (0/53, $z = 2.04$, $p = .041$). These four participants were then interviewed further, during which, the "inconsistencies" in participants' "responses were pointed out directly" (Royzman et al., 2015, p. 308). Following this interview, Royzman et al. were left with a dumbfounding estimate of 1/53 (which they claim is not significantly greater than 0/53).

It is surprising that, having made the claim that dumbfounding arises as a result of social pressure, providing convincing evidence for this claim required a follow up interview, in which participants are exposed to social pressure. Using the same logic employed by Royzman et al. it would not be surprising if participants revised their responses after being “advised to carefully review and, if appropriate, revise” their responses (Royzman et al., 2015, p. 308). From this, it appears that incidences of dumbfounding can be reduced by changing the demands of the social situation. In effect, Royzman et al. (2015) have shown that moral dumbfounding is sensitive to social pressure. Demanding consistency between judgement and the endorsing of principles that may be relevant for a judgement reduces incidences of dumbfounding, whereas demanding consistency between a judgement and information contained in the vignette leads to increased dumbfounding. This is not the same as their claim that moral dumbfounding is caused by social pressure. Furthermore, the role of social pressure in the reduced incidences of dumbfounding observed is not acknowledged.

Secondly, following this interview, Royzman et al. (2015) are still left with one participant who, by their own criteria, can be identified as dumbfounded (Royzman et al., 2015, p. 308). No explanation for the responding of this participant is offered, and cannot be explained by the theoretical position adopted in the conclusion. It is argued that one participant from a sample of 53 is not significantly greater than 0/53, $z = 1.00$, $p = .32$. Disregarding this estimate of moral dumbfounding as not statistically significant, $p = .32$, avoids offering an explanation for a response that is inconsistent with the argument made in the paper.

Thirdly, and most importantly, dumbfounding has been identified as a rare demonstration of the separation between intuitions and reasons for these intuitions. Practical challenges to demonstrating this separation have already been identified (a) post-hoc rationalisation and identification of reasons that are consistent with a judgement; (b) the possibility that the intuition emerged as a result of a well-rehearsed reasoned response). The work presented by Royzman et al. (2015) may be viewed as a practical demonstration of this first challenge; helping participants identify reasons that are consistent with their judgement and providing them with an opportunity to endorse these reasons. The endorsing of a reason does not imply that the reason contributed to the judgement. This view of moral dumbfounding presents two methodological considerations that need to be addressed before accepting the claim that judgements in the dumbfounding paradigm can be attributed to either norm-based reasons or harm-based reasons. The first relates to participants' ability to articulate either harm-based or norm-based reasons. The second relates to the consistency with which these reasons guide judgements.

Beginning with the first methodological consideration, the final study reported by Royzman et al. (2015) does not report whether or not participants who endorsed either norm-based reasons or harm-based reasons also articulated the same reason. The mere endorsing of a principle or reason does not provide evidence that this principle guided the making of a judgement. To illustrate this point, consider the following scenario:

> Two friends (John and Pat) are bored one afternoon and trying to think of something to do. John suggests they go for a swim. Pat declines stating that it's too much effort – to get changed, and then to get dried and then washed and dried again after; he says he'd rather do something that requires less effort. John agrees and adds "Oh yeah, and there's that surfing competition on today so the place will be mobbed". To which Pat replies "Yeah exactly!"

When John mentioned the surfing competition Pat immediately adopted it as another reason not to go for a swim however it is clear that this reason played no part in Pat's original judgement. It is possible that in identifying other reasons that are consistent with a particular judgement researchers may falsely attribute the judgement made to these reasons.

The studies described by Royzman et al (2015) do not sufficiently guard against the possibility of falsely attributing judgements to reasons endorsed, allowing for the possibility that some participants were falsely excluded from analysis. Two ways to reduce the possibility of false exclusion are identified and examined in Studies 4 and 5 below. Firstly, as illustrated by the Pat and John example, endorsing a reason post-hoc does not provide any evidence that this reason was guiding a judgement. The articulation of a reason independent of a prompt provides much stronger evidence that that reason was guiding the judgement. For this reason the inclusion of an open-ended string response option immediately after the presenting of the vignette, in which participants are invited to provide the reason(s) for their judgement should reduce the possibility of false exclusion. Participants are then only excluded from analysis if they both articulated and endorsed a given principle.

The second methodological consideration relates specifically to the harm-based reasons, or the application of the harm principle. Royzman et al. (2015) argue that if participants do not believe that harm did not come as a result of the actions of Julie and Mark then concerns of harm may be considered a legitimate reason for judging the behaviour as wrong. Essentially, they have identified the harm principle as "it is wrong for two people to engage in an activity whereby harm may occur". Royzman et al. (2015) argue that the application of this harm principle provides participants with a reason for their judgements. If the harm principle is guiding the judgements of participants, then it should be applied consistently across differing contexts. Royzman et al. do not demonstrate that the participants in their sample consistently apply this principle across differing contexts (e.g., contact sports/boxing). Assessing the consistency with which the harm principle is applied across differing contexts will inform the degree to which this principle guides some participants' judgements, and reduce the false exclusion of participants based on the incorrect attribution of their judgements to the harm principle.

4.3 Reasons or Rationalisations

The aim of the studies in the current chapter is to test the claim by Royzman et al. (2015) that participants' judgements in the *Incest* scenario can be attributed to the harm principle or the norm principle. These studies address the final tenet of the first research question (Is moral dumbfounding a real phenomenon?), that is, "Is dumbfounded responding truly indicative of a state of dumbfoundedness or can it be attributed to features of experimental design?". Furthermore, one aspect of the second research question "How can the existence (or absence) of dumbfounding inform theories of moral judgement?" is also addressed to some degree in this chapter, namely, can moral dumbfounding be explained by rationalism? Two studies are reported in this chapter. Study 4 introduced an open-ended response option to assess whether or not participants articulated either principle. Study 5 included three targeted questions, to assess the consistency with which participants apply the harm principle. A principle is only deemed to be guiding a judgement if it is (a) consistently endorsed and articulated (Study 4), or (b) consistently endorsed and articulated, and, in the case of the harm principle, applied across differing contexts (Study 5).

4.4 Study 4: Articulating and Endorsing

Study 4 was an extension of Royzman et al. (2015), using largely the same materials. One moral judgement vignette (*Incest*) was taken from Haidt et al. (2000)

(Appendix A). Targeted questions, designed to assess participants' endorsements of the harm principle or the norm principle, were taken directly from Royzman et al. (2015). In addition to this, an open-ended response option was included immediately after the presenting of the vignette to assess whether or not participants could articulate these principles. The inclusion or exclusion of participants from analysis, in studies of dumbfounding, should account for both a participant's endorsing and articulation of a given principle. It was hypothesised that participants' endorsing of a principle would not be predictive of their ability to articulate this principle, and that by accounting for this, dumbfounded responding will be observed.

4.4.1 Method.

4.4.1.1 Participants and design. Study 4 was a frequency-based extension of Royzman et al. (2015). The aim was to identify if dumbfounded responding could be evoked, when both the endorsement and articulation of harm-based and norm-based principles was controlled for.

A combined sample of 110 (60 female, 49 male, 1 other; M_{age} = 32.44, min = 18, max = 69, *SD* = 11.28) took part. Fifty eight (25 female, 32 male, 1 other; M_{age} = 38.47, min = 19, max = 69, *SD* = 12.34) were recruited through MTurk.[7] Participation was voluntary and participants were paid 0.50 US dollars for their participation. Participants were recruited from English speaking countries or from countries where residents generally have a high level of English (e.g., The Netherlands, Denmark, Sweden).

[7] In order to prevent repeat participation from MTurk workers, this study and all remaining studies conducted on MTurk, were included as part of the same MTurk project as Study 3b. In addition, a probe question was included to check if participants had encountered the scenario before.

Fifty two (35 female, 17 male; $M_{\text{age}} = 25.71$, min = 18, max = 38, $SD = 3.80$) were recruited through direct electronic correspondence. Participants in this sample were undergraduate students, postgraduate students, and alumni from Mary Immaculate College (MIC), and University of Limerick (UL). Participation was voluntary and participants were not reimbursed for their participation.

4.4.1.2 Procedure and materials. Data were collected using an online questionnaire generated using Questback (Unipark, 2013). The questionnaire opened with the information sheet and consent form. The main questionnaire was only accessible once consent had been provided. Following the consent form, participants were presented with questions relating to basic demographics. Participants were then presented with two statements to assess if participants' judgements may be grounded in the norm principle. These were taken directly from Royzman et al. (2015): (i) "violating an established moral norm just for fun or personal enjoyment is wrong only in situations where someone is harmed as a result, but is acceptable otherwise."; (ii) "violating an established moral norm just for fun or personal enjoyment is inherently wrong even in situations where no one is harmed as a result.". Participants read both statements and were asked to select the statement they "identify with the most". The order of these statements was randomised. Participants who selected (ii) were then asked to elaborate on their position through an open-ended response question.

Participants were then presented with the "Julie and Mark" *Incest* vignette (Appendix A) from the original moral dumbfounding study (Haidt et al., 2000). They were asked to rate on a 7-point Likert scale how right or wrong they would rate the behaviour of Julie and Mark (where, 1 = *Morally wrong*; 4 = *neutral*; 7 = *Morally right*). They were asked to provide a reason for their judgement through open-ended response, and, rated their confidence in their judgement. Participants were then presented with a series of prepared counter-arguments designed to refute commonly used justifications for rating the behaviour as "wrong" (Appendix B).

Dumbfounding was measured using a "critical slide". The critical slide contained a statement defending the behaviour and a question as to how the behaviour could be wrong ("Julie and Mark's behaviour did not harm anyone, how can there be anything wrong with what they did?"). There were three possible answer options: (a) "There is nothing wrong"; (b) an admission of not having reasons ("It's wrong but I can't think of a reason"); and finally a judgement with accompanying justification (c) "It's wrong and I can provide a valid reason". The order of these response options was randomised. Participants who selected (c) were then prompted on a following slide to type a reason. The selecting of option (b), the admission of not having reasons, was taken to be a dumbfounded response, as was the use of an unsupported declaration as a justification for option (c).

Following the critical slide, participants rated the behaviour, and rated their confidence in their judgement again. They also indicated, on a 7-point Likert scale, how much they changed their mind. A post-discussion questionnaire containing self-report reaction to the scenario across various dimensions (confidence, confusion, irritation, etc.), taken from Haidt et al. (2000) was administered after these revised judgements had been made (Appendix C).

Two targeted questions were taken directly from Royzman et al. (2015) to assess whether or not participants' judgements may be grounded in the harm principle: (i) "Having read the story and considering the arguments presented, are you able to believe that Julie and Mark's having sex with each other will not negatively affect the quality of their relationship or how they feel about each other later on?"; (ii) "Having read the story and considering the arguments presented, are you able to believe that Julie and Mark's having sex with each other will have no bad consequences for them personally and/or for those close to them?". Participants responded "Yes" or "No" to each of these statements. The order of these questions was randomised.

Two other measures were also taken for exploratory purposes: Meaning in Life questionnaire (MLQ; Steger et al., 2008). This ten item scale is made up of two five item sub scales: presence (e.g., "I understand my life's meaning.") and search (e.g., "I am looking for something that makes my life feel meaningful."). Responses were recorded using a 7-point Likert scale ranging from 1 = *strongly disagree* to 7 *strongly agree*; and CRSi7 a seven item scale taken from The Centrality of Religiosity Scale (Huber & Huber, 2012). Participants responded to questions relating to the frequency with which they engage in religious or spiritual activity (e.g., "How often do you think about religious issues?"). Responses were recorded using a 5-point Likert scale ranging from 1= *never* to 5 = *very often*. The seven item inter-religious version of the scale was selected because some non-religious activities (such as meditation) may also have a bearing on a person's ability to reason about moral issues. All statistical analysis was conducted using R (3.4.0, R Core Team, 2017b).[8]

[8]

R (3.4.0, R Core Team, 2017b) and the R-packages *afex* (0.15.2, Singmann et al., 2015), *car* (2.1.4, Fox & Weisberg, 2011), *citr* (0.2.0, Aust, 2016), *desnum* (0.1.1, McHugh, 2017), *devtools* (1.13.1, Wickham & Chang, 2017), *dplyr* (Wickham et al., 2017), *estimability* (1.2, R. Lenth, 2016), *extrafont* (0.17, Winston Chang, 2014), *foreign* (0.8.68, R Core Team, 2017a), *ggplot2* (2.2.1, Wickham, 2009), *lme4* (1.1.13, Bates et al., 2015), *lsmeans* (2.26.3, R. V. Lenth, 2016), *lsr* (R. V. Lenth, 2016), *Matrix* (1.2.10, Bates & Maechler, 2017), metap (Dewey, 2017), *papaja* (0.1.0.9492, Aust & Barth, 2017), *plyr* (1.8.4, Wickham, 2011), *pwr* (Champely, 2018), *reshape2* (1.4.2, Wickham, 2007), *scales* (0.4.1, Wickham, 2016), *sjstats* (Lüdecke, 2018), and *wordcountaddin* (0.2.0, Marwick, n.d.).

4.4.2 Results and discussion. Eighty seven of the total sample (N = 110; 79.09%) rated the behaviour of Julie and Mark as wrong initially. The mean initial rating (immediately following presentation of the scenario) of the behaviour for the entire sample was, M = 2.04, SD = 1.45. An independent samples t-test revealed no difference in initial rating between the MTurk sample (M = 1.98, SD = 1.52), and the MIC sample, (M = 2.10, SD = 1.39), t(107.94) = -0.41, p = .683. Eighty six of the total sample, (N = 110; 78.18%) rated the behaviour as wrong after viewing the counter-arguments and the critical slide. The mean revised rating (immediately following the critical slide) of the behaviour for the entire sample was, M = 2.15, SD = 1.54. An independent samples t-test revealed no difference in revised rating between the MTurk sample, (M = 2.00, SD = 1.53), and the MIC sample, (M = 2.33, SD = 1.54), t(106.55) = -1.11 , p = .268. A paired samples t-test revealed a significant difference in rating of behaviour from time one, initial rating, (M = 2.04, SD = 1.45), to time two, revised rating, (M = 2.15, SD = 1.54), t(109) = -2.38, p = . 019; d = 0.08. This result may be due to changes in the severity of the judgements as opposed to changing the judgement. Further analysis revealed that only eight participants changed their judgement: two participants changed their judgement from “wrong” to “neutral”; one participant changed their judgement from “right” to “neutral”; four changed their judgement from “neutral” to “right”; and one participant changed their judgement from “neutral” to “wrong”. A chi-squared test for independence revealed no significant association between time of judgement and valence of judgement made, χ^2(2, N = 220) = 0.73, p = .694, V = .02. Ten participants (9%) indicated that they had encountered the scenario before. When asked to elaborate, participants provided anecdotes, or referred to previous readings (either fiction or philosophy), 2 participants (2%) indicated that they had encountered it in a previous survey.

4.4.2.1 Baseline measure of dumbfounding. Participants who selected the admission of not having reasons on the critical slide were identified as dumbfounded. Twenty participants (18.18%) were initially identified as dumbfounded by their selecting of the dumbfounded response ("It's wrong but I can't think of a reason") on the critical slide. Twenty two participants (20%) selected "There is nothing wrong"; 68 (61.82%) participants selected "It's wrong and I can provide a valid reason". Participants who selected "It's wrong and I can provide a valid reason" were required to provide a reason. The reasons provided were analysed and coded for dumbfounded responses, defined as (a) unsupported declarations or (b) tautological reasons. A total of 6 participants were identified as dumbfounded following this coding. Three participants provided an unsupported declaration, two provided a tautological reasons, and one participant provided both an unsupported declaration and a tautological reason as justification for their judgements. This brought the total number of participants providing dumbfounded responding to 26 (23.64%).

4.4.2.2 Endorsing harm and norm principles. The exclusion criteria developed by Royzman et al. (2015) were applied, and all participants who endorsed either the harm principle or the norm principle were excluded from analysis. This left a sample of fourteen who were "fully convergent" (Royzman et al., 2015, p. 308), and therefore eligible for analysis. None of these fourteen selected the dumbfounded response, ten selected "there is nothing wrong", and four selected "It's wrong and I can provide a valid reason". The responses to the critical slide are consistent with the initial and revised judgements made by these participants, with only four rating the behaviour as wrong on each occasion. The reasons provided were coded and no participants provided any other form of dumbfounded responses. This strict measure of convergence identified only fourteen participants, out of a sample of 110, who were eligible for analysis. None of these fourteen participants provided a dumbfounded response. These results are in line with Royzman et al. (2015), by excluding participants who endorse either the harm principle or the norm principle, dumbfounding can be eliminated.

4.4.2.3 Articulating harm and norm principles. The purpose of the Study 4 was to assess if participants can articulate the principles identified by Royzman et al. (2015), independently of the targeted statements/questions, as these may serve as a prompt. A revised measure of convergence is developed here. A participant's endorsement of either principle should lead to their exclusion from analysis, only if the participant also articulated this principle when given the opportunity. The open-ended responses were analysed and coded for any mention of either the harm principle or the norm principle. Participants were only excluded from analysis if they both endorsed and articulated either principle. For the purposes of consistency with Royzman et al. (2015), unsupported declarations and tautological responses, previously identified as dumbfounded responses were coded as an articulation of the norm principle. As predicted, the number of participants who both articulated and endorsed either principle was much lower than the number of participants who only endorsed either principle. Fifty two participants were eligible for analysis according to the revised exclusion criteria. Figure 4.1 shows the responses to the critical slide for the entire sample and for participants eligible for analysis according to each measure of convergence.

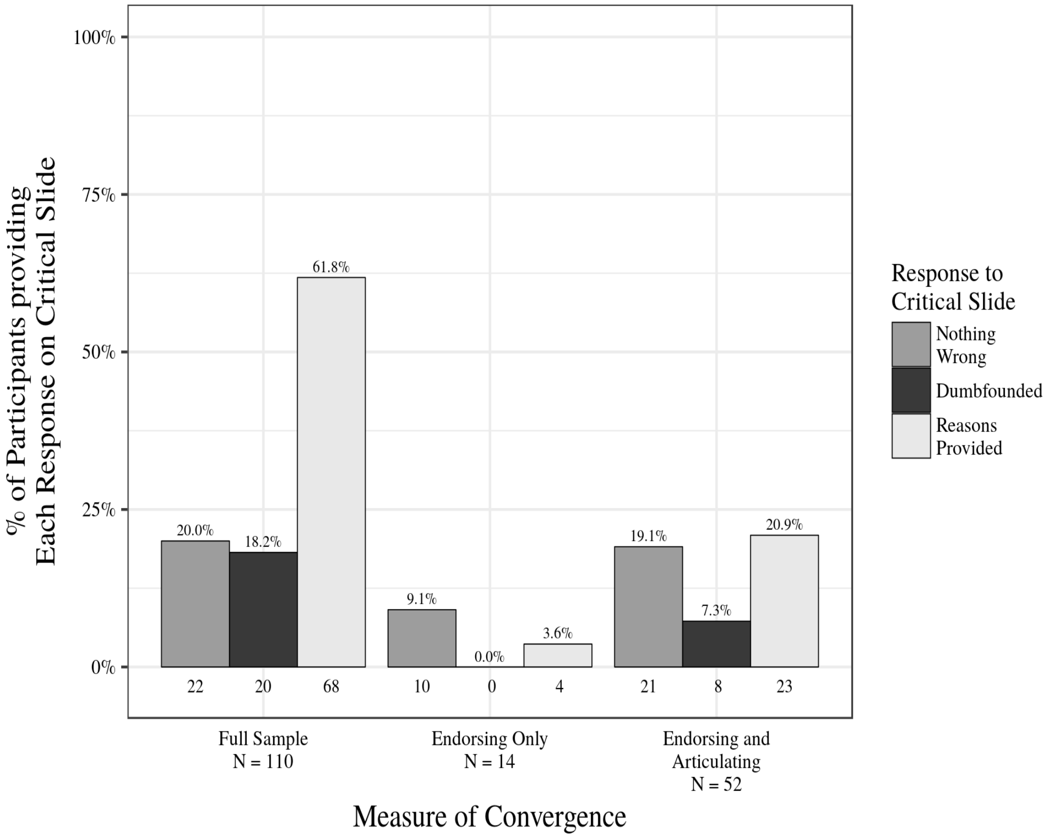

Figure 4.1: Study 4 – Responses to critical slide for the Entire Sample, and for each measure of convergence: (i) Endorsing only, and (ii), Endorsing and Articulating (all percentages are expressed in relation to the entire sample).

4.4.2.4 Revised rates of dumbfounding. Responses to the critical slide for the 52 participants who were eligible for analysis were as follows. Eight of these participants (15.38%) selected the dumbfounded response, the admission of not having reasons. Twenty one participants (40.38%) selected “There is nothing wrong” and Twenty three participants (44.23%) selected “It's wrong and I can provide a valid reason”. The reasons provided were analysed and coded. Four of these reasons were identified as possible dumbfounded responses (two unsupported declarations, e.g., “They are siblings and this behavior is not right”; and two tautological reasons, e.g., “They are brother and sister.”). However, as outlined above, and following the claim by Royzman et al. (2015), that referencing norm principles is sufficient as a justification, these participants were not classified as dumbfounded. According to the revised criteria for exclusion, and a stricter measure of dumbfounded responding (admissions of having no reason only), a total of eight participants from a sample of 110 (7.27%) were be identified as dumbfounded.

4.4.2.5 Judgement-principle consistency. The measure of convergence developed by Royzman et al. (2015) (endorsing only), led to a large proportion of participants who selected "There is nothing wrong" to be excluded from analysis (12 participants; 54.55% of the 22 participants who selected this option). Both the harm principle and the norm principle provide legitimate reasons for participants to judge the behaviour as wrong (Royzman et al., 2015). It follows that if a participant endorses either principle, they would also judge the behaviour as wrong. It is surprising that, 12 of the 22 participants who selected "There is nothing wrong" on the critical slide, also endorsed either the harm principle or the norm principle. The endorsing of these principles meant that these participants were excluded from analysis on the grounds they had a legitimate reason to rate the behaviour as wrong. However, these participants did not rate the behaviour as wrong. This demonstrates an inconsistency between the endorsing of the principles through targeted questions and statements and the apparent use of these principles as reasons guiding the participants' judgements. The endorsing only measure of convergence, using the targeted questions and statements developed by Royzman et al. (2015) led to participants being falsely excluded from analysis.

According to the revised criteria for exclusion, in which participants are only excluded from analysis if they were also able to articulate the principle that they endorsed, only one of the participants who was excluded from analysis selected "There is nothing wrong". In contrast, the original criteria for exclusion, endorsing only, developed by Royzman et al. (2015) falsely excluded 12 participants from analysis. The revised measure of convergence developed in Study 4 shows a dramatically reduced incidence of false exclusion of participants who selected "There is nothing wrong". This suggests that accounting for both the articulating and the endorsing of principles provides more accurate (though still not quite perfect) exclusion criteria.

The aim of Study 4 was to address limitations identified in Royzman et al. (2015). They excluded participants from analysis based on their endorsing of either the harm principle or the norm principle through targeted questions/statements. Using these criteria for exclusion, they found minimal dumbfounded responding (1 participant from a sample of 53, Royzman et al., 2015, p. 309). It was hypothesised that their exclusion criteria were too broad, and that participants' endorsing of either principle does imply that participants can articulate the given principle. Revised criteria for exclusion were developed which accounted for both the endorsing and the articulation of either the harm principle or the norm principle. Our initial analysis replicated the findings of Royzman et al. (2015). Further analysis, using the revised measure of convergence demonstrated considerably more consistency in the exclusion/inclusion of participants who selected "There is nothing wrong". These revised criteria identified eight participants as dumbfounded. Study 4 identified inconsistency in the endorsing and articulation of the harm principle and the norm principle, a second study was devised to assess the consistency in the application of the harm principle across differing contexts, along with the endorsing, and articulation of each principle.

4.5 Study 5: Applying Moral Principles Across Contexts

In Study 4 we tested if participants could articulate the harm principle and the norm principle as identified by Royzman et al. (2015). In Study 5 we investigated the role of the harm principle in the making of judgements. Specifically, we examined if the harm principle can legitimately be said to be guiding the judgements of participants. This was done by assessing whether or not the harm principle is applied consistently across different contexts

Drawing on Royzman et al. (2015), the harm principle may summarised as follows "it is wrong for two people to engage in an activity whereby harm may occur". According to the argument proposed by Royzman et al. (2015), participants' moral judgements are grounded in this principle, such that applying this principle to the *Incest* dilemma gives people a good reason to judge the behaviour of Julie and Mark as wrong. If this principle is to be considered as guiding participants' judgements, it should be consistently applied across differing contexts. Study 5 tested if this was the case by including a set of targeted questions relating to the generalisation and application of the harm principle across different contexts (the rest of the materials were largely the same as those used in Study 4). We hypothesised that participants' responses to these targeted questions would reveal inconsistency in the application of the harm principle across differing contexts. Any exclusion criteria based on the harm principle should account for the endorsing of the principle (Royzman et al., 2015), articulating the principle (Study 4), and the application of the principle (Study 5). A truer measure of moral dumbfounding should reflect these more detailed exclusion criteria.

4.5.1 Method.

4.5.1.1 Participants and design. Study 5 was a frequency-based extension of Study 4. The aim was to identify if participants were consistent in endorsing and applying the harm principle. A combined sample of 111 (67 female, 44 male; M_{age} = 34.23, min = 19, max = 74, *SD* = 11.42) took part.

Sixty one (36 female, 25 male; M_{age} = 39.08, min = 20, max = 74, *SD* = 12.25) were recruited through MTurk. Participation was voluntary and participants were paid 0.50 US dollars for their participation. Participants were recruited from English speaking countries or from countries where residents generally have a high level of English (e.g., The Netherlands, Denmark, Sweden).

Fifty (31 female, 19 male; M_{age} = 28.32, min = 19, max = 48, *SD* = 6.65) were recruited through direct electronic correspondence. Participants in this sample were undergraduate students, postgraduate students, and alumni from Mary Immaculate College (MIC), and University of Limerick (UL). Participation was voluntary and participants were not reimbursed for their participation.

4.5.1.2 Procedure and materials. Data were collected using an online questionnaire generated using Questback (Unipark, 2013). The questionnaire in Study 5 was the same as that presented in Study 4, with the inclusion of three additional targeted questions which aimed to assess the consistency with which participants generalise and apply the harm principle. The questions were: (a) "How would you rate the behaviour of two people who engage in an activity that could potentially result in harmful consequences for either of them?"; (b) "Do you think boxing is wrong?"; (c) "Do you think playing contact team sports (e.g. rugby; ice-hockey; American football) is wrong?". Responses to (a) were recorded on a 7-point Likert scale (where, 1 = *Morally wrong*; 4 = *neutral*; 7 = *Morally right*). Responses to (b) and (c) were recorded using a binary “Yes/No” option. These questions were presented sequentially, in randomised order. The randomised sequence was grouped as Block A. Similarly all slides and questions directly relating the moral scenario were grouped as Block B. Block B also included the targeted questions relating to the endorsing of the harm principle. The order of presentation of these blocks was randomised.

As with Study 4, the questionnaire opened with the information sheet, and the main body of the questionnaire could not be accessed until participants consented to continue. Once consent was given participants were asked a number of questions relating to basic demographics. They were then presented with the two targeted statements relating to the norm principle (in randomised order) and asked to select the statement they “identify with the most”. Participants were then presented with either Block A (containing the targeted questions relating to the application of the harm principle) or Block B (containing the moral scenario, related questions, and targeted questions relating to the endorsing of the harm principle). Following this participants were presented with the second block. As in Study 4, the questionnaire ended with the MLQ (Steger et al., 2008); and CRSi7 (Huber & Huber, 2012).

4.5.2 Results and discussion. Seventy nine of the total sample (*N* = 111; 71.17%) rated the behaviour of Julie and Mark as wrong initially. The mean initial rating of the behaviour (immediately following presentation of the scenario) for the entire sample was, *M* = 2.35, *SD* = 1.67. An independent samples t-test revealed no difference in initial rating between the MTurk sample (*M* = 2.08, *SD* = 1.48), and the MIC sample, (*M* = 2.68, *SD* = 1.83), *t*(93.31) = 1.86, *p* = .066. Sixty seven of the total sample, (N = 111; 60.36%) rated the behaviour as wrong after viewing the counter-arguments and the critical slide. The mean revised rating of the behaviour (immediately following the critical slide) for the entire sample was, *M* = 2.62, *SD* = 1.71. An independent samples t-test revealed a significant difference in revised rating between the MTurk sample, (*M* = 2.31, *SD* = 1.53), and the MIC sample, (*M* = 3, *SD* = 1.84), *t*(95.40) = 2.11 , *p* = .037. A paired samples t-test revealed a significant difference in rating of behaviour from time one, initial rating, (*M* = 2.35, *SD* = 1.67), to time two, revised rating, (*M* = 2.62, *SD* = 1.54), *t*(110) = -3.47 , *p* < .001. Further analysis revealed that although 15 participants changed their judgement, only two participants changed fully the valence of their judgement, changing their judgement from “wrong” to “right”. Of the other changes in judgement, ten participants changed their judgement from “wrong” to “neutral”; two participants changed their judgement from “right” to “neutral”; and one changed their judgement from “neutral” to “right”. A chi-squared test for independence revealed no significant association between time of judgement and valence of judgement made, χ^2(2, *N* = 222) = 3.40, *p* = .183, *V* = .12.

Eighteen participants (16%) indicated that they had encountered the scenario before. As, in Study 4, when asked to elaborate, participants provided anecdotes, or referred to previous readings/TV (either fiction or philosophy), 8 participants (7%) indicated that they had encountered it in a previous survey. The number of participants indicating previous experience with the scenario was higher than in Study 4 and as such the possibility that it may have confounded the results was investigated. An independent samples t-test revealed no difference in judgement between participants who had previously seen the scenario, (*M* = 2.83, *SD* = 1.86), and participants who had not previously seen the scenario, (*M* = 2.26, *SD* =1.62), *t*(22.31) = 1.228, *p* = 0.232. Furthermore, a chi-squared test for independence revealed no significant association between previous experience with the scenario and response to the critical slide, $\chi^2(2, N = 111) = 3.16, p = .206, V = .17$. On this basis these participants were not excluded from the sample.

Recall that the questions were blocked for randomisation. Block A contained the targeted questions relating to the application of the harm principle, and Block B contained the scenario and the related questions (including the critical slide) along with the credulity check regarding the harm principle. Tests for order effects revealed no difference in initial rating of the behaviour depending on order of blocks, $t(107) = -1.64, p = .104$. There was no difference in responding to the critical slide depending on order of blocks, $\chi^2(2, N = 111) = 4.76, p = .093$. Of the three questions relating to the application of the harm principle, there was no difference in response to the generic potential harm question ("How would you rate the behaviour of two people who engage in an activity that could potentially result in harmful consequences for either of them?") depending on the order of the bocks, $t(85) = -1.02, p = .312$. A chi-squared test for independence revealed no significant association between order of blocks and judgements of boxing ("Do you think boxing is wrong?"), $\chi^2(1, N = 111) = 2.86, p = .091, V = .16$, or the question regarding contact team sports ("Do you think playing contact team sports (e.g., rugby; ice-hockey; American football) is wrong?"), $\chi^2(1, N = 111) = .19, p = .660, V = .04$.

The order of the questions regarding the application of the harm principle was also randomised. A one-way ANOVA revealed a significant difference in responses to the question "How would you rate the behaviour of two people who engage in an activity that could potentially result in harmful consequences for either of them?" (1 = *Extremely wrong*; 4 = *neutral*; 7 = *Extremely right*) depending on when it was presented $F(2, 104) = 4.757, p = .011$, partial $\eta^2 = .080$. Tukey's post-hoc pairwise revealed that, when this question was responded to first, participants ratings were significantly lower (*M* = 2.80, *SD* = 1.43) than when it was responded to second (*M* = 3.57, *SD* = 1.21), *p* = .040, or third (*M* = 3.68, *SD* = 1.31), *p* = .014; and there was no difference in responding to this question second (*M* = 3.57, *SD* = 1.21), *p* = .040, or third (*M* = 3.68, *SD* = 1.31), *p* = .932.

A chi-squared test for independence revealed no significant association between order these questions and responses to the question “Do you think boxing is wrong?”, $\chi^2(2, N = 112) = 4.88, p = .087, V = .21$. Similarly, a chi-squared test for independence revealed a significant association between order these questions and responses to the question “Do you think playing contact team sports (e.g. rugby; ice-hockey; American football) is wrong?”, $\chi^2(2, N = 112) = 1.7822, p = .409, V = .13$.

The order of the blocks had no influence on the any of the responses of interest. There were differences in responding to the question relating to the general application of the harm principle (“How would you rate the behaviour of two people who engage in an activity that could potentially result in harmful consequences for either of them?”). This question was more abstract than the two questions it appeared with, in which participants were asked to judge a named behaviour (boxing or contact team sports). The description in the general question could apply to either of the named behaviours. Participants who responded to this question first rated the behaviour as more wrong than participants who responded to it after reading one or both of the named behaviours. It seems likely that the named behaviours provided an example of a situation in which the behaviour described in the general question may be acceptable, leading participants to respond more favourably to the general question.

4.5.2.1 Baseline measure of dumbfounding. Participants who selected the admission of not having reasons on the critical slide were identified as dumbfounded. Twenty one participants (18.92%) were initially identified as dumbfounded by their selecting of the dumbfounded response ("It's wrong but I can't think of a reason") on the critical slide. Thirty six participants (32.43%) selected "There is nothing wrong"; 54 (48.65%) participants selected "It's wrong and I can provide a valid reason". Participants who selected "It's wrong and I can provide a valid reason" were required to provide a reason. The reasons provided were analysed and coded for dumbfounded responses, defined as (a) unsupported declarations or (b) tautological reasons. A total of 6 participants were identified as dumbfounded following this coding. Four participants provided an unsupported declaration (e.g., "Incest is wrong"; "100% wrong"), one provided a tautological reasons ("They are brother and sister which make it incest"), and one participant provided both an unsupported declaration and a tautological reason as justification for their judgements. This brought the total number of participants identified as dumbfounded to 27 (24.32%).

This baseline measure of dumbfounding revealed similar rates of dumbfounding across Studies 4 and 5, with 18.18% selecting selecting the admission of not having reasons in Study 4 compared with 18.92% in Study 5. These rates remained similar when the coded open-ended responses were included: 23.64% in Study 4 compared with 24.32% in Study 5.

4.5.2.2 Endorsing/articulating harm and norm principles. The exclusion criteria developed by Royzman et al. (2015) (the endorsing of either principle) were applied, and this left a sample of 20 who were eligible for analysis. Two of these fully convergent participants selected "It's wrong but I can't think of a reason". Two selected “It's wrong and I can provide a valid reason”. Sixteen selected “There is nothing wrong”. Recall that in Study 4 this measure of convergence resulted in a high proportion of participants who selected “There is nothing wrong” to be falsely excluded from analysis (12/22; 54.55%). This pattern can be found again in Study 5, 20 of the 36 participants who selected “There is nothing wrong” (55.56%) were falsely excluded from analysis by using the measure of convergence developed by Royzman et al. (2015).

The revised criteria for exclusion (both articulating and endorsing either principle) developed in Study 4 were then applied, and the number of participants eligible for analysis increased to 61. Of the 61 participants who were eligible for analysis according to the revised exclusion criteria, nine (14.75%) selected “It’s wrong but I can’t think of a reason”. Thirty three participants (54.10%) selected “There is nothing wrong”, and 19 participants (31.15%) participants selected “It’s wrong and I can provide a valid reason”. Again this led to a reduction in false exclusions, three of the 36 (8.33%) participants who selected “There is nothing wrong were excluded by this measure. The reasons provided were coded for additional dumbfounded responses. One participant provided an unsupported declaration (writing: “100% wrong.”), however, as in Study 4, this was not taken as a dumbfounded response because it may be viewed as an articulation of the norm principle. Observed rates of dumbfounded responding following the application of the revised exclusion criteria (14.75%) were similar to those observed in Study 4 (15.38%). According to the revised criteria for exclusion (developed in Study 4), and a stricter measure of dumbfounded responding (admissions of having no reason only), a total of nine participants from a sample of 110 (8.18%) were identified as dumbfounded.

4.5.2.3 Consistency in applying the harm principle. The purpose of Study 5 was to identify if people are consistent in applying the harm principle. Three targeted questions were included to assess this. One question related to the harm principle in general terms, with no mention of specific behaviours. Two other questions related to sports whereby harm may result; one named an individual sport with a high risk of harm (boxing), the other question listed a number of contact team sports with a moderate to high risk of harm (rugby; ice-hockey; American football). The implications of the responses to each of these questions are addressed for each question individually, before being analysed together.

4.5.2.3.1 Potential harm generally. According to the responses to the generalised potential harm targeted question, 55 (49.55%) participants believe it is wrong to "engage in an activity that could potentially result in harmful consequences for either of them". A chi-squared test for independence revealed no significant association between order of Blocks A and B and response to this question, $\chi^2(2, N = 111) = 1.02$, $p = .600$, $V = .10$. Similarly, an independent samples t-test revealed no difference in responses to this generalised potential harm question between sample that were presented with Block A (containing targeted questions relating to the application of the harm principle) first, ($M = 3.46$, $SD = 1.36$), and the sample that were presented with Block B (containing the scenario and related questions) first, ($M = 3.19$, $SD = 1.40$), $t(85.40) = -1.02$, $p = .312$.

The responses to the general potential harm question (application) were taken together with the open-ended responses (articulation) and targeted questions relating to the harm principle (endorsing). Only participants who were consistent in their responses across all three dimensions are excluded. This left a sample of 100 (90.09%) eligible for analysis. Of these, 19 participants (17.12% of the total sample) selected the dumbfounded response. When participants who endorsed and articulated (through unsupported declarations, tautological reasons, or otherwise) the norm principle were also excluded from analysis, this sample was reduced to 69. Ten (9.01% of the total sample) of these participants selected the dumbfounded response.

4.5.2.3.2 Boxing. The targeted question relating to boxing revealed that 26 participants (23.42%) believed that boxing is wrong. A chi-squared test for independence revealed no significant association between order of Blocks A and B and response to this question, $\chi^2(1, N = 111) = 2.86, p = .091, V = .16$. The responses to the boxing question were taken together with the open-ended responses (articulation) and targeted questions (endorsing) relating to the harm principle. Again, only participants who were consistent across these three dimensions were excluded from analysis, leaving a total sample of 105 participants eligible for analysis. Twenty one of these participants (18.92% of the total sample) selected the dumbfounded response. When participants who endorsed and articulated (through unsupported declarations, tautological reasons, or otherwise) the norm principle were excluded from analysis, this sample was reduced to 69. Ten (9.01% of the total sample) of these participants selected the dumbfounded response.

4.5.2.3.3 Contact team sports. The targeted question relating to contact team sports identified three participants (2.70%) who believed that such sports are wrong. Due to this small number, no test for order effects of the blocks was conducted. As above, these responses were taken together with the open-ended responses (articulation) and targeted questions (endorsing) relating to the harm principle and only participants who were consistent across these three dimensions were excluded from analysis. This left a total sample of 109 participants eligible for analysis.

Twenty one of these participants (18.92% of the total sample) selected the dumbfounded response. When participants who endorsed and articulated (through unsupported declarations, tautological reasons, or otherwise) the norm principle were excluded from analysis, this sample was reduced to 72. Ten (9.01% of the total sample) of these participants selected the dumbfounded response.

4.5.2.3.4 Applying the harm principle across contexts. The responses to the three targeted questions relating the application of the harm principle were analysed together. Only one participant was consistent in their application of the harm principle across all three targeted questions. Similarly, only one participant was consistent in the application, articulation, and, endorsing of the harm principle (as measured by the open-ended responses and the targeted questions taken from Royzman et al. (2015)). The responses to the critical slide across all measures of convergence used are displayed in Figure 4.2.

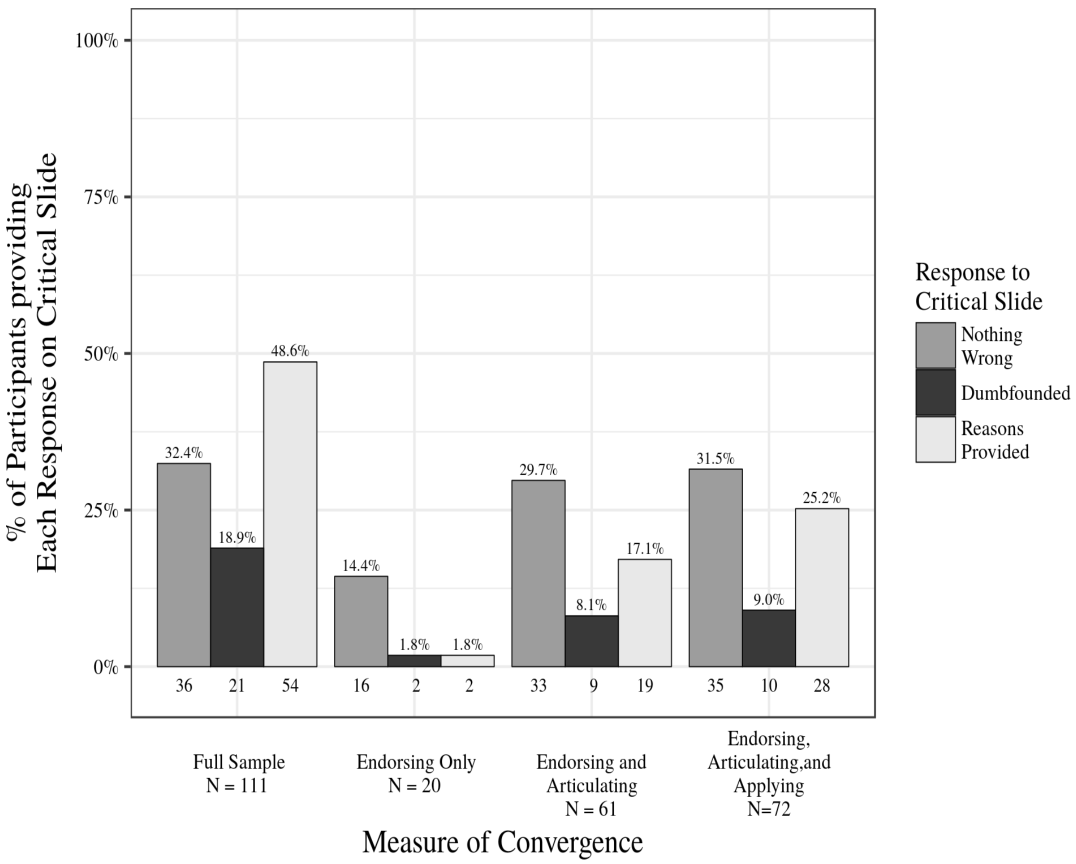

Figure 4.2: Study 5 – Responses to critical slide for the Entire Sample, and for each measure of convergence: (i) Endorsing only, (ii) Endorsing and Articulating, and (iii), Endorsing, Articulating, and Applying (all percentages are expressed in relation to the entire sample).

The number of participants excluded based on their response to the contact team sports question were noticeably lower than for the other questions relating to the harm principle. This question specifically relates to the harm principle, however there is a possibility that responses to this question do not accurately reflect the application of the harm principle as measured by the other targeted questions. In the interests of rigour, the team contact sport question was ignored and the consistency of the use of the harm principle was analysed again. When this question was removed, there were four participants who were consistent in their applying, endorsing, and articulating of the harm principle, leaving a sample of 107 eligible for analysis. A chi-squared test for independence revealed no significant association between the inclusion of the contact team sports question and the number of participants eligible for analysis, $\chi^2(1, N = 222) = 0.82, p = .366, V = .06$. As such all three questions relating to the application of the harm principle are included in the analysis that follows.

Only one participant consistently applied, articulated and endorsed the harm principle such that the total sample eligible for analysis based on the harm principle was 110. Of these participants, 21 of these participants (18.92% of the total sample) selected the admission of having no reason, presenting as dumbfounded. When the articulation and endorsing of the norm principle is also included in the criteria for exclusion the total sample is reduced to 73. Of these, ten (9.01% of the total sample) selected the admission of having no reasons.

4.5.2.4 Judgement-principle consistency. As in Study 4, the initial criteria for exclusion (endorsing only) excluded a large proportion of the participants who selected "There is nothing wrong"; 20 of the 36 (55.56%) of the participants who selected "There is nothing wrong" were excluded. When articulation of the principles was accounted for, only three (8.33%) of these 36 participants were excluded. This is higher than in Study 4 (one participant, 4.55% of those who selected "There is nothing wrong"), however in reducing the obvious false exclusion of participants who selected "There is nothing wrong" it remains an improvement on the original criteria. This suggests that accounting for participants' ability to articulate the principles endorsed provides a more accurate criteria for exclusion than accounting only for the endorsing of a given principle. Furthermore, when the applying of the harm principle was also accounted for, only one of the 36 participants who selected "There is nothing wrong" was excluded. The criteria for convergence developed here lead to greater consistency between a participant's eligibility for analysis and their judgement made than the original criteria described by Royzman et al. (2015).

Study 5 investigated the consistency with which people apply, articulate, and endorse the harm principle. Only one participant consistently applied, articulated, and endorsed the harm principle. As such, the harm principle as a basis for exclusion from analysis becomes practically redundant. The endorsing and articulation of the norm principle resulted in the exclusion of 37 participants. The degree to which the articulation or the endorsing of the norm principle may render participants ineligible for consideration as dumbfounded is unclear, this is discussed in more detail below. However, even if participants are excluded from analysis based on the norm principle, dumbfounded responding is still observed, with ten participants (13.70% of sample eligible for analysis; 9.01% of the total sample) selecting the admission of having no reason on the critical slide.

4.6 General Discussion

The aim of the Studies 4 and 5 was to assess if the judgements of dumbfounded participants can be attributed to moral principles based on their endorsing of these principles. This was done by assessing the consistency with which participants articulate and apply these moral principles. Royzman et al. (2015) argue that, if participants endorse a principle, their judgement can be attributed to that principle. They claimed that by attributing participants' judgements to particular principles in this way, moral dumbfounding can be eliminated. However, attributing judgements to reasons based on the endorsing of a related principle is problematic. Stronger evidence that a participant's judgement may be attributed to a given principle should account for (a) the participant's ability to articulate this principle, independent of a prompt; or (b) the consistency with with the participant applies the principle across differing contexts. Two studies were conducted to address these issues. It was found that participants do not consistently articulate (Study 4) or apply (Study 5) principles that they may endorse. In these cases participants' judgements were not be attributed to these principles, and evidence for dumbfounding was found. It appears that dumbfounded responding may indeed be indicative of a state of dumbfoundedness, rather than being attributed to features of the experimental design (Research Question 1.3 in Chapter 2).

4.6.1 Articulating principles. Study 4 showed that participants, who endorse a given principle, do not necessarily articulate it. Of the 52 participants who endorsed the norm principle, only 36 also articulated it when given the opportunity. Similarly, of the 87 participants who endorsed the harm principle, only 29 also mentioned harm in their open ended responses. These figures are similar for Study 5, with 60 participants endorsing the norm principle, and only 37 of these mentioning norms, and 85 endorsing the harm principle, only 23 of whom also mentioned harm. Across both studies (*N* = 221) 187 (84.62%) endorsed one of the principles, of these, only 108 (of these 187 participants; 57.75%) also articulated the same principle. This inconsistency between the endorsing and articulation of principles that are purported to be governing moral judgements suggests that endorsing alone provides a poor measure of whether these principles serve as reasons for a given judgement.

4.6.2 Applying the harm principle. It was predicted that participants would not consistently apply the harm principle across differing contexts. However, the degree to which this was found to be the case was surprising. Only 1 participant applied the harm principle consistently across all three targeted questions (potential harm generally, boxing, contact team sports). According to the argument put forward by Royzman et al. (2015; see also Gray et al., 2014; Jacobson, 2012), the judgements of dumbfounded participants can be attributed to the harm principle. If participants' judgements are grounded in the harm principle, then it follows that this principle should be applied across differing contexts (or some further statement of the principle such that situations in which it should and should not be applied are made explicit.). This was clearly shown not to be the case, and the claim that the harm principle is governing the judgements of dumbfounded participants is not supported.

4.6.3 The norm principle and unsupported declarations. In Studies 4 and 5, unsupported declarations were coded as an articulation of the norm principle. However, in Chapter 3 parallels between the providing of unsupported declarations and the providing of admissions of not having reasons were identified (similar proportion of time spent (a) smiling/laughing, (b) in silence). There is also a strong theoretical case for the inclusion of unsupported declarations as dumbfounded responses.

Finally, the theoretical framework adopted here, identifies propositional beliefs/deontological judgements as habitual/model-free intuitions. As discussed in Chapter 2, the learning and maintaining of an intuition occurs through continued and consistent type-token interpretation, which means that the emergence of a moral intuition may occur independently of the reasons for the intuition. Stating the content of the intuition, is not the same as providing a reason for the intuition. Royzman et al. (2015) argue that holding the propositional belief is justification enough for a judgement, however this is holding participants to a different standard. There is a difference between having a reason for an intuition/propositional belief and claiming that the reason for a judgement is grounded in an associated propositional belief. In view of this, it is possible that by not including unsupported declarations or tautological reasons as dumbfounded responses, the level of dumbfounding reported in this chapter are under representative of the phenomenon. However, even according to this stricter measure, evidence for dumbfounding was found.

4.6.4 Implications. The existence of moral dumbfounding and the associated support for intuitionist theories of moral judgement (e.g. Cushman, Young, & Greene, 2010; Haidt, 2001; Hauser et al., 2008; Prinz, 2005; see also Crockett, 2013; Cushman, 2013; Greene, 2008, 2013) has been challenged in recent years. The majority of these challenges are theoretical (e.g., Gray et al., 2014; Jacobson, 2012; Sneddon, 2007; Wielenberg, 2014). Royzman et al. (2015) appeared to give some empirical weight to these challenges. Two studies presented here address specific methodological limitations associated with the work by Royzman et al. (2015) and evidence for dumbfounding was found.

4.6.5 Limitations and future directions. The role of social pressure and conversational norms in the emergence of moral dumbfounding is not well understood. Royzman et al. (2015) argue that dumbfounding occurs as a result of social pressure to conform to conversational norms. The evidence they present does not support the claim that dumbfounding is caused by social pressure, however, they do show that dumbfounded responding is sensitive to social pressure. The initial estimate of incidences of dumbfounding found by Royzman et al. (2015) was 4/53 (7.55%). These four participants were then interviewed further, during which, the “inconsistencies” in participants’ “responses were pointed out directly” and they were “advised to carefully review and, if appropriate, revise” their responses (Royzman et al., 2015, p. 308). Following this interview they were left with only one participant who is presents as dumbfounded. This is a clear demonstration that dumbfounding can be reduced by social pressure to “appear consistent”. However, the studies described in this chapter provide evidence that dumbfounded responding cannot be attributed to social pressure alone. The processes by which we make moral judgements also give rise to moral dumbfounding. This means that isolating the underlying mechanisms that give rise to moral dumbfounding may contribute to our overall understanding of the making of moral judgements.

This provides the focus of the remainder of this thesis. Firstly, a possible explanation of moral dumbfounding based on dual-process theories of moral judgement (e.g., Cushman, 2013) is identified and two predictions of this explanation are tested. Secondly, an explanation of moral dumbfounding that draws on model theory (Bucciarelli et al., 2008) is examined. Finally an alternative theoretical approach to the study of moral judgement, that may provide an explanation for moral dumbfounding is explored.

4.7 Conclusion

The studies presented in this chapter addressed methodological issues with the work by Royzman et al. (2015), specifically, the exclusion criteria/measure of convergence (endorsing principles only) they developed were too broad. It was shown that participants who endorse a moral principle may not necessarily articulate that principle when given an opportunity. It was also shown that participants who endorse the harm principle, do not consistently apply it across differing contexts. In view of these findings, participants' judgements should not be attributed to particular principles based only on the endorsing of these principles. A stronger measure of convergence/criteria for exclusion should account for the articulation and the application of moral principles. Strong support for this revised measure of convergence can be found in the dramatic reduction in the false exclusion of participants who did not rate the behaviour as wrong, demonstrating greater judgement-principle consistency. Dumbfounding was measured using the stricter measure of dumbfounding developed in Chapter 3, in order to reduce ambiguity in results. Using this stricter measure of dumbfounding, evidence for dumbfounding was found across both Studies 4 and 5.

The studies described in this chapter demonstrated that moral dumbfounding occurs as a result of the processes that underlie moral judgement (as opposed to social pressure as claimed by Royzman et al., 2015) and it does not appear that dumbfounding can be adequately explained by rationalism (Chapter 2, Research Question 2.1.1). Building on this finding, and drawing on the materials and procedures that have been developed across 5 studies, the remainder of this thesis will attempt to contribute to our understanding of the making of moral judgements by attempting to provide an explanation of moral dumbfounding.

5 Chapter 5 – Influencing Dumbfounded Responding 1: Inhibiting the Identification of Reasons

The previous two chapters have identified methods for studying dumbfounding and presented evidence that dumbfounding is a genuine phenomenon that withstands the rationalist challenge. Chapter 3 demonstrated dumbfounding and developed the means to elicit and study moral dumbfounding. Chapter 4 tested a rationalist explanation of dumbfounding proposed by Royzman, Kim, and Leeman (2015) and addressed specific limitations in their methods. The studies in Chapter 4 employed a stricter measure of dumbfounding, and as such estimates of the prevalence of moral dumbfounding reported in Chapter 4 were revised downwards from those reported in Chapter 3. Nevertheless, dumbfounded responding was reliably elicited, posing a significant challenge to the rationalist explanation of dumbfounding offered by Royzman et al. (2015).

According to a rationalist perspective (e.g., Royzman et al., 2015), people do have reasons for their judgements, but they give in to the counter-arguments of the experimenter in response to social pressure to avoid appearing uncooperative or stubborn. The studies in Chapter 4 demonstrated that participants do not consistently articulate the reasons that are claimed to be guiding their judgements by Royzman et al. (2015). Furthermore, the harm principle that underlies one of these reasons, is not applied across different contexts. These findings are in clear opposition to the rationalist claim that moral judgements are based on reasons or grounded in principles.

In contrast to rationalist perspectives, dual-process approaches to moral judgement (e.g., Cushman, 2013) do not attribute the making of moral judgements to reasons or principles. In line with dual-process theories of cognition more generally (e.g., Evans, 2010) judgements can be intuitive or automatic, such that reasons for these judgements are beyond conscious awareness. After a judgement is made deliberation may be employed to seek conscious reasons to support a judgement (Evans, 2010, p. 7). Moral dumbfounding is therefore more consistent with dual-process approaches to cognition than rationalist approaches. The studies described in this chapter aim to examine two predictions of a dual-process explanation of moral dumbfounding, examining Research Question 2.1.2 (Can the existence of moral dumbfounding be adequately explained by Dual-Process approaches to moral judgement).

5.1 Moral Dumbfounding as Dual-Processes in Conflict

According to a dual-process account of the making of moral judgement, the making of moral judgements generally occurs as a result of habitual responding, while the identification of reasons for a judgement requires deliberation (Brand, 2016; Cushman, 2013; Haidt, 2001; Haidt & Björklund, 2008). This means that the identification of reasons for a judgement occurs independently of the making of the judgement. Moral dumbfounding then occurs when deliberation fails to yield a justification for a habitual response. It is possible that the failure to identify reasons for a judgement places the deliberative response in conflict with the habitual response, in that, the failure to identify reasons for a particular judgement may lead to the conclusion that the initial judgement should be revised.

This type of inconsistency between an intuitive response and a deliberative response has been identified in the dual-process literature more generally. Consider the following puzzle (taken from De Neys, 2012, p. 29):

> A psychologist wrote thumbnail descriptions of a sample of 1000 participants consisting of 995 females and 5 males. The description below was chosen at

random from the 1,000 available descriptions.

Jo is 23 years old and is finishing a degree in engineering. On Friday nights, Jo likes to go out cruising with friends while listening to loud music and drinking beer.

Which one of the following two statements is most likely?

a. Jo is a man

b. Jo is a woman

According to the description Jo appears to engage in behaviours that are stereotypically "male". The intuitive response given the description therefore would be that Jo is more likely to be man. However, this response neglects the statistical probability that has been provided in the opening section of the puzzle. The probability that Jo is male is 0.5% while the probability that Jo is female is 99.5%, therefore the correct response is that Jo is more likely to be a woman. This correct response is in conflict with the intuitive response (that Jo is more likely to be a man, based on the description of the behaviour). In order to correctly answer the puzzle, it is necessary to ignore or over-ride the initial intuition and engage in deliberation regarding the probabilities provided in the question. Essentially deliberation is required to overrule the erroneous intuition (habitual response).

The above illustration of conflict in dual-processes is known as a base rate neglect problem (where people greatly overestimate the probability of an event due the reliance on stereotypical implications while not considering the base rate enough or at all Bonner & Newell, 2010; De Neys, 2012; De Neys & Glumicic, 2008; Evans, 2007; Tversky & Kahneman, 1983). Other examples of conflict in dual-processes include the "conjunction fallacy task", and the "syllogistic reasoning task" (De Neys, 2012, p. 29). In brief, for the conjunction fallacy task people are presented with a description of a person and asked to judge the likelihood of two statements. One statement contains a single non-representative attribute (e.g., a hobby that is intuitively inconsistent with the personality of the person described) and the other statement contains the non-representative attribute in conjunction with a more representative attribute (e.g., an occupation that is consistent with the person described). The conjunction fallacy occurs when people deem the second statement as more likely than the first. In reality, the inclusion of two attributes makes the second statement statistically less likely, however, because one of the attributes is representative of the description they have read, participants are likely to select it. In the syllogistic reasoning task people are presented with a set of premises that yield a logical yet non-intuitive conclusion. People reject the logical conclusion based on their intuitions as opposed to deliberating on the logic of the conclusions from the premises. Other examples of conflict include the persistence of apparently "irrational" and compulsive behaviours (e.g., overeating, smoking, or gambling; Evans, 2008). A person may, through deliberation, judge such behaviours as at odds with their long term goals yet continue to engage in them through force of habit.

A clear example linking dual-process conflict and moral dumbfounding can be found in a study by Rozin, Markwith, and McCauley (1994; as discussed by Lerner & Goldberg, 1999, p. 634). They found that people report reduced willingness to contact various items that they believed had prior contact with an AIDS victim, someone who had been in a car accident, or a murderer, despite assurances that these items are sanitary (Rozin et al., 1994). This paradigm closely resembles one of the tasks used by Haidt, Björklund and Murphy (2000) as part of their unpublished moral dumbfounding study. Their study contained three moral judgement tasks and two non-moral tasks, all of which appeared to elicit dumbfounding. As part of one of the non-moral tasks, the experimenter dipped a sterilised cockroach in a glass of juice. Participants were then asked to drink from the glass. Much like the study by Rozin et al. (1994), participants were unwilling to drink from the glass (Haidt et al., 2000). The finding by Rozin et al. (1994) is described in terms of a conflict between implicit (habitual) processes and conscious (deliberative) responses (Lerner & Goldberg, 1999). Haidt et al. (2000) present their version of the similar task as equivalent to the moral dumbfounding they observed in the moral judgement tasks. As such, explaining dumbfounding in terms of a conflict of habitual and deliberative responding is consistent with the literature on both conflict in dual-systems and on moral dumbfounding.

In classic cases of moral dumbfounding, the habitual response is that of condemnation of the action described (Haidt et al., 2000). A person will then engage in deliberation in order to identify reasons in support of this condemnation. If this deliberation is unsuccessful, the deliberative response may be to revise the judgement, placing the deliberative response in conflict with the habitual response. Resolving this conflict involves either: the revising of the initial judgement, or, further deliberation and successful identification of a reason to support the habitual response. If a person does not change their judgement, and cannot identify reasons to justify their habitual response, they fail to resolve the conflict and present as dumbfounded. Dumbfounding then, according to this view, is the failure to resolve conflict between habitual and deliberative responses.

Normal cases of conflict can be resolved either by (a) the over-riding of the habitual response or (b) the ignoring of the inconsistent information from deliberation. Recall the puzzle discussed above; according to the description, Jo appears to present as stereotypically male. Deliberation reveals that Jo is 199 times more likely to be a woman (99.5%) than to be a man (0.5%). In this case the deliberative response is in clear conflict with the habitual response. This conflict is resolved if a person accepts that their intuition was incorrect and adopts the deliberative response. Alternatively a person may choose to ignore the inconsistent information that resulted from deliberation and maintain their initial judgement.

It is hypothesised that moral dumbfounding occurs when neither of these strategies is perceived to be available. Consider (a) the over-riding of the intuitive response: there is a rich body of research demonstrating that people appear unwilling to over-ride habitual responses on certain issues, particularly moral issues (e.g., Abelson, 1988; Kruglanski & Webster, 1996; Kruglanski, Webster, & Klem, 1993; McGregor, 2006b, 2006a; McGregor, Zanna, Holmes, & Spencer, 2001). From this, it is likely that, in cases of conflict relating to moral issues, people resolve conflict through (b) the ignoring of inconsistent information from deliberation.

Two types of dumbfounded responses were identified in Chapter 3: (a) unsupported declarations and (b) admissions of not having reasons. The salience of inconsistent information is much greater when a person admits to not having reasons than when they simply provide an unsupported declaration. As such, an unsupported declaration is likely a much more attractive response than admitting to not having reasons for a judgement. Indeed, in Chapter 3, Study 2 (section 3.3) measuring dumbfounding as the selection of an unsupported declaration revealed remarkably high rates of dumbfounding. In the absence of explicit reminders that deliberation did not provide reasons to support their judgement, participants appeared to readily dismiss a deliberative response.

5.1.1 Dual-process classifications of responses in the moral dumbfounding paradigm. In traditional dual-process conflict studies there are generally two responses: a correct response and an incorrect response; and two types of responses: logical (deliberative) and intuitive (habitual). This binary correct/incorrect, logical/intuitive classification means that the operationalisation of dependent variables in these studies is straight forward, a responses is either correct or incorrect. The correct response always maps onto the logical (deliberative) and various manipulations can frame the intuitive response as either correct or incorrect.

In contrast, there is no clear correct or incorrect answer in moral judgement tasks. In addition to this, the study of moral dumbfounding (in the common case of an initial judgement that the target behaviour is wrong) involves at least three different responses (1) the providing of reasons; (2) accepting the counter-arguments and rating the behaviour as “not wrong”; or (3) a dumbfounded response. The positioning of these responses options in terms of habitual or deliberative responses may vary from person to person or from situation to situation.

The nature of the dumbfounding paradigm, which involves explicit arguments against the most common judgements, requires that participants engage in some deliberation. This means that all responses are likely the result of some level of deliberation in order to support (or attempt to support) an initial intuition. A simplistic view of the possible responses identified above would identify response (1: reasons) as the successful alignment of deliberation and intuition. Response (2: nothing wrong) is the over-riding of intuition by deliberation, while response (3: dumbfounded) is the failure of deliberation to rationalise an intuition. This view however presumes that there is only one intuition at play, and participants’ deliberations are an attempt to provide reasons for this single intuition.

At least one other intuition that may become salient as part of the dumbfounding paradigm has been identified. Participants may have intuitions relating to the nature of moral knowledge, for example that moral judgements should be justifiable by reasons. During the course of the discussion/presentation of slides, this intuition (or alternative intuitions) may become salient. The emergence of this intuition is consistent with research on meaning maintenance (Heine et al., 2006; Proulx & Inzlicht, 2012; see also: Cialdini, Trost, & Newsom, 1995), and generally consistent with what has been observed in studies of moral dumbfounding (in particular Study 1 in Chapter 3), whereby people appear to be motivated to identify reasons for their judgement. This means that people experience more than a conflict between habitual and deliberative responding, they may also experience competing intuitions. The emergence of these competing intuitions may be attributed to deliberation, however, it is also possible that their emergence occurs as a result of the nature of the dumbfounding paradigm independently of level of deliberation; participants are asked to provide reasons for their judgements. The need for judgements to be justifiable by reasons may be made salient by simply asking participants to provide justifications for their judgements.

Three possible types of responses in the dumbfounding paradigm were listed above, (1) providing reasons; (2) accepting the counter-arguments and rating the behaviour as “not wrong”; and (3) a dumbfounded response. However two classes of dumbfounded responses have been identified. To reflect this, response (3) may be described as (3a) an admission of not having reasons and (3b) an unsupported declaration. The discussion below attempts to describe each of the responses 1-3 may be in terms of the varying roles of deliberation and intuition, at least two relevant (competing) intuitions have been identified.

It is likely that response (1: reasons) is the most desirable, it involves successfully resolving the conflicting intuitions and providing reasons for a judgement. However, response (1: reasons) also requires that deliberation is "successful", that is that deliberation results in successfully identifying a reason for a judgement. In the dumbfounding paradigm this is made particularly difficult because commonly identified reasons for judgements are refuted during the course of the study. This means that successfully identifying a reason requires identifying a reason beyond the refuted reasons, or identifying shortcomings in the refutations. On the other hand, response (3b: unsupported declaration) an unsupported declaration arguably requires the least deliberation. Defending a judgement with a restatement of the judgement can be done without deliberation. Furthermore, in restating a judgement and affirming a position, the salience of inconsistent intuitions may be reduced, much like the "seizing and freezing" behaviours described by Kruglanski and Webster (1996). This leaves (2: revising/nothing wrong) and (3a: admissions), it is probable that changing a judgement requires more deliberation than admitting to not having a reason for it; in changing a judgement a person would likely deliberate the strength of the counter-arguments. In Chapter 3 both types of dumbfounded responses were identified as more similar to each other than to the other types of responses. This was based on the analysis a range of non-verbal behaviours (e.g., laughing, smiling, silence). Differences were found across these measures depending on type of response, where the prevalence of these behaviours was significantly different in cases of dumbfounded responding when compared with cases of providing reasons or changing judgements in line with the counter-arguments. Crucially, no differences were found depending on type of dumbfounded response provided. It is likely that the observed similarity between both dumbfounded responses would extend to levels of deliberation. As such, dumbfounded responding is hypothesised to involve the least amount of deliberation, and providing reasons requires the most amount of deliberation, with changing judgement involving more deliberation than dumbfounded responding, but not as much deliberation as successfully providing reasons.

The only response that can be positioned as habitual or deliberative with any certainty is response (1: reasons), which is certainly deliberative, as it illustrates successful deliberation such that two competing intuitions may be aligned. The relative roles of intuition and deliberation in each of the other responses are less clear. There are at least two competing intuitions that may give rise to the remaining responses, that the behaviour is wrong, and that moral judgements should be grounded in reasons. Providing an unsupported declaration (response 3b) is a clear endorsing of the first of these intuitions (that the behaviour is wrong) over the second intuition. Arriving at this response is possible without deliberation. If a person does not see a need for judgements to be justified by reasons they are unlikely to engage in a deliberative search for reasons.

The remaining responses (2: revising/nothing wrong) and (3a: admissions) may be viewed as instances of selecting one of the competing intuitions following a failed deliberation. Providing response (2: revising/nothing wrong), a revised judgement may also be viewed as selecting one of the competing intuitions based on, and informed by deliberation, acknowledging the value of this deliberation process. For this reasons it is possible that response (2: revising/nothing wrong) may, in some cases, involve slightly more deliberation than response (3a: admissions).

The level of deliberation involved in the remaining responses (2: admission) and (3a) may vary depending on the individual. Some people may readily change their judgement based on new information, while others may not see the need to justify their judgement by reasons. The undesirability of an admission of having no reason is evidenced by the low rates of selecting/providing it in previous studies. Similarly, people appear reluctant to change their judgement. During the interview in Study 1, one participant changed their judgement on the “Julie and Mark” vignette, and when asked if they were happy with their decision they exclaimed “No!”.

In view of the above discussion, the responses may be (tentatively) ranked in order of the relative role of deliberation. Beginning with the highest level of deliberation and ending with the lowest, the responses may be ranked as follows: providing reasons (successful deliberation), accepting the counter-arguments and rating the behaviour as “not wrong” (failed deliberation/deliberation over-riding an initial intuition), an admission of not having reasons (failed deliberation/rejection of value of deliberation) and, an unsupported declaration (failed deliberation/rejection of value of deliberation/deliberation absent). As noted previously, providing reasons is the only response for which claims regarding the relative role of deliberation and intuition can be made with any degree of certainty.

5.2 Influences on Moral Dumbfounding

One prediction of explaining dumbfounding as conflict in dual-processes is that under specific manipulations, responses in the moral dumbfounding paradigm should vary in predictable ways. Beyond the application of an external manipulation, responses in the moral dumbfounding paradigm may display variability that can be linked to specific individual difference variables. The studies described in this chapter aim to investigate both of these possibilities.

5.2.1 Influencing moral dumbfounding through external manipulation.

Cognitive load is widely accepted as inhibiting deliberative responding (e.g., De Neys, 2006; Evans & Curtis-Holmes, 2005; Evans & Stanovich, 2013; Schmidt, 2016). On the basis of the discussion above, I have identified providing reasons as involving more deliberation than alternative responses. This implies that cognitive load should inhibit the identification of reasons for a judgement, leading to an increase in dumbfounding or an increase in accepting the counter-arguments and revising the judgement made.

According to Greene, Morelli, Lowenberg, Nystrom and Cohen, (2008), cognitive load influences utilitarian judgements but not deontological judgements. The judgements in dumbfounding paradigms are deontological; the moral violations described in the scenarios in studies of moral dumbfounding are violations of widely accepted deontic propositions. Changing a judgement in the dumbfounding paradigm means rejecting a deontological judgement in favour of a judgement informed by a utilitarian position (the counter-arguments highlight the lack of harm in the scenarios). Drawing on Greene et al. (2008), who showed that utilitarian judgements are negatively influenced by cognitive load, it is expected that cognitive load should lead to more dumbfounded responding, rather than changing of judgements. This prediction is purported to be supported by Haidt et al. (2000). In the opening note of the original Haidt et al. (2000) report, they report that they conducted a second study in which they manipulated cognitive load. They report that they found that cognitive load lead to increased levels of dumbfounding but did not influence judgements made. Beyond a brief mention in the opening note, this cognitive load and moral dumbfounding study is not reported in full in Haidt et al. (2000) or elsewhere.

An investigation of the dumbfounding under cognitive load can test two aspects of dual-process models of moral judgement. Firstly, it is hypothesised that deliberative responding generally will be inhibited by cognitive load, leading to less identification of reasons for judgements. This inhibition may result in higher rates of selecting “there is nothing wrong” or higher rates of dumbfounded responding (or both). However, adopting the work by Greene et al. (2008) suggests that the inhibition of deliberation should result in higher rates of dumbfounded responding only.

Again, there are three possible responses in the dumbfounding paradigm: (1) providing reasons; (2) a change in judgement; and (3) a dumbfounded response. Providing reasons has been identified as requiring more deliberation than the other responses, and as such the introduction of a cognitive load manipulation should reduce the providing of reasons in favour of one of the other responses. It is not clear whether participants would be more likely to revise their judgement or provide a dumbfounded response.

5.2.2 Individual differences in moral dumbfounding. It is likely that responses in the dumbfounding paradigm will vary depending on individual differences. One individual difference variable linked to dual-process approaches to cognition, therefore may be related to susceptibility to dumbfounding is Need for Cognition (Cacioppo & Petty, 1982; Forsterlee & Ho, 1999; Petty, Cacioppo, & Kao, 1984; Petty, Feinstein, Blair, & Jarvis, 1996). The Need for Cognition Scale (NCS) is a measure of an individual’s tendency “to engage in and enjoy effortful analytic activity” (Forsterlee & Ho, 1999, p. 471; see also Cacioppo & Petty, 1982). In other words, it measures a tendency to engage in deliberation (Evans & Stanovich, 2013). It is also related to a person’s need to understand and make sense of the world (Forsterlee & Ho, 1999). It is hypothesised that people who score highly on the NCS will be more likely to provide reasons for their judgement. Related to this, people who score low on the NCS are likely to fail to identify reasons for their judgement.

That NCS is related to a need to understand and make sense of the world suggests that of the people who fail to identify reasons for their judgement, the people who revise their judgement will likely score higher on the NCS than people who provide a dumbfounded response.

The studies described in this chapter aim to investigate test two predictions of a dual-process explanation of moral dumbfounding. Testing the first prediction involves experimentally manipulating cognitive load in order to test if there is a relationship between cognitive load and participants' ability to identify of reasons for a judgement. If such a relationship exists, the prediction that dumbfounding will increase (as opposed to judgements changing Greene et al. 2008) will also be tested. The second prediction that will be tested is that a person's tendency to provide reasons will be related to their score on the need for cognition scale (Cacioppo & Petty, 1982; Petty et al., 1984).

5.3 Study 6: Dumbfounding and Cognitive Load 1 – College Sample

The primary aim of Study 6 was to investigate if a cognitive load manipulation would influence participants' ability to justify their judgement. The secondary aim of Study 6 was to investigate participants' ability to justify their judgement is related to need for cognition.

5.3.1 Method.

5.3.1.1 Participants and design. Study 6 was a between-subjects design with Need for Cognition additionally measured as a potential correlate and moderator variable. The dependent variable was response to the critical slide. The independent variable was cognitive load with two levels: present and absent. It was hypothesised that cognitive load would inhibit deliberation and lead to lower rates of providing reasons. It was also hypothesised that participants providing reasons would score higher on the Need for Cognition Scale.

A total sample of 66 participants[9] (55 female, 11 male; $M_{\text{age}} = 22.42$, min = 18, max = 57, $SD = 6.86$) took part. Participants in this sample were undergraduate students, postgraduate students, and alumni from Mary Immaculate College (MIC), and University of Limerick (UL). Participation was voluntary and participants were not reimbursed for their participation. The sample size in Study 6 is constrained by collecting data for a sixth related study in a small institution – participants who took part in studies 1-5 were not eligible to take part in Study 6. The constraints on data collection led to a low sample with limited power. Study 6 only has sufficient power to detect a large effect. Any observed null effects will therefore be inconclusive.

5.3.1.2 Procedure and materials. Data were collected using an online questionnaire generated using Questback (Unipark 2013). Data collection took place in a designated computer laboratory in MIC. The experimenter remained in the laboratory for the duration of the study. Participants were first presented with an information sheet and consent form. Following this, participants completed some questions relating to basic demographics.

Two statements, assessing if participants' judgements may be grounded in the norm principle were then presented: (i) "violating an established moral norm just for fun or personal enjoyment is wrong only in situations where someone is harmed as a result, but is acceptable otherwise."; (ii) "violating an established moral norm just for fun or personal enjoyment is inherently wrong even in situations where no one is harmed as a result." (Royzman, Kim, and Leeman 2015). Participants were asked to

[9] A priori power analysis indicated that in order to detect a large effect size ($V = .5$) with 80% power, a sample of 39 participants was required. In order to detect a medium effect size ($V = .3$) with 80% power a sample of 107 participants was required.

select the statement they “identify with the most”. The order of these statements was randomised. Participants who selected (ii) were asked to elaborate on their position. Participants in the experimental condition were then presented with an eight digit number/letter string and asked to memorise the sequence. After 30 seconds, the experiment progressed to the next slide. Participants had the option to click “ok” and progress to the next slide after 15 seconds.

Participants were then presented with the target moral scenario, the “Julie and Mark” (*Incest*) vignette (Appendix A), taken from the original moral dumbfounding study (Haidt, Björklund, and Murphy 2000). Participants rated, on a 7-point Likert scale, how right or wrong the behaviour of Julie and Mark was (where, 1 = *Morally wrong*; 4 = *neutral*; 7 = *Morally right*). They were then provided with an open ended response option and asked to provide a reason for their judgement. They then rated their confidence in their judgement. Following this, participants were presented with a series of counter-arguments, which refuted commonly used justifications for rating the behaviour as “wrong” (Appendix B). After each counter-argument, participants were asked “Do you (still) think it is wrong?”, with a binary “yes/no” response option; and then they were asked “Do you have a reason for your judgement?”, with three possible response options “Yes, I have a reason”, “No I have no reason”, and “Unsure”.

Dumbfounding was measured using the critical slide. The critical slide contained a statement defending the behaviour and a question as to how the behaviour could be wrong (“Julie and Mark’s behaviour did not harm anyone, how can there be anything wrong with what they did?”). There were three possible answer options: (a) “There is nothing wrong”; (b) an admission of not having reasons (“It’s wrong but I can’t think of a reason”); and finally a judgement with accompanying justification (c) “It’s wrong and I can provide a valid reason”. The order of these response options was randomised. Participants who selected (c) were then prompted on a following slide to type a reason. The selecting of option (b), the admission of not having reasons, was taken to be a dumbfounded response. Following the critical slide, participants rated the behaviour, and rated their confidence in their judgement again. They also indicated, on a 7-point Likert scale, how much they had changed their mind.

Participants in the experimental condition were then required to reproduce the eight digit number-letter string sequence presented previously. They were also asked to rate on a 7-point Likert scale how difficult they found the memory task. Following this a post-discussion questionnaire, taken from Haidt et al. (2000) was administered (Appendix C).

Two targeted questions relating to the harm principle (Royzman, Kim, and Leeman 2015) were then presented: (i) “Having read the story and considering the arguments presented, are you able to believe that Julie and Mark’s having sex with each other will not negatively affect the quality of their relationship or how they feel about each other later on?”; (ii) “Having read the story and considering the arguments presented, are you able to believe that Julie and Mark’s having sex with each other will have no bad consequences for them personally and/or for those close to them?”. Participants responded “Yes” or “No” to each of these statements (Royzman et al., 2015). The order of these questions was randomised.

Three targeted questions assessing the consistency of applying the harm principle were then presented. The questions were: (a) “How would you rate the behaviour of two people who engage in an activity that could potentially result in harmful consequences for either of them?”, responses were recorded on a 7-point

Likert scale (1 = *Morally wrong*; 4 = *neutral*; 7 = *Morally right*); (b) “Do you think boxing is wrong?”, with a binary “Yes/No” response option.; and (c) “Do you think playing contact team sports (e.g., rugby; ice-hockey; American football) is wrong?”, with binary “Yes/No” response option.

It was also hypothesised that ability to provide reasons for a judgement may be moderated by individual differences, and in particular Need for Cognition (Cacioppo & Petty, 1982; Petty et al., 1984). The short Need for Cognition scale was included (Petty et al., 1984). This is an 18 item scale containing questions relating to motivation to engage in thinking (e.g., “I would prefer complex to simple problems”). Responses were recorded on a -4 to +4 Likert-type scale, where -4 = *very strong disagreement* and +4 = *very strong agreement*.

5.3.2 Results and discussion. Forty six participants (69.7 %) rated the behaviour of Julie and Mark as wrong initially. The mean initial rating of the behaviour was, $M = 2.38$, $SD = 1.87$. Forty one participants, (62.12 %) rated the behaviour as wrong after viewing the counter-arguments and the critical slide. The mean revised rating of the behaviour was, $M = 2.82$, $SD = 1.91$. A paired samples t-test revealed a significant difference in rating from time one ($M = 2.38$, $SD = 1.87$), to time two ($M = 2.82$, $SD = 1.91$), $t(65) = -3.029$, $p = .004$. This result may be due to changes in the severity of the judgements as opposed to changing the judgement. Further analysis revealed that 12 participants changed the valence of their judgement: 7 participants changed their judgement from “wrong” to “neutral”; 1 participant changed their judgement from “wrong” to “right”; 1 participant changed their judgement from “neutral” to “right”; and 3 participants changed their judgement from “neutral” to “wrong”. A chi-squared test for independence revealed no significant association between time of judgement and valence of judgement made, $\chi^2(2, N = 66) = 0.85, p = .625, V = .11$.

5.3.2.1 Baseline rates of dumbfounding. Participants who selected the admission of not having reasons on the critical slide were identified as dumbfounded. Thirteen participants (19.7%) selected “It's wrong but I can't think of a reason”. Thirty three participants (50%) selected “It's wrong and I can provide a valid reason”; and 20 participants (30.3%) selected “There is nothing wrong”. Table 5.1 shows the responses to the critical slide for each condition.

Table 5.1: Rates of selecting each response to the critical slide in Study 6

	Cognitive load		Control	
Response to critical slide	N	percent	N	percent
There is nothing wrong.	15	45%	5	15%
It's wrong but I can't think of a reason.	6	18%	7	21%
It's wrong and I can provide a valid reason.	12	36%	21	64%

5.3.2.1.1 Dumbfounding and coded string responses. Participants who selected “It's wrong and I can provide a valid reason” were required to provide a reason. The reasons provided were coded for unsupported declarations or tautological reasons. An additional 12 participants were identified as dumbfounded following this coding. Three participants provided unsupported declarations (e.g., “Siblings cannot have sexual relationships with one another”), 4 participants provided tautological reasons (e.g., “Incest”), and 5 participants provided unsupported declarations accompanied by tautological reasons (e.g., “They r related. thats wrong”). Taking the coded string responses into account brought the total number of participants identified as dumbfounded to 25 (37.88%). In line with the limitations with taking unsupported declarations and tautological reasons as dumbfounded responses, identified in Chapter 4, the following analysis takes the selecting of an admission of not having a reason on the critical slide as the only measure of dumbfounding.

5.3.2.1.2 Dumbfounding and endorsing harm or norm principles. The exclusion criteria developed by Royzman et al., (2015) (endorsing of either the harm principle or the norm) were applied, and this resulted in a sample of 5 participants who were eligible for analysis. Just 1 of these fully convergent participants selected “It’s wrong but I can’t think of a reason”. No participants selected “It's wrong and I can provide a valid reason”; and 4 participants selected “There is nothing wrong”.

5.3.2.1.3 Dumbfounding and articulating, endorsing, and applying harm or norm principles. The revised exclusion criteria developed previously (articulating, endorsing, and applying of either the harm principle or the norm principle) were applied, and this resulted in a sample of forty participants who were eligible for analysis. Of these, 6 participants selected “It’s wrong but I can’t think of a reason”, 14 participants selected “It's wrong and I can provide a valid reason”, and 20 participants selected “There is nothing wrong”.

5.3.2.2 Cognitive load manipulation check. Initial check of responses to the memory task revealed that 7 participants (21.21%) successfully remembered the sequence of numbers and letters. Responses to the manipulation check question revealed that 5 participants (15.15%) found the memory task easy. Of these, 4 participants both found the task easy and got the answer right. The responses to the critical slide for these 4 were checked. One participant selected “There is nothing wrong”, 1 participant selected “It's wrong but I can't think of a reason”, and 2 participants selected “It's wrong and I can provide a valid reason”. Responses to the critical slide were spread evenly within this group of 4 participants, so these participants were not excluded from analysis. Further analysis revealed that all participants correctly remembered at least 2 digits in their correct location in the sequence. The mean number of correctly remembered digits was $M = 5.82$, $SD = 1.94$.

5.3.2.3 Cognitive load and engagement with the task. The cognitive load manipulation took place before the presenting of the vignette describing the behaviour to be judged. This allowed for the possibility that participants under cognitive load may not have engaged fully with the vignette when compared to the control group. An independent samples t-test revealed no significant difference in initial rating in the cognitive load group, ($M = 2.67$, $SD = 1.9$), and the control group, ($M = 2.55$, $SD = 1.92$), $t(63.9) = 1.256$, $p = .214$. An independent samples t-test revealed no significant difference in initial confidence in the cognitive load group, ($M = 5.42$, $SD = 1.3$), and the control group, ($M = 5.09$, $SD = 1.83$), $t(57.8) = 0.854$, $p = .396$. In view of this, we concluded that both groups engaged equally with the task.

5.3.2.4 Cognitive load and eligibility for analysis. In order to establish which measure of dumbfounding should be used in comparing the cognitive load and control groups, we investigated if there was any relationship between cognitive load and the exclusion criteria. Firstly, the final criteria for exclusion (articulating, endorsing, and applying harm and norm principles) were investigated. A chi-squared test for independence revealed a significant association between cognitive load and eligibility for analysis, $\chi^2(2, N = 66) = 9.44$, $p = .009$, $V = .38$, with 26 participants (78.79%) presenting as eligible for analysis in the cognitive load group, and only 14 participants (42.42%) presenting as eligible for analysis in the control group.

The criteria for exclusion required either, the articulating and endorsing and norm principle, or the articulating, endorsing and applying of the harm principle. The relationship between each of these elements and cognitive load was investigated separately.

5.3.2.4.1 Cognitive load and applying the harm principle. Three chi-squared tests for independence revealed no significant association between (a) cognitive load and applying the harm principle generally, $\chi^2(2, N = 66) = 1.787, p = .409, V = .13$; (b) cognitive load and applying the harm principle to boxing, $\chi^2(1, N = 66) = 0.135$, $p = .714$, $V = .05$; or (c) cognitive load and applying the harm principle to rugby, $\chi^2(1, N = 66) = 0.548, p = .459, V = .09$.

5.3.2.4.2 Cognitive load and endorsing and articulating the harm principle. Further chi-squared tests for independence also revealed no significant association between cognitive load and the endorsing of the harm principle,$\chi^2(1, N = 66) = 2.37$, $p = .124$; or, between cognitive load and the articulating of the harm principle, $\chi^2(1, N = 66) = < .001, p > .999$.

5.3.2.4.3 Cognitive load and endorsing and articulating the norm principle. A final series of chi-squared tests for independence revealed no significant association between cognitive load and the endorsing of the norm principle, $\chi^2(1, N = 66) < .001, p > .999$. However, a significant association between cognitive load and the articulating of the norm principle was found, $\chi^2(1, N = 66) = 6.061, p = .014, V = .30$, with 11 participants (33.33%) mentioning norms in the cognitive load group and with 22 participants (66.67%) mentioning norms in the control group. The apparent relationship between cognitive load and the articulation of norms can be seen in Figure 5.1.

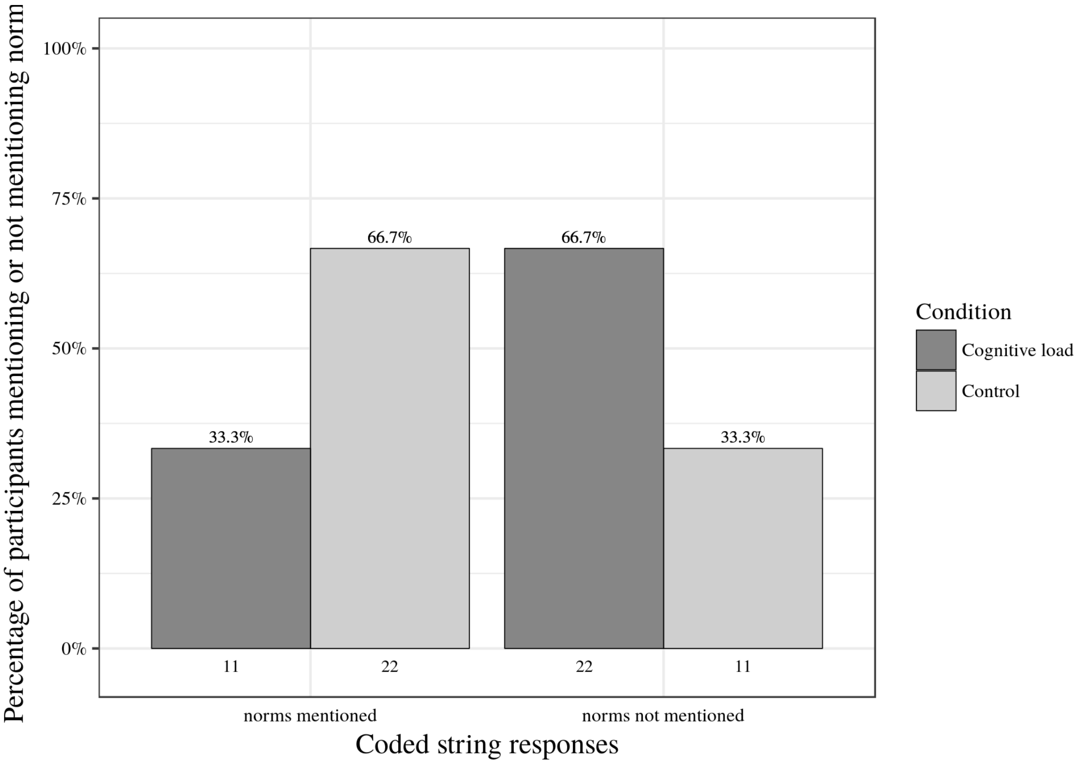

Figure 5.1: Study 6: Apparent relationship between cognitive load and mentioning norms

5.3.2.5 Cognitive load and responses to critical slide. The responses to the critical slide for the experiment group and the control group were analysed separately. In the group under cognitive load, 15 participants (45.45%) selected "There is nothing wrong", 6 participants (18.18%) selected "It's wrong but I can't think of a reason", and 12 participants (36.36%) selected "It's wrong and I can provide a valid reason". In the control group, 5 participants (15.15%) selected "There is nothing wrong", 7 participants (21.21%) selected "It's wrong but I can't think of a reason", and 21 participants (63.64%) selected "It's wrong and I can provide a valid reason". A chi-squared test for independence revealed a significant association between experimental condition and response to the critical slide, $\chi^2(2, N = 66) = 7.531$, $p = .023$, $V = .27$: under cognitive load more participants (15) selected "There is nothing wrong" than in the control group (5; see Table 5.2), the observed power was .69. Figure 5.2 shows the responses to the critical slide depending on cognitive load.

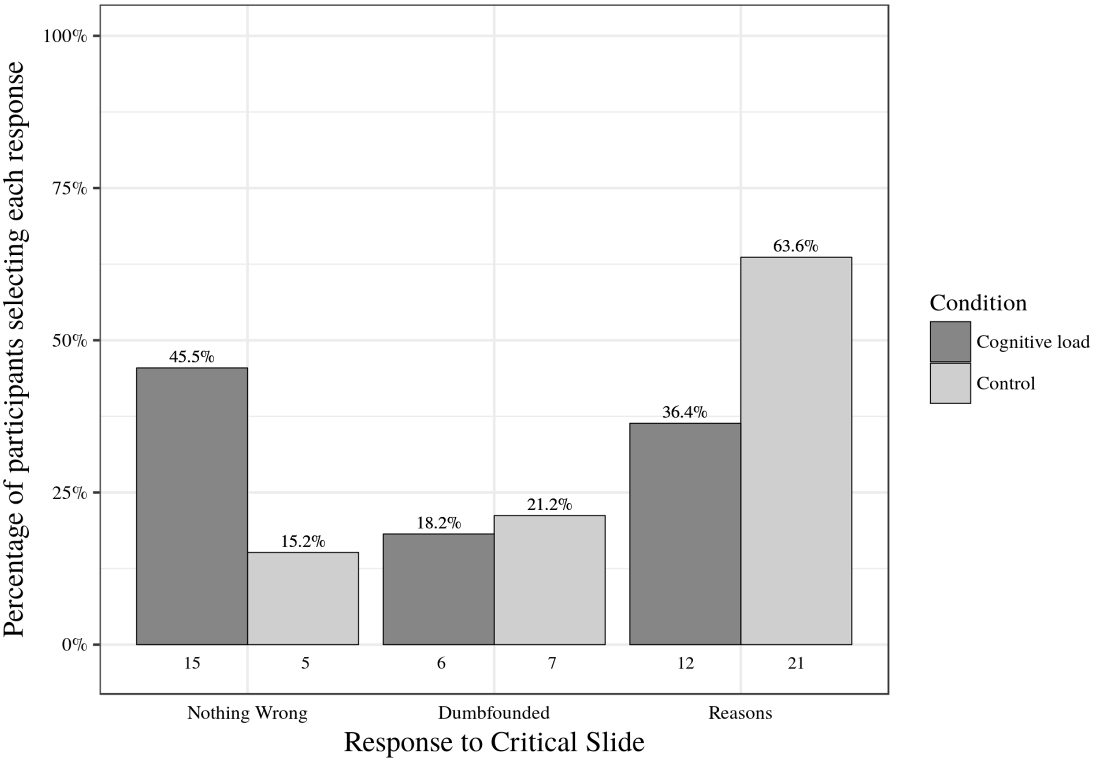

Figure 5.2: Study 6: Responses to critical slide and cognitive load

Table 5.2: Study 6 – Observed counts, expected counts, and standardised residuals for each response to the critical slide depending on cognitive load

Response		Cognitive Load	Control
Observed count	Nothing wrong	15	5
	Dumbfounded	6	7
	Reasons	12	21
Expected count	Nothing wrong	10	10
	Dumbfounded	6.5	6.5
	Reasons	16.5	16.5
Standardised residuals	Nothing wrong	2.68*	-2.68*
	Dumbfounded	-0.32	0.32
	Reasons	-2.22*	2.22*

*Note: * = sig. at p < .05 (|z| > 1.96); ** = sig. at p < .001 (|z| > 3.11)*

5.3.2.6 Cognitive load and judgements made. Twenty participants (60.61 %) rated the behaviour of Julie and Mark as wrong initially in the cognitive load group, and 26 participants (78.79 %) rated the behaviour as wrong in the control group. An independent samples t-test revealed no significant difference in initial rating in the cognitive load group, (M = 2.67, SD = 1.90), and the control group, (M = 2.55, SD = 1.92), t(63.9) = 1.256, p = .214. A chi-squared test for independence revealed no significant association between revised judgement and condition, χ^2(2, N = 66) = 3.166, p = .205, V = .22.

Eighteen participants (54.55 %) rated the behaviour of Julie and Mark as wrong after the critical slide in the cognitive load group, and 23 participants (78.79 %) rated the behaviour as wrong after the critical slide, in the control group. An independent samples t-test revealed no significant difference in revised rating between the cognitive load group, (M = 3.09, SD = 1.89), and the control group, (M = 2.55, SD = 1.92), t(63.99) = 1.161, p = .250. Similarly, a chi-squared test for independence revealed no significant association between revised judgement and condition, χ^2(2, N = 66) = 1.844, p = .398, V = .17.

5.3.2.7 Cognitive load and change in judgement. A paired samples t-test revealed no significant difference in rating from time one, (M = 2.09, SD = 1.83), to time two, (M = 2.55, SD = 1.92), t(32) = -1.844, p = .074 in the control group. However, in the cognitive load group, a significant difference in rating was found between time one (M = 2.67, SD = 1.90), and time two, (M = 3.09, SD = 1.89), t(32) = -2.701, p = .011.

As discussed previously, this difference may be due to changes in the severity of the judgements as opposed to changing the judgement. Further analysis revealed that 5 participants changed their judgement: 3 participants changed their judgement from “wrong” to “neutral”; one participant changed their judgement from “neutral” to “right”; 1 participant changed their judgement from “neutral” to “wrong”; crucially, no participants changed their judgement from “wrong” to “right”;. A chi-squared test for independence revealed no significant association between time of judgement and valence of judgement made, $\chi^2(2, N = 66) = 0.847, p = .654, V = .11$.

5.3.2.7.1 Changed Judgement, cognitive load, and critical slide. The revised judgements were binned to wrong versus not wrong. In the first analysis “neutral” judgements were included with the “not wrong” judgements. A chi-squared test for independence revealed no significant association between time of judgement and valence of judgement made, $\chi^2(1, N = 66) = 0.08, p = .773, V = .03$. A second analysis excluded “neutral” judgements and again a chi-squared test for independence revealed no significant association between time of judgement and valence of judgement made, $\chi^2(1, N = 66) = 0.15, p = .699, V = .05$. Interestingly, in the control group, 10 participants did not rate the behaviour as wrong, while only 5 participants selected “There is nothing wrong” on the critical slide. Seemingly, in the control group five participants changed their judgement between the critical slide and the revised judgement. Further analysis confirmed that 3 participants changed their judgement from a dumbfounded response to a judgement of “neutral”, 1 participant who provided reasons for their judging the behaviour as wrong proceeded to provide a neutral judgement, and 1 participant who provided a reason for their judgement proceeded to provide a judgement of “right”. Conversely, under cognitive load, 1 participant changed their judgement from a dumbfounded response to a neutral response, and 1 changed their judgement from “There is nothing wrong” to “wrong” between the critical slide and the final judgement.

5.3.2.8 Individual differences and providing reasons. The hypothesised relationship between Need for Cognition and responses to the critical slide was investigated. This analysis was exploratory. It was hypothesised that higher Need for Cognition would be related to higher rates of providing reasons, while lower Need for Cognition would be associated with dumbfounded responding. A multinomial logistical regression was conducted and no statistically significant association between Need for Cognition and response to the critical slide was found, $\chi^2(2, N = 66) = 4.86, p = .088$. The observed power was .49.

An independent samples t-test revealed no difference in Need for Cognition between and the cognitive load group, ($M = 5.42$, $SD = 1.08$), and the control group, ($M = 5.23$, $SD = 1.02$), $t(63.81) = 0.75$, $p = 0.456$. The observed power was .11.

5.3.3 Study 6 discussion. The aim of the Study 6 was to investigate if dumbfounded responding was influenced by cognitive load. Specifically, adopting a dual-systems model of moral judgement, it was hypothesised that cognitive load would lead to reduced levels of deliberative responding leading to a reduction in successfully identifying reasons for judgements. This may lead to (a) increased levels of dumbfounding, or (b) increased selecting of the "nothing wrong" response. Initial analysis found increased selecting of the "nothing wrong" response in the cognitive load group compared to the control group. Interestingly, at the point of the revised judgement, this difference was no longer present, however, the primary response of interest was the response to the critical slide which did show a difference between the cognitive load group and the control group. The individual difference variable Need for Cognition did not appear to be related to participants' susceptibility to moral dumbfounding.

It is hypothesised here that moral dumbfounding occurs when people fail to resolve a conflict between an initial intuition and an intuition that emerges following deliberation. Under cognitive load a higher proportion of participants failed to provide reasons for their judgements, and a higher proportion of participants selected the "There's nothing wrong" response. Under cognitive load, it may be that the persistent reminders of the inconsistency in reasoning (i.e. conflict), eventually leads people to change their judgement. There was no difference in rates of dumbfounding between the cognitive load group and the control group. It appears that when faced with an inability to justify a judgement with reasons, it is preferable to revise the judgement than to acknowledge the inconsistencies in the form of a dumbfounded response.

Taking the initial judgements and the revised judgements, there was no difference in valence of judgement from time one to time two. However, there was a difference in severity of judgement from time one to time two. This difference was present for the entire sample, however when the cognitive load group and the control group were analysed separately, it emerged that this difference between time one and time two, occurred only in the cognitive load group. It is possible that, with further questioning, the control group would eventually display significant differences in rating of the behaviour from the initial judgement. This claim is supported by the disappearing of any difference between the groups from the critical slide to the revised judgement (with no change in the cognitive load group). It is possible that persistent reminders of inconsistency in reasoning lead to the changing of judgements and that this occurs faster under cognitive load. The in-depth analysis of whether or not reasons were reported suggests that levels of dumbfounding would not be influenced by cognitive load, even if it was to be measured at an earlier stage.

One further finding was that the reasons provided as justifications for judgements appeared to be qualitatively different depending on the condition. Norms were mentioned more frequently in the control group than in the cognitive load group. There are various reasons for this that could be speculated, however, given the small sample size it is possible that occurred due to chance. This finding may prove interesting if it successfully replicates. Further qualitative analysis of the content of the open-ended responses will depend on successful replication of this finding.

5.4 Study 7: Dumbfounding and Cognitive Load 2 – Online Replication

Study 6 demonstrated interesting variability in responses to the critical slide depending on cognitive load. The aim of Study 7 is to assess the replicability of the results of Study 6, using an online sample. In Study 6, the experimenter was in the room with the participants. This made it more difficult for participants to cheat on the memory task. This is not possible with an online sample. An alternative cognitive load manipulation was taken from De Neys and Schaeken (2007), whereby a dot pattern is briefly presented to participants, and participants are required to reproduce the dot pattern at a later stage.

5.4.1 Method.

5.4.1.1 Participants and design. Study 7 was a between-subjects design with Need for Cognition additionally measured as a potential correlate and moderator variable. The dependent variable was response to the critical slide. The independent variable was cognitive load with two levels: high and low.

A total sample of 100 participants[10] (56 female, 44 male; M_{age} = 38.38, min = 19, max = 72, *SD* = 12.41) took part. Participants in this sample were recruited using Amazon's MTurk (Amazon Web Services Inc. 2016). Participants were paid $0.50 for their participation. Participants were recruited from English speaking countries or from countries where residents generally have a high level of English (e.g., The Netherlands, Denmark, Sweden).

5.4.1.2 Procedure and materials. Data were collected using an online questionnaire generated using Questback (Unipark 2013). Materials were largely the same as in Study 6, with a change to the cognitive load manipulation. Cognitive load was manipulated using a dot-pattern memory task (De Neys & Schaeken, 2007).

Participants were presented with a 3 x 3 grid containing a dot pattern. This image disappeared after one second. Participants then answered a question relating to the moral judgement task. Following the moral judgement question, participants were asked to reproduce the dot-pattern. In line with the manipulation employed by De Neys and Schaeken (2007), all participants took part in the memory task, and cognitive load was manipulated by varying the complexity of the patterns presented. The control group were presented with simple patterns, containing three dots in a line, while the experimental group were presented with more complex dot patterns containing 4 dots, see Figure 5.3.

[10] As in Study 6, a priori power analysis indicated that in order to detect a large effect size (*V* = .5) with 80% power, a sample of 39 participants was required. In order to detect a medium effect size (*V* = .3) with 80% power a sample of 107 participants was required.

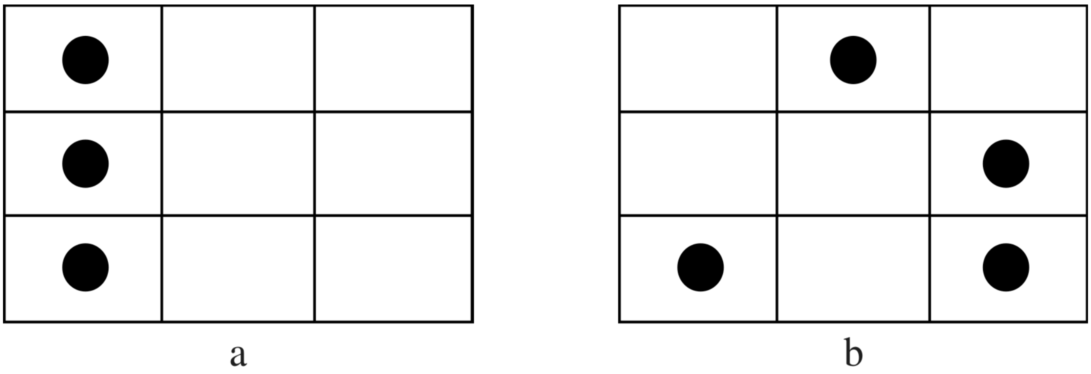

Figure 5.3: Sample dot patterns – higher complexity for the experimental condition (a) and more simple for the control group (b)

Participants were first presented with an information sheet and consent form. Following this, participants completed some questions relating to basic demographics. The two statements relating to the norm principle were then presented. Following this, participants took part in a practice memory task. The *Incest* vignette was presented after this practice memory task. There were four target questions during which participants were engaged in the memory task. A different pattern was presented before each of the following: the initial judgement, the initial opportunity to provide reasons, the critical slide, and the revised judgement. After each of these questions participants were required to reproduce the pattern.

As in Study 6, dumbfounding was measured using the critical slide. Following the critical slide and the revised judgement, the survey continued as in Study 6, with participants being asked to rate their confidence in their judgement, the degree to which they changed their mind, how difficult they found the memory task. The post-discussion questionnaire (Appendix C), taken from Haidt et al. (2000) was also administered, and the targeted questions relating to the endorsing and the applying of the harm principle were presented. Participants then completed the short (18 item) need for cognition scale (Cacioppo and Petty 1982; Petty, Cacioppo, and

Kao 1984).

5.4.2 Results and discussion. Seventy seven participants (77 %) rated the behaviour of Julie and Mark as wrong initially. The mean initial rating of the behaviour was, $M = 2.13$, $SD = 1.54$. Seventy participants, (70 %) rated the behaviour as wrong after viewing the counter-arguments and the critical slide. The mean revised rating of the behaviour was, $M = 2.35$, $SD = 1.65$. A paired samples t-test revealed a significant difference in rating from time one, ($M = 2.13$, $SD = 1.54$), to time two, ($M = 2.35$, $SD = 1.65$), $t(99) = -2.846$, $p = .005$. This result may be due to changes in the severity of the judgements as opposed to changing the judgement. Further analysis revealed that 10 participants changed the valence of their judgement: 6 participants changed their judgement from “wrong” to “neutral”; 1 participant changed their judgement from “wrong” to “right”; 2 participants changed their judgement from “neutral” to “right”; and 1 participant changed their judgement from “right” to “neutral”. A chi-squared test for independence revealed no significant association between time of judgement and valence of judgement made, $\chi^2(2, N = 100) = 1.26, p = .532, V = .11$.

5.4.2.1 Baseline rates of dumbfounding. Participants who selected the admission of not having reasons on the critical slide were identified as dumbfounded. Twenty six participants (26%) selected “It's wrong but I can't think of a reason”. Fifty participants (50%) selected “It's wrong and I can provide a valid reason”; and 24 participants (24%) selected “There is nothing wrong”. Table 5.3 shows the responses to the critical slide for each condition.

Table 5.3: Rates of selecting each response to the critical slide in Study 7

	Cognitive load		Control	
Response to critical slide	N	percent	N	percent
There is nothing wrong.	11	22%	13	27%
It's wrong but I can't think of a reason.	15	29%	11	22%

	Cognitive load		Control	
It's wrong and I can provide a valid reason.	25	49%	25	51%

5.4.2.1.1 Dumbfounding and coded string responses. Participants who selected “It's wrong and I can provide a valid reason” were required to provide a reason. The reasons provided were coded for unsupported declarations or tautological reasons. An additional 15 participants were identified as dumbfounded following this coding. Six participants provided unsupported declarations (e.g., “incest is always wrong”), 4 participants provided tautological reasons (e.g., “they are brother and sister”), 4 participants provided unsupported declarations accompanied by tautological reasons (e.g., “They are brother and sister. Wrong, wrong, wrong!”), and 1 participant admitting to not having reasons (“No valid reason”). Taking the coded string responses into account brought the total number of participants identified as dumbfounded to 41 (41%). As in Study 6, the remaining analysis used the stricter measure of dumbfounding, the selecting of an admission of having no reason only. The participant who admitted to having no reasons is also identified as dumbfounded for the remaining analysis.

Table 5.4: Rates of dumbfounded responding following analysis of the string responses

	Cognitive load		Control	
Response to critical slide	N	percent	N	percent
There is nothing wrong.	11	22%	13	27%
It's wrong but I can't think of a reason.	16	31%	11	22%
It's wrong and I can provide a valid reason.	24	47%	25	51%

5.4.2.1.2 Dumbfounding and endorsing harm or norm principles. The exclusion criteria developed by Royzman et al., (2015) (endorsing of either the harm principle or the norm) were applied, and this resulted in a sample of 13 participants who were eligible for analysis. None of these fully convergent participants selected

"It's wrong but I can't think of a reason". Two participants selected "It's wrong and I can provide a valid reason"; and 11 participants selected "There is nothing wrong".

5.4.2.1.3 Dumbfounding and articulating, endorsing, and applying harm or norm principles. The revised exclusion criteria developed previously (articulating, endorsing, and applying of either the harm principle or the norm) were applied, and this resulted in a sample of 61 participants who were eligible for analysis. Of these, 16 participants selected "It's wrong but I can't think of a reason", 23 participants selected "It's wrong and I can provide a valid reason"; and 22 participants selected "There is nothing wrong".

5.4.2.2 Cognitive load and responses to critical slide. The responses to the critical slide for the experiment group and the control group were analysed separately. In the group under cognitive load, 11 participants (21.57%) selected "There is nothing wrong", 16 participants (31.37%) selected "It's wrong but I can't think of a reason", and 24 participants (47.06%) selected "It's wrong and I can provide a valid reason". In the control group, 13 participants (26.53%) selected "There is nothing wrong", 11 participants (22.45%) selected "It's wrong but I can't think of a reason", and 25 participants (51.02%) selected "It's wrong and I can provide a valid reason". A chi-squared test for independence revealed no association between experimental condition and response to the critical slide, $\chi^2(2, N = 100) = 1.07$, $p = .585$, $V = .10$. The observed power was .14. Figure 5.4 shows the responses to the critical slide (with the admission provided also identified as a dumbfounded response) depending on cognitive load.

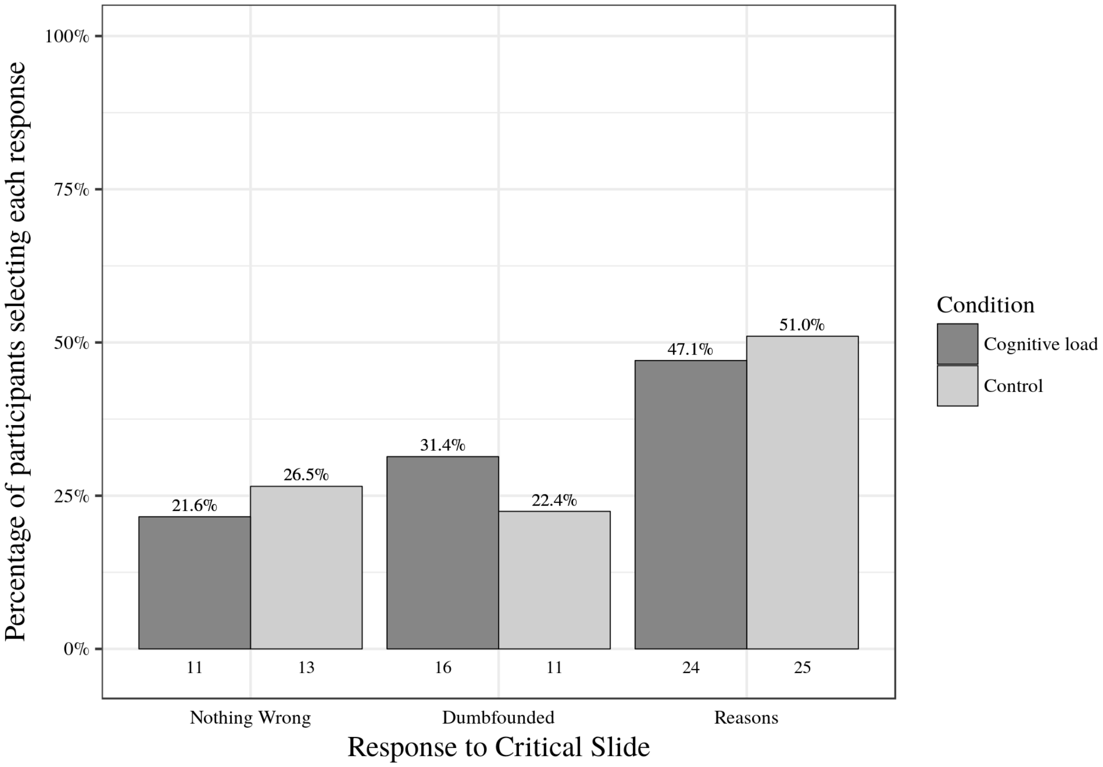

Figure 5.4: Study 7: Responses to critical slide and Cognitive Load (including admissions provided)

5.4.2.3 Engagement with the memory task. The aim of Study 7 was to replicate Study 6 with an online sample. An alternative cognitive load manipulation was employed in order to minimise the risk that participants could cheat with the memory task. This cognitive load manipulation was applied in the same manner as in De Neys and Schaeken (2007). Participants in both groups took part in a memory task. However, using this method, no differences between the cognitive load group and the control group were observed. Recall that in Study 6, the control group did not engage in any task. It is possible that simply engaging in a memory task led to differences in responses, and that level of difficulty (the manipulation that was employed) was irrelevant. Indeed, the responding to the critical slide in the control group in Study 7 is more similar to the responding in the experimental group in

Study 6 than to the control group in Study 6.

The participants in the De Neys and Schaeken (2007) study completed the study in small groups. It is not stated whether or not participants were supervised in their participation, however, given that they were undergraduate students taking part for course credit, and that they completed it in small groups, it is likely that there was some element of oversight or supervision. The participants in Study 7, recruited through MTurk, were unsupervised in their completion of the study. De Neys and Schaeken (2007) report high rates of successful reproducing of the dot patterns (with 94% of the complex patterns and 97% of the simple patterns being reproduced correctly De Neys & Schaeken, 2007, p. 130). Rates of successful reproduction of the dot patterns in Study 7 were much lower, with 25% of the complex patterns and 65% of the simple patterns that accompanied the critical slide being reproduced correctly. From this it is apparent that engagement with the memory task in Study 7 was much lower than in De Neys and Schaeken (2007). Failure to engage with a task that is intended as a cognitive load manipulation doubtless reduces the effectiveness of the task in manipulating cognitive load. In view of this, participants' engagement with the memory task was investigated, and participants deemed not to have engaged with the task were compared against participants who did engage with the task.

During the recall phase of the memory task, participants were presented with a 3 x 3 grid and asked to check the correct spaces where the dots were placed, to reproduce the pattern. For the scoring of this task, each of the nine places in the grid could be marked/not marked correctly or incorrectly, making 9 the total possible number of correct responses. If a person misplaced one dot in the pattern this would count for 2 incorrect places in the grid: the mark in the incorrect place, and the absence of a mark in the place it should have been. A participant who received a score of 7, could reasonably be taken to have engaged with the task, and simply made a slip. As such, participants who scored 7 or greater were deemed to have engaged with the memory task, and participants who scored below 7 were deemed to have not engaged with the task. Taking the responses for the memory task, for the critical slide, across both conditions, 56 participants were identified as engaging with the memory task, and 44 were identified as not engaging with the task.

5.4.2.4 Engagement with the memory task and eligibility for analysis. Having identified an alternative measure of cognitive load, controlling for engagement with the memory task, a series of tests were conducted to assess if this new measure was related to eligibility for analysis, as measured by the criteria developed in Chapter 4. Firstly, unlike Study 6, a chi-squared test for independence revealed no significant association between engagement with the memory task and overall eligibility for analysis, $\chi^2(1, N = 100) = 1.207$, $p = .272$, $V = .11$, with 31 participants (55.36%) presenting as eligible for analysis in the engaged group, and 30 participants (68.18%) presenting as eligible for analysis in the not engaged.

5.4.2.4.1 Cognitive load and applying the harm principle. Three chi-squared tests for independence revealed no significant association between (a) engagement with the memory task and applying the harm principle generally, $\chi^2(2, N = 100) = 1.345$, $p = .510$, $V = .12$; (b) engagement with the memory task and applying the harm principle to boxing, $\chi^2(1, N = 100) = 1.61$, $p = .204$, $V = .13$; or (c) engagement with the memory task and applying the harm principle to rugby, χ^2 $(1, N = 100) = 0.08$, $p = .782$, $V = .03$.

5.4.2.4.2 Cognitive load and endorsing and articulating the harm principle. Two chi-squared tests for independence also revealed no significant association between engagement with the memory task and the endorsing of the harm principle, $\chi^2(1, N = 100) = 0.25, p = .617, V = .02$; or, between engagement with the memory task and the articulating of the harm principle, χ^2 1, $N = 100) = 0.299, p = .584, V = .$ 08.

5.4.2.4.3 Cognitive load and endorsing and articulating the norm principle. A final series of chi-squared tests for independence revealed no significant association between cognitive load and the endorsing of the norm principle, $\chi^2(1, N = 100) = 0.143, p = .705, V = .04$, or the articulating of the norm principle, $\chi^2(1, N = 100) = 0.546, p = .460, V = .07$. This suggests that the apparent relationship between cognitive load and articulating the norm principle observed in Study 6 may have occurred due to chance.

5.4.2.5 Dumbfounding and engagement with the task. The responses to the critical slide for participants who engaged with the memory task and participants who did not engage with the memory task were analysed separately. In the group who engaged, 13 participants (23.21%) selected “There is nothing wrong”, 21 participants (37.50%) selected “It's wrong but I can't think of a reason”, and 22 participants (39.29%) selected “It's wrong and I can provide a valid reason”. In the group who did not engage with the memory task, 11 participants (19.64%) selected “There is nothing wrong”, 6 participants (10.71%) selected “It's wrong but I can't think of a reason”, and 27 participants (48.21%) selected “It's wrong and I can provide a valid reason”. A chi-squared test for independence revealed an association between engagement in the memory task and response to the critical slide, $\chi^2(2, N = 100) = 7.68, p = .021, V = .28$. The observed power was .70. Figure 5.5 shows the variation in responses to the critical slide depending on engagement with the memory task. Table 5.5 shows the observed counts, expected counts, and standardised residuals.

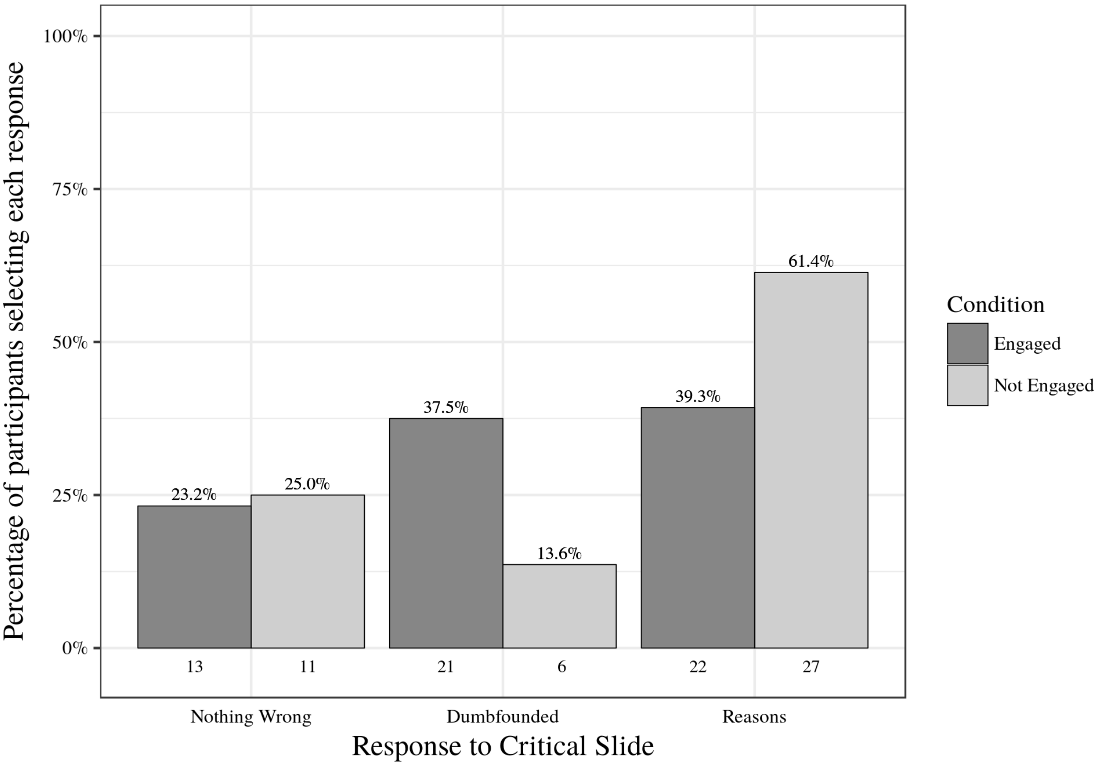

Figure 5.5: Study 7: Responses to critical slide and Engagement with memory task

Table 5.5: Study 7 – Observed counts, expected counts, and standardised residuals for each response to the critical slide depending on cognitive load

Response		Cognitive Load / (engaged)	Control / (not-engaged)
Observed count	Nothing wrong	13	11
	Dumbfounded	21	6
	Reasons	22	27
Expected count	Nothing wrong	13.44	10.56
	Dumbfounded	15.12	11.88
	Reasons	27.44	21.56
Standardised residuals	Nothing wrong	-0.21	0.21
	Dumbfounded	2.67*	-2.67*
	Reasons	-2.19*	2.19*

*Note: * = sig. at p < .05 (|z| > 1.96); ** = sig. at p < .001 (|z| > 3.11)*

When engagement with the memory task was accounted for the rates of providing reasons dropped significantly. This drop in rates of providing reasons is consistent with the prediction that a cognitive load manipulation would hinder the identification of reasons. Unlike Study 6, the reduction in providing reasons resulted in higher rates of dumbfounding as opposed to higher rates of selecting "There is nothing wrong". This may provide suggestive evidence in support of the hypothesis that changing a judgement requires more deliberation than providing a dumbfounded response. All participants in Study 7 identified as under cognitive load in this study were identified as having engaged with the task, whereas in Study 6 participants who received the cognitive load manipulation were identified as under cognitive load, without assessing their engagement. This may suggest that deliberation was more inhibited for participants identified as having engaged with the memory task in Study 7 than participants in the manipulation group in Study 6. This is speculative but may provide a useful question for future research.

5.4.2.6 Individual differences and providing reasons. As in Study 6, the possibility of a relationship between Need for Cognition and susceptibility to dumbfounded responding was investigated. Again, this analysis was primarily exploratory. It was hypothesised that higher Need for Cognition scores would be predict the providing of reasons, and lower scores would predict dumbfounded responding. A multinomial logistical regression was conducted and no statistically significant association between Need for Cognition and response to the critical slide was found, $\chi^2(2, N = 100) = 2.19, p = .334$. The observed power was .24.

An independent samples t-test revealed no difference in Need for Cognition between the cognitive load group, ($M = 5.33$, $SD = 1.47$), and the control group, ($M = 5.79$, $SD = 1.69$), $t(95.03) = -1.449, p = 0.151$. The observed power was .30.

5.4.3 Study 7 discussion. The aim of Study 7 was to replicate Study 6 with an online sample. Initial analysis indicated that this was unsuccessful. However, specific limitations with the experimental manipulation were identified. In particular, the rates of correct responding to the memory task were considerably lower than expected based on previous research (De Neys & Schaeken, 2007). In view of this, level of engagement with the memory task was used as an alternative IV. Participants who engaged with the memory task (as measured by scoring 7/9 or higher) were compared against participants who did not engage with the memory task.

Interestingly, responses to the critical slide varied depending on engagement with the memory task. As expected, the rates of providing reasons was significantly lower for participants deemed to be under cognitive load (engaged) than for participants who did not engage fully with the memory task. Rates of dumbfounding were higher for engaged participants than for participants who did not engage with the memory task, and the rates of selecting “There is nothing wrong” were similar. This finding is different from what was observed in Study 6, whereby a decrease in providing reasons led to an increase in selecting “There is nothing wrong”. This finding may point to a difference in the level of deliberation associated dumbfounded responding and changing judgements (with changing judgements requiring more deliberation), though this is currently speculation. One prediction of this claim is that participants who provided a dumbfounded response in Study 6 correctly remembered more digits than participants who selected “There is nothing wrong”. Indeed, the mean number of correctly remembered digits for dumbfounded participants was, $M = 6.17$, $SD = 2.32$, while the mean number of correctly remembered digits for participants who selected “There is nothing wrong was, $M = 5.47$, $SD = 2.00$. An independent samples t-test revealed that this difference was not significant, $t(8.15) = -0.65$, $p = 0.534$.

The stated aim of Study 7 was to replicate Study 6. As a replication it failed. However, interesting variability was observed when engagement with the memory task was accounted for. Two problems with the cognitive load manipulation were identified. Firstly, the control (a simpler memory task) did not serve as an appropriate control. For future studies, the control group should not engage in any memory task. Secondly, (and unsurprisingly) the manipulation is only effective if participants engage with the memory task. Given that supervision of participants is impractical when collecting data using MTurk, the level of engagement with the memory task may be determined by assessing the rates of correct responding.

5.5 Study 8: Dumbfounding and Cognitive Load 3 – Revised Online Replication

Study 7 failed to replicate the findings of Study 6. However, the role of engagement with the memory task emerged as an important moderator in the effectiveness of the cognitive load manipulation. Study 8 was conducted in order to test if cognitive load affects participants' ability to identify reasons for their judgements, when account for engagement with the memory task.

5.5.1 Method.

5.5.1.1 Participants and design. Study 8 was a between-subjects design with Need for Cognition additionally measured as a potential correlate and moderator variable. The dependent variable was response to the critical slide. The independent variable was cognitive load with two levels: present and absent.

A total sample of 163 participants[11] (92 female, 71 male; M_{age} = 40.07, min = 20, max = 72, *SD* = 13.11) took part. Participants in this sample were recruited through MTurk. Participation was voluntary and participants were paid $US 0.50 for their participation. Participants were recruited from English speaking countries or from countries where residents generally have a high level of English (e.g., The Netherlands, Denmark, Sweden).

5.5.1.2 Procedure and materials. Study 8 was the same as Study 7 with two changes. The control group did not take part in a memory task. The selection of dot patterns presented to participants in the control group was changed following piloting. In order to avoid task fatigue, the dot patterns presented, alternated between the easy 3-dot patterns and the more complex 4-dot patterns. The study was set up such that participants were always presented with a complex 4-dot pattern ahead of the critical slide.

Following the finding in Study 7 that some participants do not appear to engage with the memory task, and that engagement with the task appeared to moderate the effect the task had on responses to the critical slide, engagement with the task was measured as a score of 7 or higher on the memory task that accompanied the critical slide. Only participants who engaged with the task were eligible for analysis. For this reason, more participants took part in the experimental condition than took part in the control condition, in order to ensure that the final groups being analysed were of similar size, following the exclusion of participants who did not engage with the memory task. Other than the two changes described

[11] As in Studies 6 and 7, a priori power analysis indicated that in order to detect a large effect size (*V* = .5) with 80% power, a sample of 39 participants was required. In order to detect a medium effect size (*V* = .3) with 80% power a sample of 107 participants was required.

above, Study 8 was the same as Study 7.

5.5.2 Results and discussion. Sixty eight of the 102 (66.67%) participants in the experimental condition engaged with the memory task, scoring 7 or higher on for the task that accompanied the critical slide. Thirty four participants identified as not engaging with the task were excluded from analysis.

A final sample of 129 participants (74 female, 55 male; M_{age} = 40.26, min = 20, max = 72, *SD* = 13.04) was studied. Sixty eight participants in the cognitive load condition were compared against 61 participants in the control group. Ninety five participants (73.64%) rated the behaviour of Julie and Mark as wrong initially. The mean initial rating of the behaviour was, *M* = 2.27, *SD* = 1.75. Ninety four participants, (72.87%) rated the behaviour as wrong after viewing the counter-arguments and the critical slide. The mean revised rating of the behaviour was, *M* = 2.35, *SD* = 1.74. A paired samples t-test revealed no significant difference in rating from time one, (*M* = 2.27, *SD* = 1.75), to time two, (*M* = 2.35, *SD* = 1.74), *t*(128) = -1.148, *p* = .253. Similarly, a chi-squared test for independence revealed no significant association between time of judgement and valence of judgement made, χ^2(2, *N* = 129) = 0.633, *p* = .729, *V* = .07.

5.5.2.1 Baseline rates of dumbfounding. Participants who selected the admission of not having reasons on the critical slide were identified as dumbfounded. Twenty two participants (17.05%) selected “It's wrong but I can't think of a reason”. Seventy seven participants (59.69%) selected “It's wrong and I can provide a valid reason”; and 30 participants (23.26%) selected “There is nothing wrong”. Table 5.6 shows the frequency of each response on the critical slide depending on condition.

Table 5.6: Rates of selecting each response to the critical slide in Study 8

	Cognitive load		Control	
Response to critical slide	N	percent	N	percent
There is nothing wrong.	15	22%	15	25%
It's wrong but I can't think of a reason.	17	25%	5	8%
It's wrong and I can provide a valid reason.	36	53%	41	67%

5.5.2.2 Dumbfounding and coded string responses. Participants who selected “It's wrong and I can provide a valid reason” were required to provide a reason. The reasons provided were coded for unsupported declarations or tautological reasons. An additional 14 participants were identified as dumbfounded following this coding. Nine participants provided unsupported declarations (e.g., “It is morally wrong for a brother and sister to make Love.”), 3 participants provided tautological reasons (e.g., “It is incest.”), and 2 participants provided unsupported declarations accompanied by tautological reasons (e.g., “they are bother and sister that is wrong”). Taking the coded string responses into account brought the total number of participants identified as dumbfounded to 36 (27.91%). The number of additional participants identified as dumbfounded following the coding of the open-ended responses was the same in each condition (7 in the control group, and 7 in the cognitive load group). In view of this, and following from Studies 6 and 7, the remaining analysis used the stricter measure of dumbfounding, the selecting of an admission of having no reason only.

5.5.2.3 Cognitive load and eligibility for analysis. As in Study 7, a series of tests was conducted to assess if there was a relationship between cognitive load and eligibility for analysis as measured by the criteria developed in Chapter 4. A chi-squared test for independence revealed no significant association between cognitive load and overall eligibility for analysis, $\chi^2(1, N = 129) = 0.025, p = .874, V = .01$.

5.5.2.3.1 Cognitive load and applying the harm principle. Three chi-squared tests for independence revealed no significant association between (a) cognitive load and applying the harm principle generally, $\chi^2(2, N = 129) = 1.668, p = .434, V = .06$; (b) cognitive load and applying the harm principle to boxing, $\chi^2(1, N = 129) < .001$, $p > .999$; or (c) cognitive load and applying the harm principle to rugby, $\chi^2(1, N = 129) = 0.023, p = .880, V = .01$.

5.5.2.3.2 Cognitive load and endorsing and articulating the harm principle. Two chi-squared tests for independence also revealed no significant association between cognitive load and the endorsing of the harm principle, $\chi^2(1, N = 129) = 0.556, p = .456, V = .07$; or, between cognitive load and the articulating of the harm principle, $\chi^2(1, N = 129) < .001, p > .999$.

5.5.2.3.3 Cognitive load and endorsing and articulating the norm principle. A final series of chi-squared tests for independence revealed no significant association between cognitive load and the endorsing of the norm principle, $\chi^2(1, N = 129) = 0.122, p = .727, V = .03$, or the articulating of the norm principle, $\chi^2(1, N = 129) = 0.764, p = .382, V = .08$. Again, the variation in articulating the norm principle with cognitive load observed in Study 6 was not present, indicating that it was more than likely due to chance.

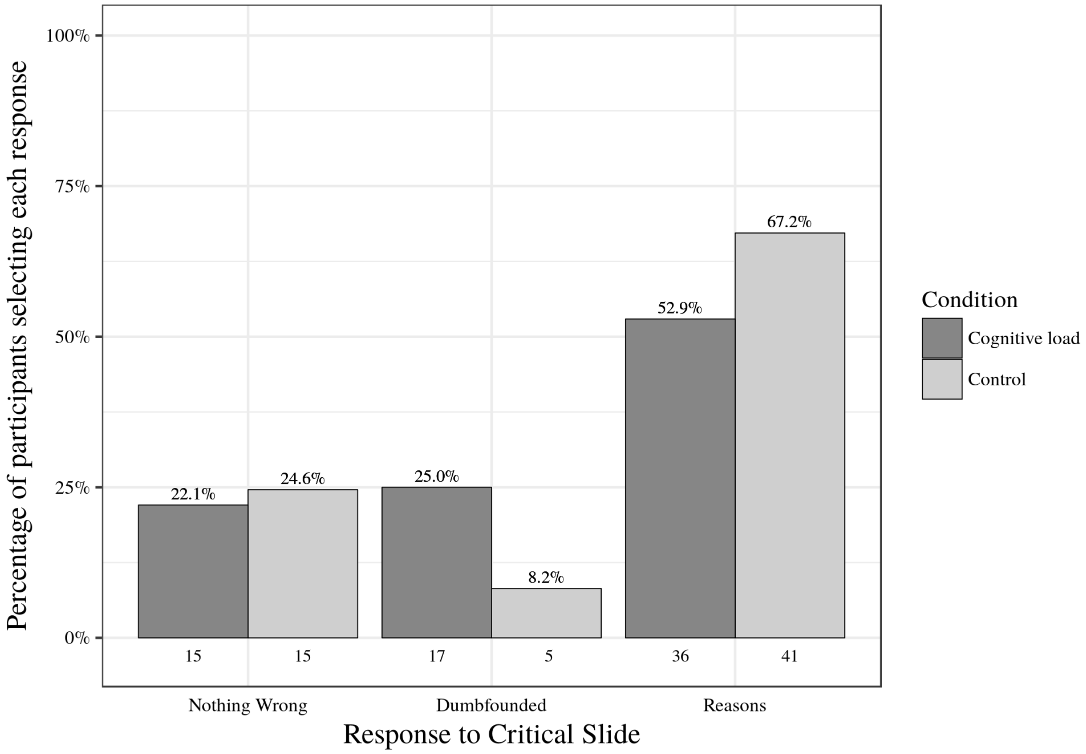

Figure 5.6: Study 8: Responses to critical slide and Cognitive Load

5.5.2.4 Cognitive load and responses to critical slide. The responses to the critical slide for the experiment group and the control group were analysed separately. In the group under cognitive load, 15 participants (22.06%) selected "There is nothing wrong", 17 participants (25%) selected "It's wrong but I can't think of a reason", and 36 participants (52.94%) selected "It's wrong and I can provide a valid reason". In the control group, 15 participants (24.59%) selected "There is nothing wrong", 5 participants (8.20%) selected "It's wrong but I can't think of a reason", and 41 participants (67.21%) selected "It's wrong and I can provide a valid reason". A chi-squared test for independence revealed a significant association between experimental condition and response to the critical slide, $\chi^2(2, N = 129) = 6.51$, $p = .039$, $V = .223$. The observed power was .63. Figure 5.6 shows the responses to the critical slide depending on cognitive load. Table 5.7 shows the observed counts, expected counts, and standardised residuals for each response depending on cognitive load.

Table 5.7: Study 8 – Observed counts, expected counts, and standardised residuals for each response to the critical slide depending on cognitive load

Response		Cognitive Load / (engaged)	Control / (not-engaged)
Observed count	Nothing wrong	15	15
	Dumbfounded	17	5
	Reasons	36	41
Expected count	Nothing wrong	15.81	14.19
	Dumbfounded	11.60	10.40
	Reasons	60.59	36.41
Standardised residuals	Nothing wrong	-0.34	0.34
	Dumbfounded	2.53*	-2.53*
	Reasons	-1.65	1.65

*Note: * = sig. at p < .05 (|z| > 1.96); ** = sig. at p < .001 (|z| > 3.11)*

The degree to which cognitive load inhibited the identification of reasons in this study is not as convincing as in Study 6 or Study 7 (when engagement with the memory task was accounted for). That said, a significant difference between participants in the control group and participants in the cognitive load (and engaged) group was observed. As in Study 7, rates of dumbfounded responding were significantly higher in the cognitive load group than in the control group. Again, this was different from Study 6 whereby reduced identification of reasons led to higher rates of selecting “There is nothing wrong”. One possible explanation for this, identified in Study 7, is that changing a judgement requires more deliberation than providing a dumbfounded response. By only including participants that clearly engaged in the manipulation task, it is possible that the participants who experienced inhibited deliberative responding experienced it to a greater degree than the participants in Study 6. Again this is speculation, and the differences involved are likely so small, and susceptible to other factors (e.g., individual differences) that testing this would be particularly problematic.

5.5.2.5 Individual differences and providing reasons. An exploratory analysis investigated possibility of a relationship between Need for Cognition and ability to provide reasons for a judgement. It was hypothesised that higher Need for Cognition scores would be predict the providing of reasons, and lower scores would predict dumbfounded responding. A multinomial logistical regression was conducted and a statistically significant association between Need for Cognition and response to the critical slide was found, $\chi^2(2, N = 129) = 6.43, p = .040$. The observed power was .61. Need for Cognition explained between 4.9% (Cox and Snell R square) and 5.8% (Nadelkerke R squared) of the variance in responses to the critical slide. As Need for Cognition increased, participants were significantly more likely to provide reasons than to present as dumbfounded, Wald = 6.08, $p = .014$, odds ratio = 1.46, 95% CI [1.08, 1.97]. There was no significant relationship observed between Need for Cognition and selection “There is nothing wrong” when compared to dumbfounded responding, Wald = 2.71, $p = .100$, odds ratio = 1.33, 95% CI [.95, 1.88].

An independent samples t-test revealed no difference in Need for Cognition between cognitive load group, ($M = 5.92$, $SD = 1.55$), and the control group, ($M = 5.91$, $SD = 1.79$), $t(110.63) = 0.032, p = 0.975$.

5.5.3 Study 8 discussion. Study 8 demonstrated the predicted relationship between engagement with a cognitive load task and providing reasons for a judgement. As expected, engagement with a cognitive load task reduced the rates of providing reasons. Consistent with what was observed in the revised analysis in

Study 7, the reduction in providing reasons led to an increase in dumbfounded responding. Speculative reasons for the increase in dumbfounded responding as opposed to selecting "There is nothing wrong" have been outlined, however it is not possible to test these with the current materials. Need for Cognition did not appear to be related to the judgements made, or with responses to the critical slide. However, participants who consistently stated that they had reasons for their judgement scored significantly higher on Need for Cognition than participants who indicated at least once that they may not have a reason for their judgement. This is consistent with what was observed in Study 6 (though not in Study 7). It appears that Need for Cognition may play some role in the providing of reasons, whereby people who score higher in Need for Cognition are either better at identifying reasons or more motivated to provide reasons, possibly more motivated to believe that reasons for the judgement exist.

Study 8 utilised a cognitive load manipulation that was developed for use with online samples following Study 7 whereby engagement with the cognitive load manipulation task is accounted for. The various strengths and weaknesses of the use of online sample and recruitment using MTurk have been identified elsewhere (Crump, McDonnell, & Gureckis, 2013; Goodman, Cryder, & Cheema, 2013). In particular Rand (2012) has identified complex experimental manipulations, with specific reference to cognitive load manipulations, as impractical for use on MTurk. The manipulation used in Study 8, controlling for engagement with the task, appears to have been successful, however there was no objective manipulation check used. As such a follow up study was conducted that included an objective manipulation check.

5.6 Study 9: Dumbfounding and Cognitive Load 4 – Online Replication with Manipulation Check

Significant challenges in manipulating cognitive load with online samples have been identified in Studies 7 and 8. Controlling for engagement in the manipulation task appeared to allay these challenges to some degree, however neither Study 7 nor Study 8 included an objective manipulation check. Study 9 was conducted to address this limitation by replicating Study 8 with the inclusion of a manipulation check.

5.6.1 Method.

5.6.1.1 Participants and design. Study 9 was a between-subjects design with Need for Cognition additionally measured as a potential correlate and moderator variable. The dependent variable was response to the critical slide. The independent variable was cognitive load with two levels: present and absent.

A total sample of 156 participant[12] (98 female, 58 male; $M_{age} = 41.41$, min = 21, max = 76, $SD = 13.74$) took part. Participants in this sample were recruited through MTurk. Participation was voluntary and participants were paid 0.50 US dollars for their participation. Participants were recruited from English speaking countries or from countries where residents generally have a high level of English (e.g., The Netherlands, Denmark, Sweden).

5.6.1.2 Procedure and materials. The materials for Study 9 were largely the same as for Study 8, with the inclusion of a manipulation check. To recap, the order of events is as follows (1) a general question relating to the norm principle,

[12] As in Studies 6-8, a priori power analysis indicated that in order to detect a large effect size ($V = .5$) with 80% power, a sample of 39 participants was required. In order to detect a medium effect size ($V = .3$) with 80% power a sample of 107 participants was required.

participants select the statement they agree with most; (2) presenting of the Julie and Mark vignette; (3) initial judgement; (4) open-ended providing of reasons; (5) series of 3 counter-arguments, for each counter-argument participants: (i) agree/disagree with each counter-argument, (ii) rate the behaviour again, and (iii) indicate whether or not they have reasons for their judgement; (6) the critical slide (measure of dumbfounding); (7) participant rate the behaviour again (revised judgement).

A prose paragraph was included after participants made their revised judgements. Participants were then asked three comprehension questions relating to the prose paragraph. It was expected that participants in the control group would perform better at this task than participants under cognitive load (Just & Carpenter, 1992; Kahneman, 1973).

5.6.2 Results and discussion. Sixty four of the 93 (68.82%) participants in the experimental condition engaged with the memory task, scoring 7 or higher on the task that accompanied the critical slide. Twenty nine participants identified as not engaging with the task were excluded from analysis.

A final sample of 127 participants (84 female, 43 male; M_{age} = 41.19, min = 21, max = 74, *SD* = 13.91) was studied. Sixty four participants in the cognitive load condition were compared against 63 participants in the control group. Ninety eight participants (77.17%) rated the behaviour of Julie and Mark as wrong initially. The mean initial rating of the behaviour was, *M* = 2.09, *SD* = 1.62. Ninety two participants, (72.44 %) rated the behaviour as wrong after viewing the counter-arguments and the critical slide. The mean revised rating of the behaviour was, *M* = 2.31, *SD* = 1.79. A paired samples t-test revealed a significant difference in rating from time one, (*M* = 2.09, *SD* = 1.62), to time two, (*M* = 2.31, *SD* = 1.79), *t*(126) = -3.142 , *p* = .002. As in previous studies, this may reflect to changes in the severity of the judgements as opposed to changing the judgement. Further analysis revealed that 13 participants changed the valence of their judgement: 6 participants changed their judgement from “wrong” to “neutral”; 1 participant changed their judgement from “wrong” to “right”; 4 participant changed their judgement from “neutral” to “right”; 1 participant changed their judgement from “right” to “neutral”; and 1 participant changed their judgement from “right” to “wrong”. A chi-squared test for independence revealed no significant association between time of judgement and valence of judgement made, $\chi^2(2, N = 127) = 1.281, p = .973, V = .10$.

5.6.2.1 Baseline rates of dumbfounding. Participants who selected the admission of not having reasons on the critical slide were identified as dumbfounded. Twenty participants (15.75%) selected “It's wrong but I can't think of a reason”. Seventy six participants (59.84%) selected “It's wrong and I can provide a valid reason”; and 31 participants (24.41%) selected “There is nothing wrong”. Table 5.8 shows the frequency of each response on the critical slide depending on condition.

Table 5.8: Rates of selecting each response to the critical slide in Study 9

	Cognitive load		Control	
Response to critical slide	N	percent	N	percent
There is nothing wrong.	19	30%	12	19%
It's wrong but I can't think of a reason.	10	16%	10	16%
It's wrong and I can provide a valid reason.	35	55%	41	65%

5.6.2.2 Dumbfounding and coded string responses. Participants who selected “It's wrong and I can provide a valid reason” were required to provide a reason. The reasons provided were coded for unsupported declarations or tautological reasons. An additional 19 participants were identified as dumbfounded following this coding. Eleven participants provided unsupported declarations (e.g., “incest is wrong”), 3 participants provided tautological reasons (e.g., “They are blood brother and sister”), and 4 participants provided unsupported declarations accompanied by tautological reasons (e.g., “it's wrong because it's incest! You don't have sex with family members, you just don't!”). Taking the coded string responses into account brought the total number of participants identified as dumbfounded to 34 (26.77%). As in the previous studies, the remaining analysis used the stricter measure of dumbfounding, the selecting of an admission of having no reason only.

5.6.2.3 Cognitive load manipulation check. Study 9 included a manipulation check to assess whether the cognitive load manipulation employed was successful. Participants read a short vignette and answered questions relating to the content of the vignette they read. There was no difference in the number of correct answers to these questions between the cognitive load group and the control group $F(1, 124) = .33$ $p = .569$, partial $\eta^2 = .003$. There was also no difference in time taken to read the vignette between the groups $F(1, 125) = 2.57$ $p = .112$, partial $\eta^2 = .020$.

5.6.2.4 Cognitive load and eligibility for analysis. As in previous studies, a series of tests was conducted to assess if there was a relationship between cognitive load and eligibility for analysis. A chi-squared test for independence revealed no significant association between cognitive load and overall eligibility for analysis, $\chi^2(2, N = 127) = 1.123, p = .570, V = .09$.

5.6.2.4.1 Cognitive load and applying the harm principle. Three chi-squared tests for independence revealed no significant association between (a) cognitive load and applying the harm principle generally, $\chi^2(2, N = 127) = 0.605, p = .739, V = .07$; (b) cognitive load and applying the harm principle to boxing, $\chi^2(1, N = 127) = 0.475, p = .491, V = .06$; or (c) cognitive load and applying the harm principle to rugby, $\chi^2(1, N = 127) = 1.624, p = .202, V = .11$.

5.6.2.4.2 Cognitive load and endorsing and articulating the harm principle. Two chi-squared tests for independence also revealed no significant association between cognitive load and the endorsing of the harm principle, $\chi^2(1, N = 127) = 0.117, p = .732, V = .03$. There was a significant association between cognitive load and articulating the harm principle, $\chi^2(1, N = 127) = 3.895, p = .048, V = .18$, with harm being mentioned more frequently (19 times, 30%) in the control group than in the cognitive load group (9 times, 14%). This relationship is surprising, and was not observed in any of the previous studies. It is probable that this occurred due to chance, particularly in view of the failure to replicate the association between cognitive load and articulating the norm principle observed in Study 6.

5.6.2.4.3 Cognitive load and endorsing and articulating the norm principle. A final series of chi-squared tests for independence revealed no significant association between cognitive load and the endorsing of the norm principle, $\chi^2(1, N = 127) = 0.008, p = .930, V = .008$, or the articulating of the norm principle, $\chi^2(1, N = 127) < .001, p > .999$. Again, this suggests that the relationship observed in Study 6 occurred due to chance.

5.6.2.5 Cognitive load and responses to critical slide. The responses to the critical slide for the experiment group and the control group were analysed separately. In the group under cognitive load, 19 participants (29.69%) selected “There is nothing wrong”, 10 participants (15.62%) selected “It's wrong but I can't think of a reason”, and 35 participants (54.69%) selected “It's wrong and I can provide a valid reason”. In the control group, 12 participants (19.05%) selected “There is nothing wrong”, 10 participants (15.87%) selected “It's wrong but I can't think of a reason”, and 41 participants (65.08%) selected “It's wrong and I can provide a valid reason”. A chi-squared test for independence revealed no association between experimental condition and response to the critical slide, $\chi^2(2, N = 127) = 2.047, p = .359, V = .13$. The observed power was .23. Figure 5.7 shows the responses to the critical slide depending on cognitive load. Table 5.9 shows the observed counts, expected counts, and standardised residuals for each response depending on cognitive load.

The predicted relationship between engagement with a cognitive load task and response to the critical slide that was observed in previous studies was not found in Study 9. The predicted lower rate of providing reasons appears to be present to some degree, however this difference is not statistically significant.

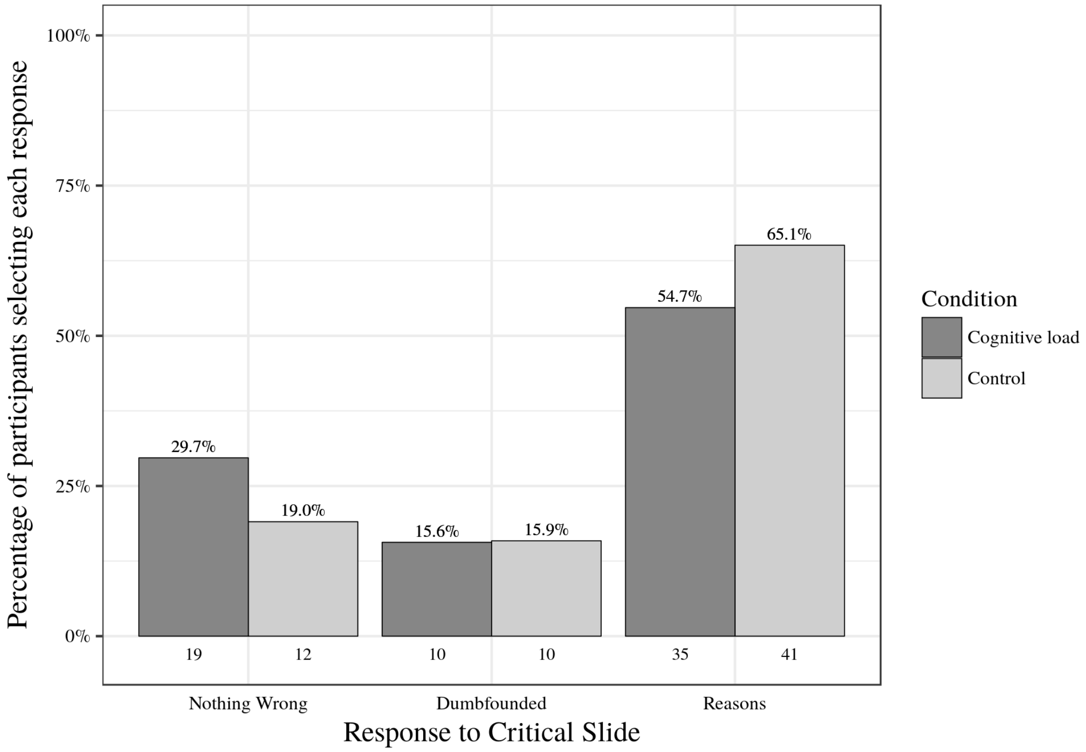

Figure 5.7: Study 9: Responses to critical slide and Cognitive Load

Table 5.9: Study 9 – Observed counts, expected counts, and standardised residuals for each response to the critical slide depending on cognitive load

Response		Cognitive Load / (engaged)	Control / (not-engaged)
Observed count	Nothing wrong	19	12
	Dumbfounded	10	10
	Reasons	35	41
Expected count	Nothing wrong	15.62	15.38
	Dumbfounded	10.08	9.92
	Reasons	38.30	37.70
Standardised residuals	Nothing wrong	1.40	-1.40
	Dumbfounded	-0.04	0.04
	Reasons	-1.19	1.19

*Note: * = sig. at p < .05 (|z| > 1.96); ** = sig. at p < .001 (|z| > 3.11)*

5.6.2.6 Individual differences and providing reasons. The hypothesised relationship between Need for Cognition and responses to the critical slide was investigated. As in previous studies, this analysis was primarily exploratory. The key hypothesis was that higher Need for Cognition scores would be predict the providing of reasons, and lower scores would predict dumbfounded responding. A multinomial logistical regression was conducted and no statistically significant association between Need for Cognition and response to the critical slide was found, $\chi^2(2, N = 127) = 1.50, p = .472$. The observed power was .18.

An independent samples t-test revealed no difference in Need for Cognition between the cognitive load group, ($M = 5.75$, $SD = 1.55$), and the control group, ($M = 1.57$, $SD = 1.65$), $t(123.49) = 0.677$, $p = 0.500$.

5.6.3 Study 9 discussion. The relationship between cognitive load and providing reasons observed in previous studies was not observed in Study 9. Rates of providing reasons in the cognitive load group appeared lower than in the control group, though not significantly. Interestingly any decrease in providing reasons appears to have led to an increase in selecting “There is nothing wrong”. It has been speculated throughout that a selecting “There is nothing wrong” in cases where this involves a change of mind requires more deliberation than a dumbfounded response. If this is true than the increase in selecting “There is nothing wrong” as opposed to dumbfounded responding (observed in Studies 7 and 8) may suggest that the cognitive load manipulation in this study did not inhibit deliberation to the same degree as it did in Studies 7 and 8. It is also worth noting that the manipulation check did not reveal any differences between the control group and the experimental group. This provides suggestive evidence that the failure of Study 9 to replicate previous results may be due to an ineffective manipulation of cognitive load. Difficulties in using cognitive load manipulations with online samples have identified elsewhere (Rand, 2012), and these difficulties have been evident throughout Studies 7, 8, and 9.

Furthermore, the dependent variable in these studies is nominal/categorical. This variable type is not well suited for identifying small or subtle effects. Consider the extensive research identifying disgust effects on moral judgements (e.g., Cameron et al., 2013; David & Olatunji, 2011; Landy & Goodwin, 2015; Rozin et al., 2009; Russell & Giner-Sorolla, 2011b, 2013; Sabo & Giner-Sorolla, 2017; Schnall et al., 2008; Wheatley & Haidt, 2005). May (2014) noted that the effect of incidental disgust is largely on the severity of a judgement, as opposed to altering the valence of the judgement. The use of a binary right/wrong measure of judgements therefore would fail to identify any effect for disgust on moral judgements. As such, the measures used in the dumbfounding paradigm mean that identifying influences on dumbfounding is particularly challenging, because minor influences on people's ability to provide reasons may not yield measurable effects.

Given the inconclusive and inconsistent results observed in Studies 6 to 9, the remainder of this chapter combines the results of all four studies in an attempt to identify what effect, if any, cognitive load, and engagement with a cognitive load task can reasonably be said to have on the dumbfounding paradigm.

5.7 Cognitive Load – Combined Results

5.7.1 Cognitive load and moral dumbfounding. The combined results for Studies 6-9 are displayed in Table 5.10, Table 5.11, and Figure 5.8. Table 5.10 shows the responses to the critical slide for each condition for each study, and for all four studies combined. Following Study 7 the initial failed replication attempt it was observed that the effectiveness of the cognitive load manipulation appeared to be related to engagement with the memory task. It is hypothesised that any influence of cognitive load on responses to the critical slide is dependent on level of engagement with the task. For the purposes of the individual analysis of each study a cut-off point of 7 or more correct answers on the dot pattern accompanying the critical slide was selected as the criterion for determining engagement with the task. The sample sizes of these studies meant that employing a stricter measure of engagement (e.g., 8 or 9 correct answers) may result in excluding too many participants to the point that the data set becomes unusable. This risk is considerably reduced in the combined analysis conducted here. As such a second measure of engagement is employed for the combined analysis. Table 5.11 shows the responses to the critical slide for each condition for each study, and for all four studies combined, when controlling for level engagement with the dot-pattern memory task (Studies 7, 8, and 9). For the following analyses, two measures of engagement employed, the primary measure (engaged a) follows from the analysis for each individual study and defines engagement as scoring 7 or more correct answers on the dot pattern accompanying the critical slide. The larger numbers of participants in the combined analysis allowed for a second, stricter measure of engagement (engaged b) to be investigated, whereby engagement is identified as scoring 8 or more correct answers. Figure 5.8 shows the percentage of participants who selected “It's wrong and I can provide a valid reason” for each condition, when controlling for engagement with the dot-pattern memory task. In addition, Figure 5.9 shows the rates of dumbfounding for each condition in each study. The rates of selecting each response to the critical slide for each study depending on experimental condition, without controlling for engagement with the task are displayed in Table 5.10. Table 5.11 shows the rates of selecting each response to the critical slide for each study when controlling for each measure of engagement.

Table 5.10: Responses to critical slide for all cognitive load studies (raw)

		Cognitive load		Control	
Study	Response	N	percent	N	percent
Study 6*	Nothing wrong	15	45%	5	15%
	Dumbfounded	6	18%	7	21%
	Reasons	12	36%	21	64%
Study 7	Nothing wrong	11	22%	13	27%
	Dumbfounded	15	29%	11	22%
	Reasons	25	49%	25	51%
Study 8*	Nothing wrong	21	21%	15	25%
	Dumbfounded	27	26%	5	8%
	Reasons	54	53%	41	67%
Study 9	Nothing wrong	23	25%	12	19%
	Dumbfounded	15	16%	10	16%
	Reasons	55	59%	41	65%
Studies 6, 7, 8, and 9	Nothing wrong	70	25%	45	22%
	Dumbfounded	63	23%	33	16%
	Reasons	146	52%	128	62%

Note. *significant variation at p < .05; **significant variation at p < .001

Table 5.11: Responses to critical slide for all cognitive load studies (controlling for engagement)

		Cognitive load		Control	
Study	Response	N	percent	N	percent
Study 6*	Nothing wrong	15	45%	5	15%
	Dumbfounded	6	18%	7	21%
	Reasons	12	36%	21	64%
Study 7 (engaged a)*	Nothing wrong	13	23%	11	25%
	Dumbfounded	20	36%	6	14%
	Reasons	23	41%	27	61%
Study 8 (engaged a)**	Nothing wrong	15	22%	15	25%
	Dumbfounded	17	25%	5	8%
	Reasons	36	53%	41	67%
Study 9 (engaged a)	Nothing wrong	19	30%	12	19%
	Dumbfounded	10	16%	10	16%
	Reasons	35	55%	41	65%
All studies (engaged a)*	Nothing wrong	62	28%	43	21%
	Dumbfounded	53	24%	28	14%
	Reasons	106	48%	130	65%
All studies (engaged b)**	Nothing wrong	62	29%	43	21%
	Dumbfounded	52	25%	28	14%
	Reasons	98	46%	130	65%

Note. *significant variation at p < .05; **significant variation at p < .001

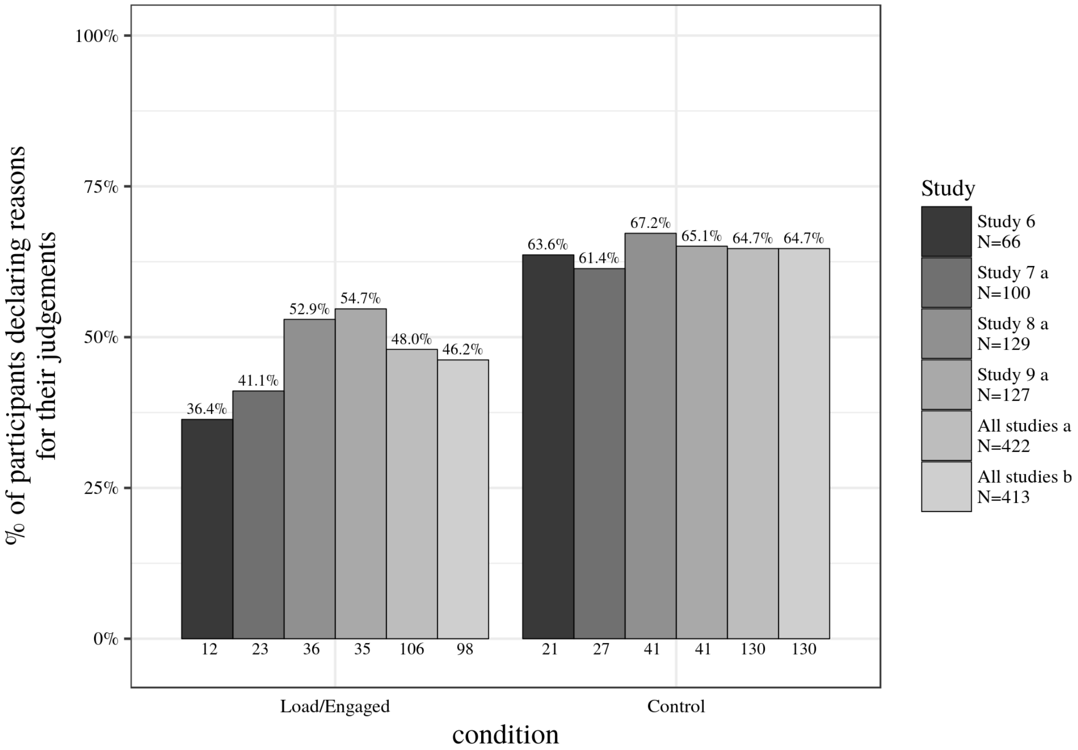

Figure 5.8: Rates of declaring reasons and cognitive load across each study

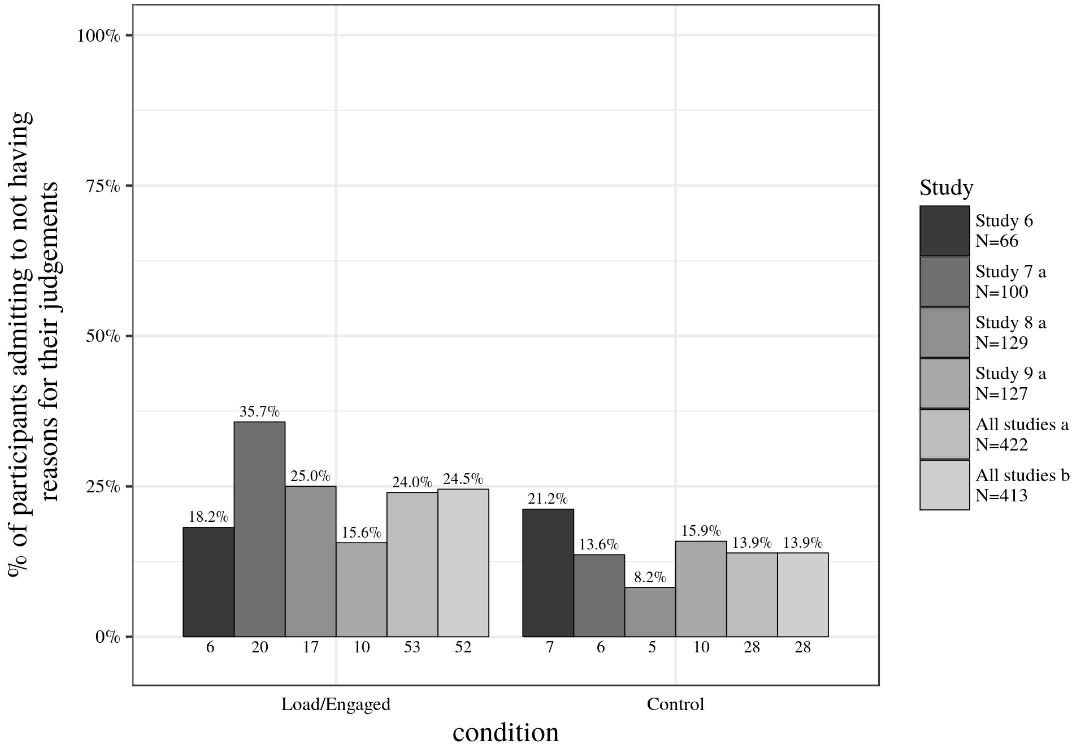

Figure 5.9: Rates of dumbfounding and cognitive load across each study

5.7.1.1 Studies 8 and 9 combined. Studies 8 and 9 were the most similar, with all materials identical, with the exception of the inclusion of a manipulation check in Study 9. The data from Study 8 and Study 9 were combined (*N* = 319, 195 in the experimental condition and 124 in the control condition). A chi-squared test for independence revealed no association between experimental condition and response to the critical slide, χ^2(2, *N* = 319) = 5.128, *p* = .077, *V* = .13. The observed power was .51. When engagement with the memory task was accounted for, as measured the primary measure of engagement, (*N* = 256, 132 in the experimental condition and 124 in the control condition), there was still no association, χ^2(2, *N* = 256) = 4.777, *p* = .092, *V* = .14. The observed power was .48. However, when the stricter measure of engagement was employed, (*N* = 247, 123 in the experimental condition and 124 in the control condition), a significant association between cognitive load and response to the critical slide was found, χ^2(2, *N* = 247) = 6.24, *p* =

.044, *V* = .16. The observed power was .60.

5.7.1.2 Studies 6, 8, and 9 combined. The data from Study 6 was then included in the analysis. Study 6 was included before Study 7 because the control in Study 6 more closely resembled the controls in Studies 8 and 9 than the control in Study 7 did (the control in Study 7 included a memory task while the controls in the other studies did not). Across the three studies (*N* = 385, 228 in the experimental condition and 157 in the control condition), a chi-squared test for independence revealed a significant association between experimental condition and response to the critical slide, $\chi^2(2, N = 319) = 6.233, p = .044, V = .14$. The observed power was .60. When engagement with the memory task was accounted for, as measured by the primary measure of engagement, (*N* = 322, 165 in the experimental condition and 157 in the control condition), there was still a significant association, $\chi^2(2, N = 256) = 7.724, p = .021, V = .17$. The observed power was .70. When the stricter measure of engagement was employed (*N* = 313, 156 in the experimental condition and 157 in the control condition), a significant association between cognitive load and response to the critical slide was found, $\chi^2(2, N = 247) = 9.821, p = .007$, *V* = . 20. The observed power was .80.

5.7.1.3 All Studies combined. Finally, the data from all four studies was combined. Taking the four studies together, without controlling for engagement with the memory task revealed no association between cognitive load and responses to the critical slide, $\chi^2(2, N = 485) = 5.121, p = .077, V = .16$. The observed power was .51. When engagement with the memory task was accounted for (using the primary measure of engagement (*N* = 422, 221 in the experimental condition and 201 in the control condition), a significant association between cognitive load and responses to the critical slide was found, $\chi^2(2, N = 422) = 12.675, p = .002, V = .17$. The observed power was .90. When the stricter measure of engagement was employed (N = 413, 212 in the experimental condition and 201 in the control condition), a significant association between cognitive load and response to the critical slide was found, χ^2(2, N = 413) = 14.847, p < .001, V = .19. The observed power was .94. This variation in responses can be seen in Table 5.10, Table 5.11, and Figure 5.8. Table 5.12

Table 5.12: Studies 6- 9 – Observed counts, expected counts, and standardised residuals for each response to the critical slide depending on cognitive load, when controlling for engagement (a)

Response		Cognitive Load / (engaged)	Control / (not-engaged)
Observed count	Nothing wrong	62	43
	Dumbfounded	53	28
	Reasons	106	130
Expected count	Nothing wrong	54.99	50.01
	Dumbfounded	42.42	38.58
	Reasons	123.59	112.41
Standardised residuals	Nothing wrong	1.58	-1.58
	Dumbfounded	2.62*	-2.62*
	Reasons	-3.45**	3.45**

*Note: * = sig. at p < .05 (|z| > 1.96); ** = sig. at p < .001 (|z| > 3.11)*

5.7.2 Need for Cognition and providing reasons. When analysed individually, for three of the four studies, there was no significant relationship between providing reasons and Need for Cognition. When the data for the four studies were combined, a multinomial logistical regression revealed a statistically significant association between Need for Cognition and response to the critical slide, χ^2(2, N = 413) = 10.29, p = .006. The observed power was .82. Need for Cognition explained between 2.4% (Cox and Snell R square) and 2.8% (Nadelkerke R squared) of the variance in responses to the critical slide. As Need for Cognition increased, participants were significantly more likely to provide reasons than to present as dumbfounded, Wald = 8.68, *p* = .003, odds ratio = 1.28, 95% CI [1.09, 1.50]. As Need for Cognition increased, participants were also significantly more likely to select "There is nothing wrong" than to present as dumbfounded, Wald = 7.77, *p* = . 005, odds ratio = 1.31, 95% CI [1.08, 1.58].

5.8 General Discussion

The aim of this chapter was to test a dual-process explanation of moral dumbfounding and address the research question identified in Chapter 2: "Can the existence of moral dumbfounding be explained by dual-process approaches to moral judgement" (2.2.1). Building on the initial insight of Cushman (2013) and drawing on the dual-process literature more generally (Bonner & Newell, 2010; De Neys, 2014; De Neys & Glumicic, 2008; Evans, 2007) it was hypothesised that moral dumbfounding emerges as a result of conflict between a habitual response (intuition) and deliberation. The relative roles of intuition and deliberation in producing each of the possible responses in the dumbfounding paradigm (1: providing reasons; 2: accepting the counter-arguments and revising judgement accordingly; and 3: a dumbfounded response) were identified, and these responses were ranked according to level of deliberation involved in each. It was hypothesised that, in requiring a successful deliberative search for reasons, response 1 (providing reasons) involves the most deliberation. The relative roles of intuition and deliberation in the remaining responses were less clear, particularly because of the possibility of that competing intuitions may emerge. If deliberation fails to identify justifications for an initial intuition, a person may find that they must choose between two competing intuitions (initial judgement; that moral judgements should be justifiable). Despite this limitation, a clear candidate for relying on deliberation the most was identified (response 1: providing reasons).

Having identified a response that clearly relied on deliberation more than the other responses, an experimental manipulation that is known to influence deliberative responding was selected for study (cognitive load, e.g., De Neys, 2006; Evans & Curtis-Holmes, 2005; Evans & Stanovich, 2013; Schmidt, 2016). In addition to this, an individual difference variable that has to been linked to a tendency to engage in deliberation was identified and selected for investigation (Need for Cognition, Cacioppo & Petty, 1982; Evans & Stanovich, 2013; Petty et al., 1984). Four studies were conducted investigating cognitive load and moral dumbfounding. The target response was “It's wrong and I can provide a valid reason” on the critical slide. It was hypothesised that cognitive load would inhibit participants' ability to provide reasons for their judgements, and that under cognitive load reduced incidences of participants stating they could provide a reason would be observed. It was also hypothesised that providing reasons would be related to Need for Cognition scores, such that participants who provided reasons would score higher on Need for Cognition than participants providing other responses on the critical slide.

Study 6 showed a clear reduction in providing reasons for a judgement depending under cognitive load cognitive load (resulting in an increase in selecting “There is nothing wrong”). The results of Study 7 indicated that any effect cognitive load had on ability to generate reasons for a judgement was related to level of engagement with the cognitive load task. Study 8 demonstrated that when controlling for engagement with the cognitive load task, cognitive load led to lower rates of providing reasons, however, Study 9 failed to replicate this finding. Three studies described here therefore showed that participants actively engaging with cognitive load task were consistently less effective at identifying reasons for their judgements. Study 9 showed a similar pattern of responses, however this was not statistically significant. There was no observed relationship between need for cognition and the critical slide for Studies 6, 7, and 9. In Study 8, higher need for cognition was associated more with providing reasons than dumbfounded responding. In the combined analysis, higher need for cognition scores were predicted higher rates of both providing reasons and selecting "there is nothing wrong" and reduced rates of dumbfounded responding.

The combined results of Studies 6-9 appear to indicate that engaging in a cognitive load task leads to reduced incidences of providing reasons (as hypothesised). However, given the failure of Study 9 to replicate this effect, caution is advised when interpreting this result. Furthermore, dumbfounded responding appeared to be linked to lower Need for Cognition scores. Again, caution is advised in interpreting this result, particularly given that only 1 (Study 8) of the Studies 6-9 individually found any relationship between Need for Cognition and response to the critical slide.

5.8.1 Implications. The results of the studies described in this chapter do not provide convincing evidence for the conflict in dual-processes explanation of moral dumbfounding identified here. That said, neither do the studies described in this chapter provide convincing evidence that this explanation of moral dumbfounding is wrong. The wider implications of this inconclusive result are unclear, beyond a that it seems that moral dumbfounding is likely to be more complex than the "conflict in dual-processes" explanation allows for. The complexity of moral dumbfounding becomes apparent when attempting to describe the various possible responses in the paradigm in terms of the relative roles of deliberation and intuition. The possibility that a failure of deliberation to provide reasons for an intuition may lead to alternative and competing intuitions to become salient means that in some cases, moral dumbfounding may involve conflict between competing intuitions.

The apparent conflict between competing intuitions is more complicated than conflict between intuition and deliberation. Recall that the examples of conflict identified at the beginning of this chapter clearly pitted an intuitive response against a deliberative response, and resolution of this conflict was relatively straight forward. For example, in base-rate neglect problems there is only one correct answer. The intuitive response is incorrect and deliberation leads to an alternative response. This conflict is easily resolved because deliberation has clearly identified the intuitive response as incorrect. In contrast moral judgements are not clearly correct or incorrect and the failure of deliberation to identify reasons for an intuition is not necessarily evidence that the intuition is incorrect.

That moral dumbfounding is more complicated than classic cases of conflict in dual-processes does not necessarily mean that attempting to explain moral dumbfounding as dual-process conflict is without merit. It is possible that conflict in dual-processes is still implicated in moral dumbfounding. However, where conflict is normally resolved through suppression of erroneous intuitive responses, in the dumbfounding paradigm this conflict is not easily resolved leading participants are faced with a conflict between competing intuitions. Given that this explanation identifies dumbfounding as arising from conflict between competing intuitions (same underlying mechanisms) as opposed to conflict between intuition and deliberation (different underlying mechanisms) means that testing this explanation is particularly problematic.

5.8.2 Limitations and future directions. There are a number of related limitations of the studies conducted in this chapter that may have contributed to the inconclusive results in different ways. Firstly, it is possible that the inconclusive result is due to limitations in the method of data collection employed across Studies 7, 8, and 9. Online data collection through MTurk is useful for ease of access to willing participants resulting in highly efficient data collection. However, the absence of oversight or supervision of participants taking part in the studies means that complex manipulations such as cognitive load manipulations may not be as successful as in a controlled laboratory setting (Crump et al., 2013; Goodman et al., 2013). In response to the difficulties in manipulating cognitive load in an online study, the studies described in the next chapter employed an alternative experimental manipulation. Another limitation that may also have contributed to the inconclusive result is that the identification of the relative roles of intuition and deliberation in the possible responses in the dumbfounding paradigm was more difficult because of the possibility of additional competing intuitions. This possibility of additional competing intuitions may mean that moral dumbfounding is more complex than simply a conflict between intuition and deliberation. In response to the apparent complexity of moral dumbfounding, additional individual difference variables were investigated on an exploratory basis in an attempt to identify possible alternative causes or influences of moral dumbfounding.

5.9 Conclusion

The studies described in this chapter aimed to investigate if cognitive load influenced the degree to which people successfully identified reasons for their judgements. The individual difference variable Need for Cognition was also investigated.

For 3 of the 4 studies conducted, there did not appear to be a relationship between Need for Cognition and response to the critical slide, however in Study 8, and the final analysis of all the studies together, lower need for cognition scores were associated with dumbfounded responding. The effect size was quite small, and this effect only emerged in 1 study, and when all the data were combined. However, this may prove useful in furthering out understanding of moral dumbfounding and the making of moral judgements more generally. That dumbfounding may be weakly related to scoring lower on Need for Cognition is interesting and may provide a useful area for follow-up work.

The primary aim of this chapter was to test a dual-process conflict explanation of moral dumbfounding. According to this view, providing reasons for a judgement is grounded in deliberation, and dumbfounded responding is grounded in habitual responding. The extent to which revising a judgement is grounded in deliberation or habitual responding is unclear, and may vary depending on the individual. The key prediction of this explanation is that inhibiting deliberative responding, through a cognitive load manipulation, would inhibit the identification of reasons.

Despite significant methodological issues encountered with conducting complex experimental manipulations using online samples, the four studies presented in this chapter offer suggestive evidence that cognitive load inhibits the identification of reasons in the dumbfounding paradigm. Engagement with the cognitive load manipulation on the part of the participants is essential for it to successfully inhibit the identification of reasons. The pattern of responses across the studies is consistent with the possibility that revising a judgement requires more deliberation than providing a dumbfounded response, however this claim was not tested directly, and there is no clear evidence in support of it. Evidence in support of the primary prediction, that inhibiting deliberative responding should inhibit the identification of reasons was found.

Significant methodological challenges were encountered, and the evidence for the dual-process explanation of moral dumbfounding was not conclusive. The methodological challenges can largely be attributed to the difficulty in implementing an effective cognitive load manipulation with an online sample. In order to test the dual-process explanation of moral dumbfounding more rigorously, an alternative manipulation and methods are required. Chapter 6 examines, a second prediction of the dual-process explanation of moral dumbfounding, that facilitating deliberation should lead to an increase in providing reasons for judgements.

6 Chapter 6 – Influencing Dumbfounded Responding 2: Facilitating the Identification of Reasons

The previous chapter attempted to test two predictions of a conflict in dual-processes explanation of moral dumbfounding: that a cognitive load task designed to inhibit deliberation should inhibit the identification of reasons, and that a tendency to identify reasons would be related to Need for Cognition. The findings were inconclusive, cognitive load appeared to inhibit the identification of reasons in some cases, Need for Cognition only appeared to be related to dumbfounded responding in 1 out of 4 studies, and when all studies were combined. It seems likely that moral dumbfounding is more complex than conflict between intuition and deliberation, particularly in view of the possibility that it may involve competing intuitions. It is also possible that the failure to produce convincing results in Chapter 5 may be attributed to methodological challenges arising from the manipulation employed (cognitive load). The aims of this chapter are twofold. Firstly this chapter will attempt to address the methodological challenges surrounding the experimental manipulation employed in Chapter 5, identifying alternate experimental manipulations, to test specific predictions of two related explanations of moral dumbfounding. Secondly, given that a weak association between dumbfounded responding and Need for Cognition was only observed when the data from all Studies 6-9 were combined, alternative individual difference variables are identified and explored in this chapter. This further investigation of individual differences is informed by dual-process theories of cognition (e.g., Chaiken & Trope, 1999), and potential explanations of dumbfounding discussed in Chapter 4.

The primary aim of this chapter is to examine the degree to which moral dumbfounding can be explained by dual-process approaches to moral judgement.

The theoretical position adopted and being tested in this chapter is essentially the same as in Chapter 5, that moral dumbfounding emerges as a result of dual-process conflict. From this, the providing of reasons remains the response of interest in that it is clearly identifiable as being grounded in deliberation to a greater extent than the other possible responses in the dumbfounding paradigm. Having failed to convincingly inhibit this response in the studies in Chapter 5, two alternative manipulations designed to facilitate deliberation and the identification of reasons are identified and tested in this chapter.

Firstly, drawing on the dual-process literature more generally (Research Question 2.1.2), the degree to which the identification of reasons can be facilitated by increased psychological distance is tested (Studies 10 and 11). Secondly, drawing specifically on model theory (Research Question 2.1.3), the degree to which prompting participants with information relevant to identifying reasons facilitates their providing of reasons is tested (Study 12). Three individual difference variables are investigated in this chapter, Cognitive Reflection Test (CRT: Frederick, 2005; Thomson & Oppenheimer, 2016; Toplak, West, & Stanovich, 2011), Need for Closure (NFC: Kruglanski, 2013; Kruglanski, Atash, De Grada, Mannetti, & Pierro, 2013; Kruglanski & Webster, 1996), and social desirability (Ballard, 1992; Crowne & Marlowe, 1960; Fischer & Fick, 1993; Strahan & Gerbasi, 1972). The hypothesised relationship between two of these measures (CRT and NFC) and susceptibility to dumbfounding/ability to provide reasons is informed by a dual-process explanation of moral dumbfounding. The social desirability scale is included to investigate the claims that moral dumbfounding occurs as a result of the social situation (e.g., Royzman et al., 2015; see Chapter 4 for discussion).

6.1 Facilitating the Identification of Reasons

6.1.1 Psychological distancing and deliberation. The concept of psychological distancing has been linked to construal level theory whereby increased psychological distance is associated with higher level construals (Liberman & Trope, 1998, 2008; Liberman, Trope, & Stephan, 2007; Trope & Liberman, 2003). The level of psychological distance refers to the degree to which something is removed from direct experience. Four dimensions of psychological distance have been identified (Liberman et al., 2007), temporal distance (thinking about past/future events), spatial distance (thinking about spatially remote locations), social distance (perspective taking), and hypotheticality (thinking about hypothetical situations). According to construal level theory, increases in psychological distance involve higher level construals while decreased psychological distance involves lower level construals. Construal level is related to abstraction, whereby higher levels of construal are associated with more abstract thinking with reduced emotional influence. A study by Kross and Ayduk (2008) investigated differences between immersed or distanced analysis of a negative emotional experience. Participants were provided with instructions in how to analyse the experience, and construal level was manipulated through different emphases in these instructions. It was found that level of negative affect associated with the analysis was significantly lower when using a distanced analysis than an immersed analysis.

A series of studies by Liberman, Sagristano, and Trope (2002) manipulated psychological (temporal) distance by requiring participants to think about things in the immediate future or in the distant future. Participants were presented with a hypothetical scenario (e.g., “Imagine that you will be having a yard sale...”) that was framed in the near future (“this coming Friday”) or the distant future (“sometime next summer), along with 38 related objects. Participants were asked to group the 38 items. It was found that when the scenarios were framed in the distant future, participants created fewer broader groups. This link between increased psychological distance and increased levels of abstraction has been demonstrated elsewhere (e.g., Liberman & Trope, 1998, 2008; Liberman et al., 2007; Trope & Liberman, 2003).

Abstract thinking (and higher level construals) is also associated with deliberative responding (more commonly known as ‘System 2’ or analytical thinking e.g., Evans, 2008). This link between deliberative responding, abstract thinking, and higher levels of construal means that higher levels of construal may be viewed as involving more deliberation. A corollary of this is that increased psychological distance, in leading to higher level construals and abstract thinking, facilitates deliberative responding. This link between psychological distance and deliberative responding can be seen in a study by Fujita, Trope, Liberman, and Levin-Sagi (2006). They show that increased psychological distance is associated with increased levels of self control. Drawing on Metcalf and Mischel (1999) Fujita et al. identify self control as involving the “cool system” (2006, p. 2). Or in the language of the dual-process approach adopted here the “cool system” may be equated with deliberation while intuition resides in the “hot system” (Fujita et al., 2006; Metcalfe & Mischel, 1999). Drawing on this link between psychological distance and deliberation, it is hypothesised here that experimentally manipulating psychological distance in the dumbfounding paradigm will facilitate the identification of reasons, leading to reduced rates of dumbfounding. This hypothesis is tested in Studies 10 and 11.

A second, potentially confounding, way in which manipulations of psychological distance may influence responses in the dumbfounding paradigm is that increases psychological distance are associated with reduced emotional influence (e.g., Kross & Ayduk, 2008). The emotional reaction that participants may have to the dumbfounding scenarios has been noted as potentially contributing to the occurrence moral dumbfounding (e.g., Haidt, 2001; Prinz, 2005). Manipulating psychological distance may therefore lead to reduced emotional reactions to the scenarios, and as such, reducing participants' susceptibility to dumbfounding. This means that any observed influence of psychological distancing on the dumbfounding paradigm will be subject to follow-up study to assess if this influence can be attributed to (a) the facilitation of analytical thinking, or (b) the reduced influence of emotion.

6.1.2 Mental models and providing reasons. The main focus of model theory (Bucciarelli et al., 2008) is on conscious reasoning (deliberation). According to model theory, people reason using mental models, specifically, for discussions of moral reasoning, people use mental models to reason about deontic propositions. Crucially, for studying moral dumbfounding, mental models are generally incomplete. Incomplete mental models generally contain sufficient information for the current goal, however in some cases the incomplete nature of the mental model results in errors. Recall the apple and orange problem discussed in Chapter 1 (taken from Bucciarelli et al., 2008, p. 126):

> You are permitted to carry out only one of the following two actions:
>
> Action 1: Take the apple or the orange, or both.
>
> Action 2: Take the pear or the orange, or both.
>
> Are you permitted to take the orange?

Building a complete model (e.g., with a pen and paper/an excel sheet) makes it clear that taking the orange is not permitted. An example of how to approach building a complete model is sketched below. The permissible possibilities associated with each action are:

Action 1	Action 2
Take the apple	Take the pear
Take the orange	Take the orange
Take both the apple and orange	Take both the pear and orange

The instruction states that you are permitted to carry out only one of the actions. Such that if a person carries any of the permissible actions associated with Action 1, all actions associated with Action 2 are impermissible, see below, actions deemed impermissible due to overlap with Action 2 are denoted as such with parentheses:

Action 1	Action 2
Take the apple	~~Take the pear~~
(Take the orange)	~~Take the orange~~
Take both the apple (and orange)	~~Take both the pear and orange~~

Similarly, if a person carries out an action associated with Action 2, all actions associated with Action 1 are impermissible (again overlapping actions are denoted with parentheses):

Action 1	Action 2
~~Take the apple~~	Take the pear
~~Take the orange~~	(Take the orange)
~~Take both the apple and orange~~	Take both the pear (and orange)

In both cases taking the orange is impermissible. Our incomplete mental models result in the impermissibility of taking the orange being overlooked.

Deliberation may help a person to flesh out a given model in order to successfully arrive at the correct answer.

The making of moral judgements involves reasoning about deontic principles using mental models. That these mental models are incomplete provides a possible explanation for moral dumbfounding. When people make a moral judgement they form a mental model about the permissibility or impermissibility of an action based on a deontic proposition. This mental model is incomplete, and therefore it may or may not include reasons in support of the relevant deontic proposition. Mental models contain sufficient information to achieve the current goal. If the goal is simply to make a judgement, then information relating to reasons for this judgement is superfluous to the goal at hand and it is therefore unlikely to be included in the mental model. It is possible that when asked to provide a reason for a judgement people will update or flesh out their mental model to identify reasons for their judgement. It is also possible (though unlikely given the emphasis on the incompleteness of mental models) that the initial mental model does contain basic reasons. In either case however, these reasons (either identified through deliberation or part of the original model) are subsequently refuted during the presentation of the slides, rendering participants dumbfounded if they cannot identify alternative reasons.

Extensive research on mental models in reasoning has illustrated that the content of mental models can be altered by varying the instructions or description of a given reasoning problem (e.g., Johnson-Laird, 2006). Building on this, it is hypothesised that if a reason for a judgement is made salient to a participant prior to presenting them with the moral judgement task, this reason may be included in their mental model and possibly form the basis of the making of the judgement. The inclusion of this reason in participants' mental models should enable participants to provide that reason as justification for their judgement when questioned at a later stage. This hypothesis is tested in Study 12.

6.2 Individual Differences and Moral Dumbfounding

In addition to introducing alternative manipulations in order to further test predictions of the dual-process explanation of moral dumbfounding, three individual difference variables will also be investigated. The possible relationship between two of these variables and dumbfounded responding is directly informed by the dual-process approach adopted here; the third variable is included as a follow-up on the possible social influences on moral dumbfounding discussed in Chapter 4. Firstly, it is hypothesised that the Cognitive Reflection Test (CRT: Frederick, 2005; K. S. Thomson & Oppenheimer, 2016; Toplak et al., 2011) is related to moral dumbfounding such that people who score higher in CRT will be better at providing reasons for their judgements. The CRT provides a measure of people's tendency to over-ride intuitive (habitual) responses and engage in deliberation in order to ensure accuracy in responding (Toplak et al., 2011). The CRT has previously been used in moral judgement tasks, for example, Royzman, Landy, and Goodwin (2014) found that people who scored higher on CRT were less likely to rate incest as not wrong. Building on the Royzman et al. (2014) finding it is hypothesised here that beyond being related to people's judgements on incest, CRT may be related to (a) people's ability to provide reasons for their judgements, or (b) related to the degree to which people ability to justify their judgements influences their judgements. The measurement of CRT is introduced in Study 11.

Secondly, it is hypothesised that dumbfounded responding may be related to Need for Closure (NFC: Kruglanski, 2013; Kruglanski et al., 2013; Kruglanski &

Webster, 1996) such that people who score high in need for closure will be more susceptible to dumbfounding. Need for Closure is related to the degree to which people avoid ambiguity (Kruglanski, 2013, p. 6). In order to avoid ambiguity, people may engage in “seizing” or “freezing” behaviours (Kruglanski & Webster, 1996), relying on intuition over deliberation. They stick to their initial intuition and refuse to question it or engage in deliberation. This description provides an explanation of the inclusion of the term “stubborn” in original definition (Haidt et al., 2000, p. 2). Dumbfounded participants appear motivated to maintain their judgement despite evidence that they may be mistaken (an absence of reasons for their judgement). The possibility that a moral judgement may not be grounded in reasons, may provide a source of ambiguity that is aversive to people who score high in NFC. It is hypothesised here that participants who score high in NFC are be more susceptible to dumbfounding. NFC was introduced in Study 11.

Finally, in response to the claim by Royzman et al. (2015) that dumbfounding may arise as a result of social pressure, it was hypothesised that dumbfounding may be linked to social desirability (e.g., Chung & Monroe, 2003; Latif, 2000; Morris & McDonald, 2013). In order to investigate this possibility, the short version of the Marlowe-Crowne (Crowne & Marlowe, 1960) social desirability scale (devised by Strahan & Gerbasi, 1972; see also Ballard, 1992; Fischer & Fick, 1993) was introduced in Study 10.

6.3 Study 10: Distancing and Moral Dumbfounding

Study 10 was an attempt to facilitate the identification of reasons for a moral judgement through increased psychological distance, manipulated through perspective taking.

6.3.1 Method.

6.3.1.1 Participants and design. Study 10 was a between-subjects design with social desirability additionally measured as a potential correlate and moderator variable. The dependent variable was response to the critical slide. The independent variable was distancing with two levels: present and absent. Participants were randomly assigned to two conditions.

A total sample of 120 participants[13] (62 female, 58 male; M_{age} = 38.02, min = 22, max = 75, *SD* = 11.90) took part. Participants were recruited through MTurk. Participation was voluntary and participants were paid 0.50 US dollars for their participation. Participants were recruited from English speaking countries or from countries where residents generally have a high level of English (e.g., The Netherlands, Denmark, Sweden).

6.3.1.2 Procedure and materials. Data were collected using an online questionnaire generated using Questback (Unipark 2013). The substantive content of the questionnaire was the same as was used in Studies 6 through 9. It opened with questions relating to basic demographics. The two statements relating to the norm principle were then presented. At this point, the distancing group were presented with an additional set of instructions which read as follows:

> Anne is a student of philosophy. She generally shows a good understanding of the subject matter, and this is reflected in her grades. Sometimes, however, she may adopt a position on an issue and struggle (or even fail) to defend it.

[13] A priori power analysis indicated that in order to detect a large effect size (*V* = .5) with 80% power, a sample of 39 participants was required. In order to detect a medium effect size (*V* = .3) with 80% power a sample of 107 participants was required.

> She is currently taking a course in ethics and has been asked to study the following scenario.
>
> How should Anne judge the actions of the two people involved?
>
> What reasons would you use to explain why she should make that judgement?

Following this, participants read the *Incest* vignette. The distance-absent (control) group proceeded straight from the normative statements to the *Incest* vignette. The vignette was the same for both groups. Participants were then asked to rate the behaviour described in the vignette. In the distancing group, all questions were phrased in terms of "Anne", (e.g., "How should Anne rate the behaviour of Julie and Mark?"; "What reasons would you use to explain why Anne should make that judgement?"). In the control group participants were asked for their own judgements (e.g., "How would you rate the behaviour of Julie and Mark?"; "Please provide the reason for this judgement in the space below").

Participants were then presented with a series of counter-arguments designed to undermine any reasons participants may have identified in support of a judgement of "wrong". Again, in the distancing, these were phrased in relation to the judgement made by "Anne", while in the control they referred to the judgement of the participant themselves. Following the counter-arguments, participants were presented with the critical slide. There was no difference in the critical slide between the groups. As in previous studies, the critical slide contained read: "Julie and Mark's actions did not harm anyone, or negatively affect anyone. How can there be anything wrong with what they did?" with three possible response options (a) "There is nothing wrong"; (b) "It's wrong but I can't think of a reason"; (c) "It's wrong and I can provide a valid reason". The order of these response options was randomised. Participants who selected option (c) were then required to provide a reason.

Participants then rated the behaviour again, rated their confidence in their decision, and completed the same post-discussion questionnaire as in previous studies (Haidt, Björklund, and Murphy 2000). They then responded to the credulity check questions devised by Royzman, Kim, and Leeman (2015), and answered the three questions relating to the application of the harm principle, devised in Chapter 4.

Participants then completed the short version of the Marlowe-Crowne (Crowne & Marlowe, 1960) social desirability scale (devised by Strahan & Gerbasi, 1972; see also Ballard, 1992; Fischer & Fick, 1993). This consisted of ten questions (e.g., "There have been occasions when I took advantage of someone.", "I never resent being asked to return a favor.") to which participants selected "True" or "False".

6.3.2 Results and discussion. Seventy seven participants (64.17 %) rated the behaviour of Julie and Mark as wrong initially. The mean initial rating of the behaviour was, *M* = 2.88, *SD* = 2.08. Seventy three participants, (60.83 %) rated the behaviour as wrong after viewing the counter-arguments and the critical slide. The mean revised rating of the behaviour was, *M* = 2.94, *SD* = 2.03. A paired samples t-test revealed no difference in rating from time one, (*M* = 2.88, *SD* = 2.08), to time two, (*M* = 2.94, *SD* = 2.03), *t*(118) = -0.716 , *p* = .476. Further analysis revealed that 10 participants changed the valence of their judgement: six participants changed their judgement from "wrong" to "neutral"; one participant changed their judgement from "right" to "wrong"; one participant changed their judgement from "right" to "neutral"; one participant changed their judgement from "neutral" to "right"; and one participants changed their judgement from “neutral” to “wrong”. A chi-squared test for independence revealed no significant association between time of judgement and valence of judgement made, $\chi^2(2, N = 120) = 0.794, p = .788, V = .08$.

6.3.2.1 Baseline rates of dumbfounding. Participants who selected the admission of not having reasons on the critical slide were identified as dumbfounded. Twenty participants (16.67%) selected “It's wrong but I can't think of a reason”. Sixty two participants (51.67%) selected “It's wrong and I can provide a valid reason”; and thirty eight participants (31.67%) selected “There is nothing wrong”. Table 6.1 shows the frequency of each response on the critical slide depending on condition.

Table 6.1: Rates of selecting each response to the critical slide in Study 10

	Distance		Control	
Response to critical slide	N	percent	N	percent
There is nothing wrong.	15	25%	23	38%
It's wrong but I can't think of a reason.	11	18%	9	15%
It's wrong and I can provide a valid reason.	34	57%	28	47%

6.3.2.2 Dumbfounding and coded string responses. Participants who selected “It's wrong and I can provide a valid reason” were required to provide a reason. The reasons provided were coded for unsupported declarations or tautological reasons. An additional 8 participants were identified as dumbfounded following this coding. Three participants provided unsupported declarations (e.g., “Incest is wrong”), two participants provided tautological reasons (e.g., “They are brother and sister”), one participant provided an unsupported declarations accompanied by a tautological reason, one participant provided an unsupported declaration accompanied by a statement with no reason (“It is incest. Incest is morally wrong. There is nothing more to say about it.”), and one participant admitted to not having a reason (“no reason”). Taking the coded string responses into account brought the total number of participants identified as dumbfounded to 28 (23.33%). Given the concerns discussed in Chapter 4 the remaining analysis will take the stricter measure of dumbfounding only, the selecting of an admission of not having reasons, as in previous studies. The participant who stated “no reason” was is also taken as dumbfounded for the remaining analysis.

6.3.2.3 Distancing and eligibility for analysis. As in the studies discussed in Chapter 5, a series of tests was conducted to assess if there was a relationship between the experimental manipulation and eligibility for analysis. A chi-squared test for independence revealed no significant association between distancing and overall eligibility for analysis, $\chi^2(2, N = 120) = 4.41, p = .110, V = .19$.

6.3.2.3.1 Distancing and applying the harm principle. Three chi-squared tests for independence revealed no significant association between (a) distancing and applying the harm principle generally, $\chi^2(2, N = 120) = 1.141, p = .565, V = .09$; (b) distancing and applying the harm principle to boxing, $\chi^2(1, N = 120) = 0.047, p = .828, V = .02$; or (c) distancing and applying the harm principle to rugby, $\chi^2(1, N = 120) = 2.501, p = .114, V = .14$.

6.3.2.3.2 Distancing and endorsing and articulating the harm principle. Two chi-squared tests for independence also revealed no significant association between distancing and the endorsing of the harm principle, $\chi^2(1, N = 120) = 0.61, p = .435, V = .02$, or articulating the harm principle, $\chi^2(1, N = 120) = 0.215, p = .643, V = .04$.

6.3.2.3.3 Distancing and endorsing and articulating the norm principle. A final series of chi-squared tests for independence revealed no significant association between distancing and the endorsing of the norm principle, $\chi^2(1, N = 120) = 2.139, p = .144, V = .13$, or the articulating of the norm principle, $\chi^2(1, N = 120) < .001, p$

> .999.

6.3.2.4 Distancing and responses to critical slide. The responses to the critical slide for the experiment group and the control group were analysed separately. In the experimental group, 15 participants (25%) selected “There is nothing wrong”, eleven participants (18.33%) selected “It's wrong but I can't think of a reason”, and 34 participants (56.67%) selected “It's wrong and I can provide a valid reason”. In the control group, twenty three participants (38.33%) selected “There is nothing wrong”, ten participants (16.67%) selected “It's wrong but I can't think of a reason”, and 27 participants (45%) selected “It's wrong and I can provide a valid reason”. A chi-squared test for independence revealed no significant association between experimental condition and response to the critical slide, $\chi^2(2, N = 120) = 2.465$, $p = .292$, $V = .14$. The observed power was .27. Figure 6.1 shows the varying responses to the critical slide depending experimental condition (with the inclusion of the participant who admitted to not having a reason in their open-ended response).

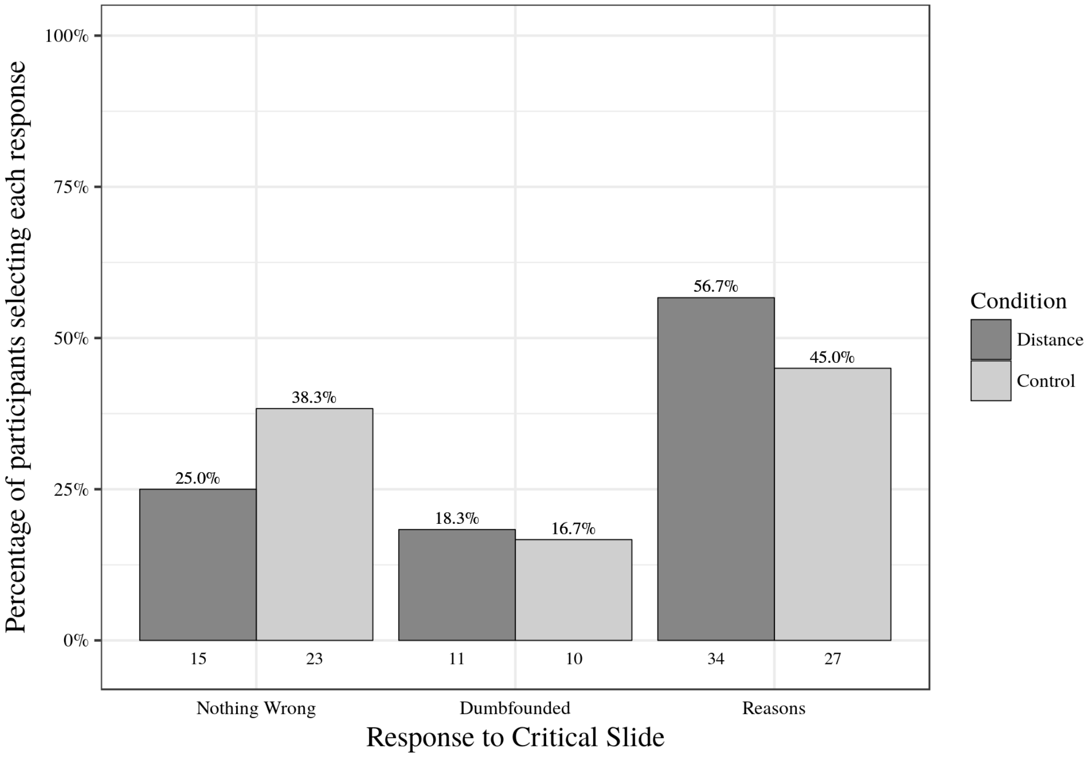

Figure 6.1: Study 10 – Dumbfounding and Distancing

The results did not identify a significant difference in ability to provide reasons between the distancing group and the control group. Interestingly, the rate of providing reasons in the control group was considerably lower than in controls for the studies in Chapter 5, and rates of selecting "There is nothing wrong" in the control group were noticeably higher than in previous studies. This is unexpected, as the procedure for participants in the control group in Study 10 was largely the same as the procedure for participants in the control groups in Chapter 5 (with the exception of Study 7, in which participants completed a simple memory task). Rates of providing reasons in the control groups in any of the studies described in Chapter 5 did not fall below 60%. In light of this trend, 45% of participants in the control group in this study providing reasons is unusual. Given that there were no substantive differences in materials or procedure it is likely that this is due to chance, perhaps the sample in Study 10 was particularly lenient. However, this finding has illustrated a further difficulty in studying moral dumbfounding, that responses to the critical slide can fluctuate from sample to sample for no apparent reason.

6.3.2.6 Social desirability and dumbfounding. An exploratory analysis was conducted to investigate if there was a relationship between social desirability and dumbfounded responding. A multinomial logistical regression revealed no statistically significant association between social desirability and response to the critical slide, $\chi^2(2, N = 120) = .84, p = .656$. The observed power was .12.

6.3.3 Study 10 discussion. The primary hypothesis of Study 10 was that a distancing manipulation would facilitate deliberation and by extension facilitate the identification of reasons, reducing dumbfounding. It was expected that participants in the distancing group would provide reasons for their judgements at a higher rate than participants in the control group. Evidence in support of this hypothesis was not found. The secondary hypothesis, that dumbfounding might be related to social desirability, was also unsupported. Regarding the primary hypothesis, two of methodological considerations must be addressed before rejecting it.

Firstly, as discussed in Chapter 5, the measures in the dumbfounding paradigm are not suited for identifying small or weak effects. Recall that the final study in Chapter 5 (Study 9) failed to replicate the previous studies, but the general trend in results could be identified. Convincing results in Chapter 5 were only found when all the studies were combined. Secondly, the distancing manipulation in Study 10 did not include a direct instruction to view the scenario from Anne's perspective. The inclusion of such an instruction may provide a stronger distance manipulation.

A further concern is the unusually low rates of providing reasons generally in Study 10. The general trend of responses between the distancing group and the control group, though not statistically significant, appeared to present as expected, however the rates of providing reasons for both groups were substantially lower than expected. The only explanation current for this is an irregular sample. However, it highlights that the underlying factors that lead participants to provide reasons, revise their judgements, or present as dumbfounded are as yet unknown. Study 10 failed to provide evidence that distancing may facilitate the identification of reasons. However, one specific methodological limitation can be addressed in a follow-up study, the inclusion of a direct instruction to view the scenario from the perspective of another person.

6.4 Study 11: Distancing 2 – Direct instruction to take alternative perspective

Study 10 did not provide evidence that distancing facilitates the identification of reasons in the moral dumbfounding paradigm. However, a specific weakness in the materials in Study 10 was identified, that there was no direct instruction to take Anne's perspective. Study 11 served to address this limitation.

6.4.1 Method.

6.4.1.1 Participants and design. Study 11 was a between-subjects design with social desirability, CRT and NFC additionally measured as potential correlate and moderator variables. The dependent variable was response to the critical slide. The independent variable was distancing with two levels: present and absent. An initial sample of 105[14] participants (49 female, 55 male, 1 other; M_{age} = 37.46, min = 19, max = 83, SD = 12.20) was collected. Participants were excluded for including nonsense text in the open-ended response questions, or for lying in the need for closure scale (a combined score of > 15 on selected items). In addition, participants

[14] As in Study 10, a priori power analysis indicated that in order to detect a large effect size (V = .5) with 80% power, a sample of 39 participants was required. In order to detect a medium effect size (V = .3) with 80% power a sample of 107 participants was required.

who did not respond to the critical slide were also excluded. This left a final sample of 67 participants (33 female, 34 male; M_{age} = 39.33, min = 19, max = 83, *SD* = 13.04) for analysis.

6.4.1.2 Procedure and materials. The materials for Study 11 were largely the same as those used in Study 10. The primary difference was the modification of the Anne vignette to include a direct instruction to view the moral judgement task from Anne's perspective. The revised vignette read as follows:

> Anne is a student of philosophy. She generally shows a good understanding of the subject matter, and this is reflected in her grades. Sometimes, however, she adopts a position on an issue in class and struggles (or fails) to defend it when challenged by others.
>
> She is currently taking a course in ethics and has been asked to study the following scenario.
>
> While reading the story on the next page, try to imagine how the philosophy student Anne will judge the actions of the two people.
>
> In particular try to think about reasons she may use to defend her judgement.
>
> Try to think about the story from Anne's perspective rather than your own.

Following the Anne vignette, participants were presented with the Julie and Mark (*Incest*) dilemma. They rated the behaviour, provided reasons for that rating, rated their confidence in that judgement, and then were presented with the counter-arguments. As in previous studies, they were then presented with the critical slide which provided a measure of dumbfounding. Following this participants rated the behaviour again, rated their confidence in their decision, and completed the post-discussion questionnaire (see previous studies and Appendix A, Haidt, Björklund, and Murphy 2000). They then responded to the credulity check questions devised by Royzman, Kim, and Leeman (2015), and answered the three questions relating to the application of the harm principle, devised in Chapter 4.

In addition to completing the short version of the Marlowe-Crowne social desirability scale (Ballard, 1992; Crowne & Marlowe, 1960; Fischer & Fick, 1993; Strahan & Gerbasi, 1972) introduced in Study 10, participants also completed the Need for Closure Scale (Kruglanski et al., 2013), and the Cognitive Reflection Test (CRT: Frederick, 2005; K. S. Thomson & Oppenheimer, 2016; Toplak et al., 2011). The Need for Closure Scale contains 47 questions (e.g., "I'd rather know bad news than stay in a state of uncertainty.") to which participants respond on a 6 point Likert scale, where 1 = *strongly disagree*, and 6 = *strongly agree*. The CRT is a brief test of analytical thinking. It contains three questions, each of which has an answer that seems intuitively correct, but is actually wrong (e.g., If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets? The intuitive answer is 100 minutes, but the correct answer is 5 minutes). Data were collected online using Questback and MTurk. The entire study lasted approximately 20 minutes.

6.4.2 Results and discussion. Fifty participants (74.63 %) rated the behaviour of Julie and Mark as wrong initially. The mean initial rating of the behaviour was, *M* = 2.43, *SD* = 2.07. Fifty-two participants, (77.61 %) rated the behaviour as wrong after viewing the counter-arguments and the critical slide. The mean revised rating of the behaviour was, *M* = 2.31, *SD* = 2.05. A paired samples t-test revealed no difference in rating from time one, (*M* = 2.43, *SD* = 2.07), to time two, (*M* = 2.31, *SD* = 2.05), *t*(63) = 0.65 , *p* = .517. Further analysis revealed that 4 participants changed the valence of their judgement: 1 participant changed their judgement from “wrong” to “neutral”; 2 participants changed their judgement from “right” to “wrong”; and 1 participant changed their judgement from “neutral” to “right”. A chi-squared test for independence revealed no significant association between time of judgement and valence of judgement made, $\chi^2(2, N = 67) = 0.108$, $p = .947$, $V = .03$.

6.4.2.1 Baseline rates of dumbfounding. Participants who selected the admission of not having reasons on the critical slide were identified as dumbfounded. Six participants (8.96%) selected “It's wrong but I can't think of a reason”. Forty nine participants (73.13%) selected “It's wrong and I can provide a valid reason”; and 12 participants (17.91%) selected “There is nothing wrong”. Table 6.2 shows the frequency of each response on the critical slide depending on condition.

Table 6.2: Study 11: Rates of selecting each response to the critical slide depending on Distancing manipulation

	Distancing		Control	
Response to critical slide	N	percent	N	percent
There is nothing wrong.	3	9%	9	28%
It's wrong but I can't think of a reason.	4	11%	2	11%
It's wrong and I can provide a valid reason.	28	80%	21	66%

6.4.2.2 Dumbfounding and coded string responses. Participants who selected “It's wrong and I can provide a valid reason” were required to provide a reason. The reasons provided were coded for unsupported declarations or tautological reasons. An additional 5 participants were identified as dumbfounded following this coding. Two participants provided unsupported declarations (e.g., “Sister and brother should no[sic] engage in sexual activities. It is wrong on all levels.”), 1 participant provided tautological reasons (“Brothers and sisters should not have sex”), and 2 participants provided an unsupported declarations accompanied by a tautological reason (e.g., "They are brother and sisters that just morally wrong"). Taking the coded string responses into account brought the total number of participants identified as dumbfounded to 12 (17.91%).

6.4.2.3 Distancing and eligibility for analysis. As in previous studies, a series of tests was conducted to assess if there was a relationship between distancing and eligibility for analysis. A chi-squared test for independence revealed no significant association between distancing and overall eligibility for analysis, $\chi^2(2, N = 67) = 4.99$, $p = .082$, $V = .27$.

6.4.2.3.1 Distancing and applying the harm principle. Three chi-squared tests for independence revealed no significant association between (a) distancing and applying the harm principle generally, $\chi^2(2, N = 67) = 1.15$, $p = .563$, $V = .13$; (b) distancing and applying the harm principle to boxing, $\chi^2(1, N = 67) < 0.01$, $p > .999$; or (c) distancing and applying the harm principle to rugby, $\chi^2(1, N = 67) < 0.01$, $p > .999$.

6.4.2.3.2 Distancing and endorsing and articulating the harm principle. Two chi-squared tests for independence also revealed no significant association between distancing and the endorsing of the harm principle, $\chi^2(1, N = 67) = 3.48$, $p = .062$, $V = .23$, or articulating the harm principle, $\chi^2(1, N = 67) = 0.02$, $p = .883$, $V = .02$.

6.4.2.3.3 Distancing and endorsing and articulating the norm principle. A final series of chi-squared tests for independence revealed no significant association between distancing and the endorsing of the norm principle, $\chi^2(1, N = 67) < 0.01$, $p > .999$, or the articulating of the norm principle, $\chi^2(1, N = 67) = 0.28$, $p = .597$, $V = .06$.

6.4.2.4 Distancing and responses to critical slide. The responses to the critical slide for the experiment group and the control group were analysed separately. In the distancing group, 3 participants (8.57%) "There is nothing wrong", 4 participants (11.43%) selected “It's wrong but I can't think of a reason”, and 28 participants (80%) selected “It's wrong and I can provide a valid reason”. In the control group, 9 participants (25.71%) selected “There is nothing wrong”, 2 participants (5.71%) selected “It's wrong but I can't think of a reason”, and 21 participants (60%) selected “It's wrong and I can provide a valid reason”. A chi-squared test for independence revealed no significant association between experimental condition and response to the critical slide, $\chi^2(2, N = 67) = 4.54, p = .103, V = .26$. The observed power was .46. Figure 6.2 shows the responses to the critical slide depending experimental condition.

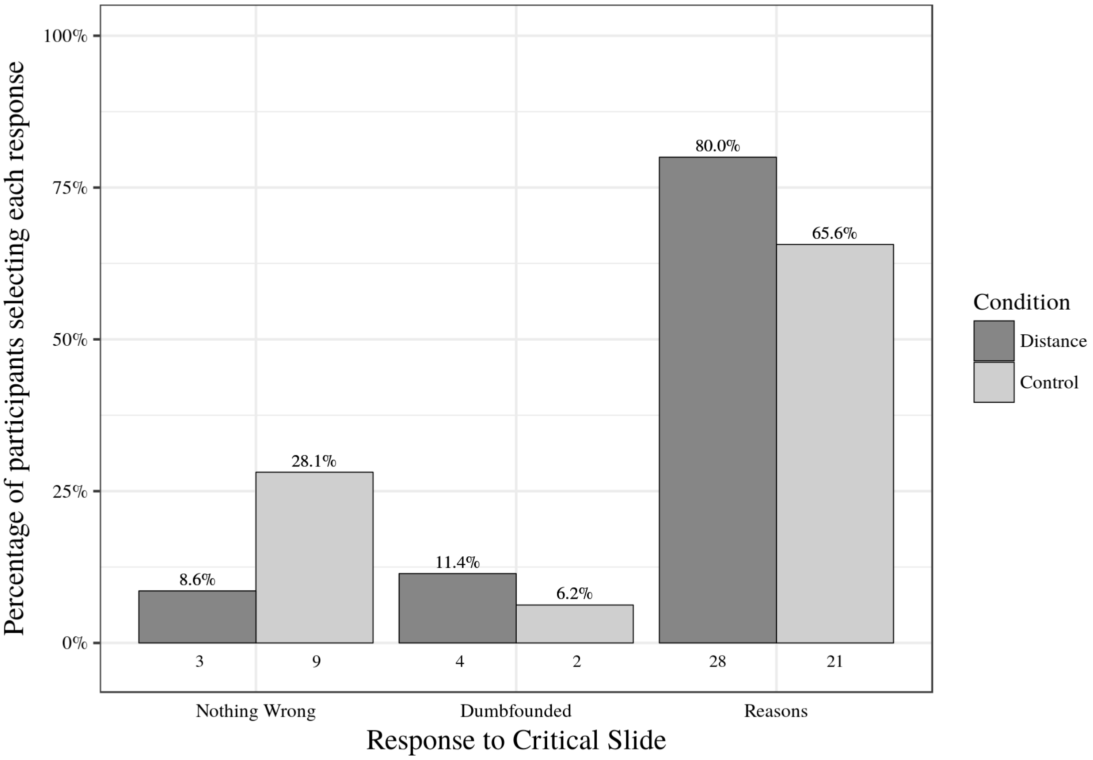

Figure 6.2: Study 11: Response to critical slide and distancing manipulation

Overall rates of dumbfounding in Study 11 were lower than in previous studies. Study 11 was the first study to use the Need for Closure scale (Kruglanski et al., 2013). This scale includes a means for eliminating participants who lied. This was not an option in previous studies. One possible explanation for this is that participants who selected the dumbfounded response in previous studies were not engaging fully with the study. Their lack of engagement went unnoticed. However, the introduction of a stricter measure of engagement with the study, checking for lying on NFC has eliminated these participants in this study. Analysis of the responses of the excluded participants suggest that this is unlikely to be the case. Twenty-six of the original 105 participants were excluded for lying in their responses on the Need for Closure scale. Of these, 5 participants (19%) selected “It’s wrong but I can’t think of a reason”. This is similar to rates of dumbfounding observed for the entire samples in previous studies, and as such it is unlikely that the existence of dumbfounding can be attributed to a small number of disengaged participants. Furthermore, dumbfounding was still present when these lying participants were eliminated.

6.4.2.5 Distancing and providing reasons. A second analysis included the coded string responses in the analysis. In the distancing group, 3 participants (8.57%) selected “There is nothing wrong”; 8 participants (22.86%) presented as dumbfounded by the selecting of “It's wrong but I can't think of a reason”, or failing to provide a reason when asked; and 23 participants (65.71%) successfully provided a reason. In the control group, 9 participants (25.71%) selected “There is nothing wrong”; 4 participants (11.43%) presented as dumbfounded by the selecting of “It's wrong but I can't think of a reason”, or failing to provide a reason when asked; and 19 participants (54.29%) successfully provided a reason. A chi-squared test for independence revealed no significant association between experimental condition and response to the critical slide, $\chi^2(2, N = 67) = 4.66, p = .097, V = .26$. Figure 6.3 shows the responses to the critical slide depending on the distancing manipulation including the coded string responses.

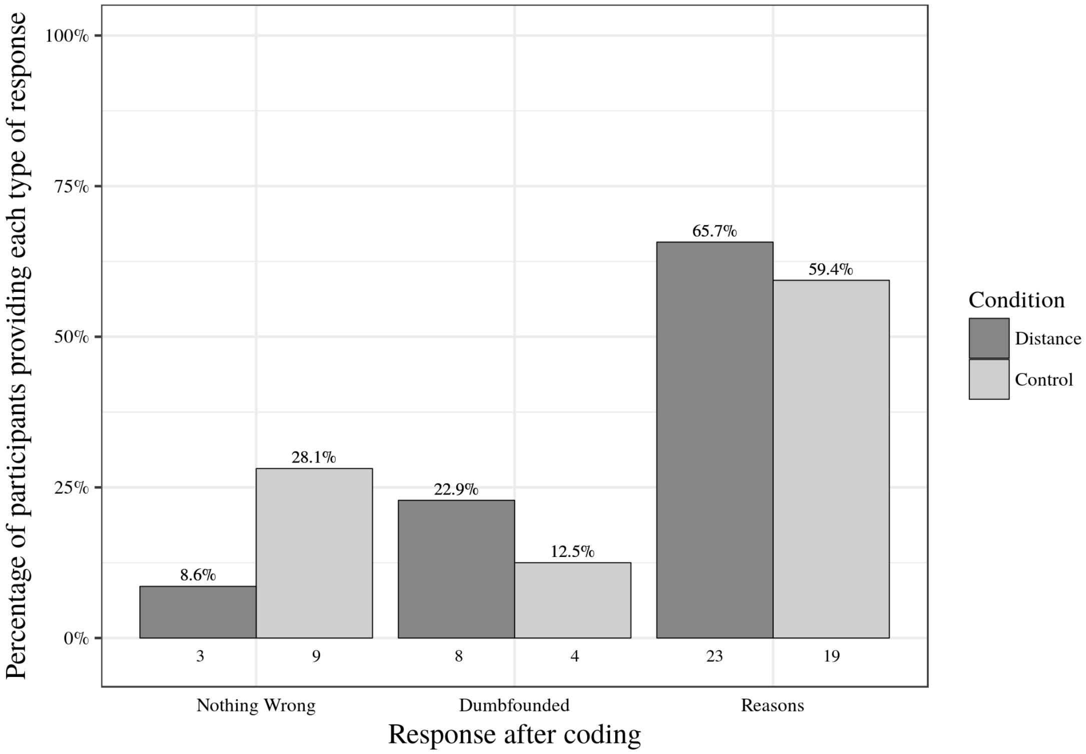

Figure 6.3: Study 11: Providing reasons and distancing manipulation

6.4.2.6 Individual differences and dumbfounding. Variation in the individual difference variables was investigated. There were three individual difference variables measured: Social Desirability, Need for Closure, and score on the Cognitive reflection test and this analysis is exploratory. These three variables were included as potential predictors of dumbfounded responding in a multinomial logistical regression model. We hypothesised that higher CRT scores would be associated with higher rates of providing reasons. We also hypothesised that higher Need for Closure scores would be associated with higher rates of dumbfounded responding. Overall, the model did not reveal any significant association between any of the individual difference variables and dumbfounded responding, $\chi^2(6, N = 67) = 5.675$, $p = .461$. The observed power was .38. Table 6.3 shows the mean scores for each variable of interest depending on score on the critical slide.

Table 6.3: Mean responses to each individual difference variable depending on response to critical slide (Study 11)

	Nothing wrong	Dumbfounded	Reason
Cognitive Reflection Test	1.65	1.27	1.49
Need for Closure	164.42	161.00	162.53
Social Desirability	3.59	4.50	4.84

6.4.3 Study 11 discussion. Study 11 did not provide evidence in support of the dual-process explanation of moral dumbfounding. It was hypothesised that a distancing manipulation would facilitate the identification of reasons lead to higher rates of participant providing reasons than in the control group. A stronger distancing manipulation was employed than in Study 10, whereby an explicit instruction to take the perspective of a third person was included. The results of Study 11 were largely in line with the results of Study 10. The distancing manipulation did not lead to a significant increase in rates of providing reasons. As in Study 10, the individual difference variable social desirability did not appear to be related to dumbfounding. Two additional individual difference variables were included in Study 11, CRT and Need for Closure. Neither appeared to be related to judgements made or ability to provide reasons for the judgements.

One strength of Study 11 over previous studies was the inclusion of the Need for Closure scale, in particular the usefulness of this scale for screening the data prior to analysis. A number of participants were excluded prior to analysis based on their responses to particular items on the Need for Closure scale (that indicated they were lying). This removed a possible source of significant error from the final dataset that was analysed. This screening along with the removal of participants who provided nonsense text in the open-ended responses, and participants who did not complete the full study meant that the final sample for analysis was much smaller than the sample collected.

6.5 Study 12: Dumbfounding as Incomplete Mental Models

Studies 6-11 tested two predictions a generalised dual-process explanation of moral dumbfounding in that the predictions that were tested were grounded in broader principles of dual-process approaches rather than specific predictions of individual dual-process theories of moral judgement. In Studies 6-9 it was hypothesised that inhibiting deliberation, through the introduction of a cognitive load task, would lead to a reduction in the rates of providing reasons. This experimental manipulation failed to consistently reduce rates of providing reasons. The studies in the current chapter employ manipulations intended to facilitate the providing of reasons. One advantage of this type of manipulation is that the intended outcome of this manipulation is congruent with the intentions of participants, participants are attempting to identify reasons and the aim of the manipulation is to aid them in that. Conversely, the aim of the manipulation in the previous chapter was to stifle participants' attempts to identify reasons, making the task more difficult. By making an already difficult more difficult, it is possible that participants may disengage from the task (getting bored as a result of the task difficulty, e.g., Acee et al., 2010).

The results of Studies 10 and 11 were similar to the results of Study 9, the predicted trend in responding appeared to be present but the result was not statistically significant. Having failed to obtain convincing evidence either for or against a generalised dual-process (Research Question 2.1.2) explanation of moral dumbfounding across 6 studies, it seems increasingly likely that moral dumbfounding is more complex than the generalised conflict in dual-processes explanation allows. Model theory offers an explanation of moral dumbfounding that is consistent with, but goes beyond, the explanation offered by dual-process approaches more generally. The key hypothesis underlying all these studies is that the identification of reasons for a judgement is grounded in deliberation to a greater extent than providing a dumbfounded response or changing a judgement. Model theory attempts to provide an account of the process of the conscious reasoning involved in deliberation, and therefore may provide additional insight into the processes by which people may identify reasons for their judgement.

As noted in the introduction of this chapter, model theory posits that people reason about deontic principles using incomplete mental models. It is hypothesised that in a moral dumbfounding task one consequence of the incompleteness of a mental model is an absence of reasons for a judgement – specifically reasons that are not subsequently refuted as part of the paradigm are not included in the mental model. It has been shown extensively that variations in instructions or descriptions of a reasoning problem provided to participants can alter the content of a mental model. This malleability of mental models means that it may be possible to vary the instructions in a moral dumbfounding study such that participants construct a mental model that includes a counter-argument immune reason for their judgement.

Study 12 tests this hypothesis. The materials are largely the same as those used in Study 11 with a minor alteration to the Anne vignette to enable a prompt for a reason to be presented to participants in a plausible and controlled manner. Given that distancing was found not to influence rates of dumbfounding/providing reasons, the version of the Anne vignette used in the manipulation group (distancing) in Study 11 provided was used as a control in Study 12. The only difference between the conditions in Study 12 was the inclusion of a prompt for a reason in the Anne vignette for one group. The reason provided in the prompt related to the possibility of "unseen consequences", or damage to relationships/feelings of regret following their actions. The individual difference variables social desirability, CRT, and NFC were also recorded in Study 12.

6.5.1 Method.

6.5.1.1 Participants and design. Study 12 was a between-subjects design with social desirability, CRT and NFC additionally measured as potential correlate and moderator variables. The dependent variable was response to the critical slide. The independent variable was model prompt with two levels: present and absent. An initial sample of 210 participants[15] (126 female, 80 male; M_{age} = 38.80, min = 20, max = 73, *SD* = 12.25) was collected. Participants in this sample were recruited through MTurk. Participation was voluntary and participants were paid $US 0.50 for their participation. Participants were recruited from English speaking countries or from countries where residents generally have a high level of English (e.g., The Netherlands, Denmark, Sweden). As in Study 11, participants were excluded for including nonsense text in the open-ended response questions, or for lying in the need for closure scale (a combined score of > 15 on selected items). This left a final sample of 118 participants (71 female, 46 male; M_{age} = 39.00, min = 20, max = 71, *SD* = 12.13) for analysis.

6.5.1.2 Procedure and materials. The materials for Study 12 were largely the same as those used in Study 11; the materials for the control group in Study 12 were unchanged from the materials that were used for the manipulation group was in

[15] As in Studies 10 and 11, a priori power analysis indicated that in order to detect a large effect size (*V* = .5) with 80% power, a sample of 39 participants was required. In order to detect a medium effect size (*V* = .3) with 80% power a sample of 107 participants was required.

Study 11. A single change was made to the materials for the experimental manipulation in Study 12. The Anne vignette was modified to include a prompt for a reason that may be used to justify a judgement condemning the behaviour of Julie and Mark. The revised vignette read as follows (prompt in italics):

> Anne is a student of philosophy. She generally shows a good understanding of the subject matter, and this is reflected in her grades. Sometimes, however, she adopts a position on an issue in class and struggles (or fails) to defend it when challenged by others.
>
> She is currently taking a course in ethics and has been asked to study the following scenario.
>
> *In the ethics course, Anne is learning to think about unseen or unanticipated consequences of actions. For instance, an action may have a positive consequence for one individual, but the action may also unexpectedly result in damaging relationships with others. Or an action may not have any immediate negative consequences, but over time feelings of regret and guilt may impact negatively on a person's life.*
>
> While reading the story on the next page, try to imagine how the philosophy student Anne will judge the actions of the two people.
>
> In particular try to think about reasons she may use to defend her judgement.
>
> Try to think about the story from Anne's perspective rather than your own.

Other than this change, the materials in Study 12 were identical to the control group in Study 11, and the experiment ran in the same way. Again, data were collected online using Questback and MTurk.

6.5.2 Results and discussion. Ninety five participants (80.51 %) rated the behaviour of Julie and Mark as wrong initially. The mean initial rating of the behaviour was, *M* = 2.19, *SD* = 1.62. Ninety three participants, (78.81 %) rated the behaviour as wrong after viewing the counter-arguments and the critical slide. The mean revised rating of the behaviour was, *M* = 2.39, *SD* = 1.73. A paired samples t-test revealed no difference in rating from time one to time two, *t*(117) = -1.69, *p* = .094. Further analysis revealed that 17 participants changed the valence of their judgement: 4 participants changed their judgement from “wrong” to “neutral”; 4 participants changed their judgement from “wrong” to “right”; 5 participant changed their judgement from “right” to “wrong”; 3 participants changed their judgement from “neutral” to “right”; and 1 participant changed their judgement from “neutral” to “wrong”. A chi-squared test for independence revealed no significant association between time of judgement and valence of judgement made, χ^2(2, *N* = 118) = 0.175, *p* = .916, *V* = .04.

6.5.2.1 Baseline rates of dumbfounding. Participants who selected the admission of not having reasons on the critical slide were identified as dumbfounded. Seventeen participants (14.41%) selected “It's wrong but I can't think of a reason”. Eighty-one participants (68.64%) selected “It's wrong and I can provide a valid reason”; and 20 participants (16.95%) selected “There is nothing wrong”. Table 6.4 shows the frequency of each response on the critical slide depending on condition. The rate selecting the admission of having no reasons returned to levels expected based on previous studies following the unexpectedly low rates observed in Study 11.

Table 6.4: Study 12: Rates of selecting each response to the critical slide depending on model prompt manipulation

	Distance		Model Prompt	
Response to critical slide	N	percent	N	percent
There is nothing wrong.	12	21%	8	13%
It's wrong but I can't think of a reason.	9	16%	8	13%
It's wrong and I can provide a valid reason.	37	64%	44	73%

6.5.2.2 Dumbfounding and coded string responses. Participants who selected “It's wrong and I can provide a valid reason” were required to provide a reason. The reasons provided were coded for unsupported declarations or tautological reasons. An additional 12 participants were identified as dumbfounded following this coding. Six participants provided unsupported declarations (e.g., “Its just wrong”; “Incest is wrong. So very wrong...”), four participants provided tautological reasons (e.g., “The two are brother and sister this is incest”), one participant provided an unsupported declaration accompanied by a tautological reason (“It is morally wrong for siblings to have sex.”), and one participant provided an unsupported declaration accompanied by a statement with no reason (“If they enjoyed themselves and have no regrets they are disgusting! It does not matter that they were both consenting adults. It is wrong period.”). Taking the coded string responses into account brought the total number of participants identified as dumbfounded to 29 (24.58%).

6.5.2.3 Model prompt and eligibility for analysis. As in the Studies 6 through 11, a series of tests was conducted to assess if there was a relationship between the model prompt and eligibility for analysis. A chi-squared test for independence revealed no significant association between experimental manipulation and overall eligibility for analysis, $\chi^2(1, N = 118) = 2.16, p = .142, V = .14$.

6.5.2.3.1 Model prompt and applying the harm principle. Three chi-squared tests for independence revealed no significant association between (a) the model prompt and applying the harm principle generally, $\chi^2(2, N = 118) = 0.33, p = .847, V = .05$; (b) the model prompt and applying the harm principle to boxing, $\chi^2(1, N = 118) = 0.28, p = .594, V = .05$; or (c) experimental manipulation and applying the harm principle to rugby, $\chi^2(1, N = 118) < .01, p > .999$.

6.5.2.3.2 Model prompt and endorsing and articulating the harm principle. A chi-squared tests for independence revealed no significant association between the model prompt and the endorsing of the harm principle, $\chi^2(1, N = 118) = 1.72, p = .190, V = .12$. There appeared to be a difference in the articulating of the harm principle, with 17% participants in the control condition mentioning harm and 34% participants in the model prompt condition mentioning harm, however this difference was not significant, $\chi^2(1, N = 118) = 3.22, p = .073, V = .17$. If a significant difference was found it may have provided evidence that people responded to the model prompt and incorporated it into their mental models (recall that the reasons the prompt were consistent with the harm principle, citing concerns of unforeseen consequences).

6.5.2.3.3 Model prompt and endorsing and articulating the norm principle. A final series of chi-squared tests for independence revealed no significant association between the model prompt and the endorsing of the norm principle, $\chi^2(1, N = 118) = 0.11, p = .740, V = .03$. There was a significant association between the model prompt and the articulating of the norm principle, $\chi^2(1, N = 118) = 5.96, p = .015, V = .22$, with 68% participants in the control condition mentioning norms and 46% participants in the model prompt condition mentioning norms. However, no significant difference in mentioning the harm principle was found so the reasons for this difference are unknown.

6.5.2.4 Model prompt and responses to critical slide. The responses to the critical slide for the experiment group and the control group were analysed separately. In the experimental group, eight participants (13.33%) selected “There is nothing wrong”, eight participants (13.33%) selected “It's wrong but I can't think of a reason”, and 44 participants (73.33%) selected “It's wrong and I can provide a valid reason”. In the control group, twelve participants (20%) selected “There is nothing wrong”, nine participants (15%) selected “It's wrong but I can't think of a reason”, and 37 participants (61.67%) selected “It's wrong and I can provide a valid reason”. A chi-squared test for independence revealed no significant association between experimental condition and response to the critical slide, $\chi^2(2, N = 118) = 1.43$, p = .489, V = .11. The observed power was .17. Figure 6.4 shows the varying responses to the critical slide depending experimental condition.

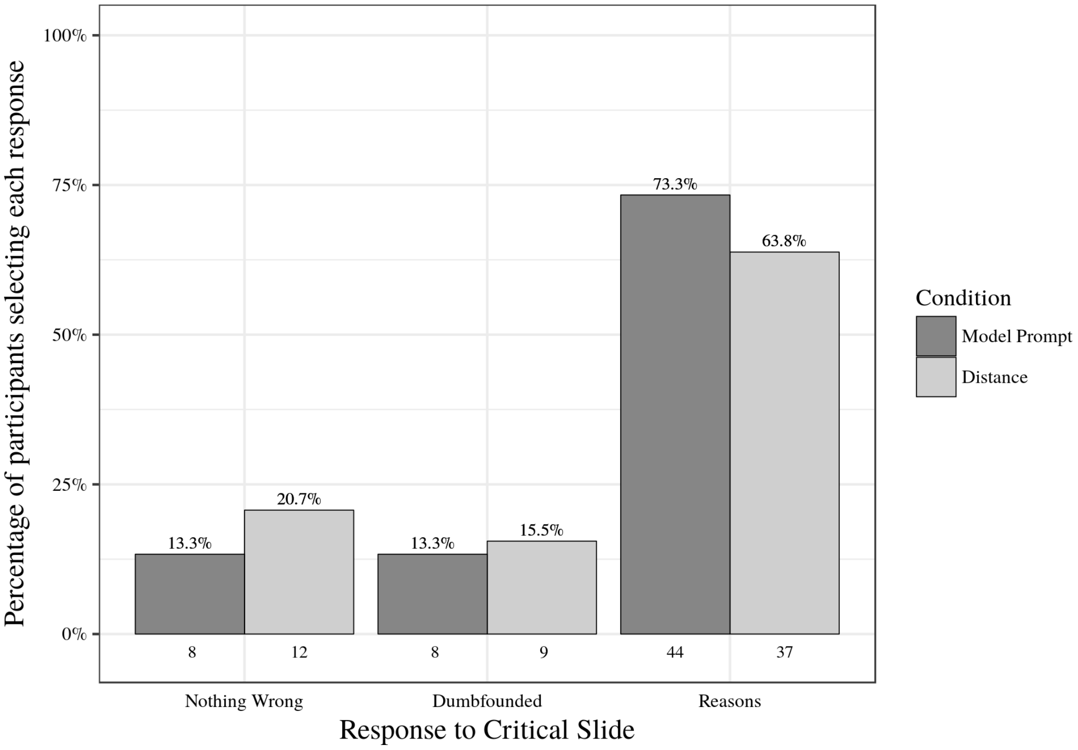

Figure 6.4: Study 12: Response to critical slide and model prompt manipulation

As in the distancing studies, the expected trend was observed, but this finding was not statistically significant, and Study 12 did not provide evidence for the mental models explanation of moral dumbfounding: a failure to construct complete models. It is possible that this failure to provide evidence for this explanation is due to limitations in the methods, however, Study 12 is the third study in this Chapter to fail to manipulate dumbfounding. It is becoming increasingly apparent that moral dumbfounding is more robust than predicted by both a generalised dual-process explanation and an explanation informed by model theory.

6.5.2.5 Model prompt and providing reasons. A second analysis included the coded string responses in the analysis. In the experimental manipulation group, 8 participants (13.33%) selected “There is nothing wrong”; sixteen participants (26.67%) presented as dumbfounded by the selecting of “It's wrong but I can't think of a reason”, or failing to provide a reason when asked; and 36 participants (60%) successfully provided a reason. In the control group, twelve participants (20%) selected “There is nothing wrong”; thirteen participants (21.67%) presented as dumbfounded by the selecting of “It's wrong but I can't think of a reason”, or failing to provide a reason when asked; and 33 participants (55%) successfully provided a reason. A chi-squared test for independence revealed no significant association between experimental condition and providing reasons, $\chi^2(2, N = 118) = 1.21, p = .547, V = .10$. The observed power was .15. Again, dumbfounding was not significantly influenced by experimental manipulation, illustrating that dumbfounding is more robust than predicted by the explanation adopted in this and the previous chapter.

6.5.2.5. Individual differences and dumbfounding. As in Study 11, three individual difference variables were measured: Social Desirability, Need for Closure, and CRT. An exploratory analysis was conducted in which all three variables were included as potential predictors in a multinomial logistical regression model. As in Study 11 we was hypothesised that higher CRT scores would be associated with higher rates of providing reasons. We also hypothesised that higher Need for Closure scores would be associated with higher rates of dumbfounded responding. Overall, the model was significantly associated with dumbfounded responding, $\chi^2(6, N = 118) = 13.51, p = .036$. The observed power was .80. The model as a whole explained between 14.1% (Cox and Snell R square) and 17.3% (Nadelkerke R squared) of the variance in responses to the critical slide. As shown in Table 6.5, the only variable that made a unique significant contribution to the model was CRT. As CRT increased, participants were significantly more likely to select "there is nothing wrong" than to present as dumbfounded, Wald = 3.942, $p = .048$, odds ratio = 2.24, 95% CI [1.01, 4.95].

Table 6.5: Study 12 – Multinomial logistic regression predicting responses to the critical slide where a dumbfounded response is the referent in each case.

								95% CI for Odds Ratio	
		B	S.E.	Wald	*df*	*p*	Odds Ratio	Lower	Upper
Social desirability	Nothing wrong	.103	.165	.392	1	.532	1.108	.803	1.530
	Reasons	.029	.136	.046	1	.830	1.030	.789	1.344
Need for Closure	Nothing wrong	-.020	.019	1.139	1	.286	.980	.945	1.017
	Reasons	.002	.015	.024	1	.877	1.002	.974	1.031
CRT	Nothing wrong	.804	.406	3.942	1	.048*	2.235	1.008	4.953
	Reasons	-.131	.265	.243	1	.622	.877	.522	1.476

*Note: * = sig. at p < .05*

This finding is consistent with the link between CRT and similar moral judgements in the existing literature (e.g., Royzman, Landy, et al., 2014). A follow-up test found that CRT was indeed significantly correlated with initial judgements, $r = .19$, $N = 118$, $p = .038$, and revised judgements following the critical slide, $r = .26$, $N = 118$, $p = .005$. It seems that the link between CRT and response to the critical slide is illustrative of a link between CRT and valence of judgement, as opposed to being predictive of ability to provide reasons.

6.5.3 Study 12 discussion. The primary aim of Study 12 was to facilitate the providing of reasons for a moral judgement by the inclusion of a prompt for a reason prior to the presenting of the moral scenario. The rates of providing reasons for judgements appeared to be higher in the model prompt group than in the control group, however this difference was not significant. This may be due to methodological issues, e.g., the dependent variable is a categorical/nominal variable, unsuited to identifying weak or small effects; or the use of online samples (e.g., Crump et al., 2013; Goodman et al., 2013). It may also be due to the limited observed power of the study. Dumbfounding did not appear to be reliably related to any of the individual difference variables, CRT, Need for Closure, or social desirability. CRT appeared to predict response to the critical slide, but this is likely confounded by the link between CRT the valence of judgement, such that participants who scored higher in CRT were more likely to select “There is nothing wrong” than to select a dumbfounded response. Study 12 corroborated the results of the previous studies, that moral dumbfounding is robust and resistant to manipulation. However, despite this apparent stability, rates of dumbfounding can fluctuate seemingly without any reason, as demonstrated in Study 10. The remainder of this Chapter will discuss the results of Studies 10-12 together.

6.6 Combined Results and Discussion

6.6.1 Facilitating the identification of reasons. The aim of each of the studies to reduce rates of dumbfounding by facilitating the identification of reasons for a judgement, firstly by facilitating deliberation through distancing, and secondly by encouraging participants to include a reason for their judgement in their mental model providing a prompt for a reason prior to presenting the moral scenario. Individually, each of these studies failed in their stated aim. They each appeared to show a general trend in the predicted direction, though none of these findings was statistically significant. Other than the limited observed power of the studies conducted, there is reason to suspect that it would be premature to reject the stated hypotheses outright based on these data.

Firstly consider the rates of providing reasons in Study 12, the final study. In the group who received the prompt, 73% of participants provided reasons for their judgements. Recall that the control used in Study 12 was the distancing vignette from Study 11 that did not lead to any significant differences in rates of dumbfounding when compared to a control with no manipulation, However there appeared to be a trend in responses in the predicted direction. What was essentially demonstrated across Studies 11 and 12 was the following: a reason prompt is not significantly different from distancing; and that distancing is not significantly different from no manipulation. However, both these appeared to display a trend in the predicted direction. It is possible that the combination of distancing and a reason prompt may be different from no manipulation. This test was not conducted because it involves treating separate independent variables as a single independent variable which is inappropriate. If such a study was conducted it is likely to reveal a significant difference in responding between no manipulation and two complimentary manipulations, each designed to facilitate the identification of reasons.

It is hypothesised here that the trends observed in Studies 10-12, though not statistically significant, are weak effects that cannot be identified due to weaknesses in the methods used (particularly the categorical nature of the dependent variable: response to the critical slide), and the limited power of the studies conducted. Support for this claim can be found in the aggregated analysis of Studies 10 and 11. Individually both studies failed to identify a significant difference in providing reasons between the distancing group and the no manipulation group. However when the data sets from Studies 10 and 11 are combined a significant difference in rates of providing reasons between distancing and the no manipulation control is found, $\chi^2(2, N = 187) = 6.01, p = .0495$. The observed power was .58. The rates of providing each response to the critical slide depending on distancing manipulation are displayed in Table 6.6.

Table 6.6: Studies 10 and 11: Distancing and rates of providing each response to the critical slide

	Distancing		Control	
Response to critical slide	N	percent	N	percent
There is nothing wrong.	18	19%	32	35%
It's wrong but I can't think of a reason.	15	16%	11	12%
It's wrong and I can provide a valid reason.	62	65%	49	53%

This significant effect for distancing when the results of both studies are combined suggests that using a distance manipulation as a control in Study 12 may have been inappropriate. That said, it is hard to conceive an alternative design that would be both plausible and controlled. The reason prompt appeared to result in a similar trend in responding to the critical slide relative to distancing as was found between distancing and the control with no manipulation. Based on this, and that the noticeably higher rate of providing reasons in the reason prompt group when compared with previous studies (only 2 samples provided reasons at a rate of greater than 70%) it is reasonable to suggest that the combination of a reason prompt and distancing facilitate the identification of reasons, as predicted by dual-process theories and model theory.

6.6.2 Individual differences across studies in Chapter 6. Studies 10, 11, and 12 were combined to investigate if social desirability influenced judgements, or ability to provide reasons. Studies 11 and 12 were analysed together to investigate if dumbfounded responding, or ability to provide reasons was related to the individual difference variables, social desirability, Need for Closure, and CRT. All analyses were exploratory.

A multinomial logistical regression revealed no statistically significant association between social desirability and response to the critical slide across Studies 10-12, $\chi^2(2, N = 305) = .09, p = .954$. The observed power was .06.

A second analysis combining Studies 11 and 12 was conducted. All three variables social desirability, Need for Closure, and CRT were included as potential predictors in a multinomial logistical regression model. Overall, the model was significantly associated with dumbfounded responding, $\chi^2(6, N = 185) = 14.88, p = .$ 021. The observed power was .84. The model as a whole explained between 10.9% (Cox and Snell R square) and 13.5% (Nadelkerke R squared) of the variance in responses to the critical slide. Again, as in Study 12, the only variable that made a unique significant contribution to the model was CRT. This is shown in Table 6.7. As CRT increased, participants were significantly more likely to select “there is nothing wrong” than to present as dumbfounded, Wald = 5.98, $p = .014$, odds ratio = 2.25, 95% CI [1.17, 4.29].

Table 6.7: Studies 11 & 12 – Multinomial logistic regression predicting responses to the critical slide where a dumbfounded response is the referent in each case.

								95% CI for Odds Ratio	
		B	S.E.	Wald	*df*	*p*	Odds Ratio	Lower	Upper
Social desirability	Nothing wrong	.024	.136	.032	1	.859	1.024	.785	1.336
	Reasons	.039	.115	.114	1	.736	1.039	.830	1.302
Need for Closure	Nothing wrong	-.006	.015	.153	1	.696	.994	.996	1.024
	Reasons	-.001	.013	.002	1	.963	.999	.975	1.024
CRT	Nothing wrong	.809	.331	5.984	1	.014*	2.245	1.174	4.290
	Reasons	-.072	.236	.093	1	.761	.931	.585	1.479

*Note: * = sig. at p < .05*

The results of the combined analysis here are consistent with what was found in Study 12. Again, the influence of CRT is likely to be due to the link between CRT and valence of judgement (e.g., Royzman, Landy, et al., 2014 have shown that people who score higher in CRT are more likely to judge incest as ok) as opposed to being related to a persons susceptibility to dumbfounding. Social desirability and Need for Closure do not appear to be related to dumbfounded responding.

6.7 General Discussion

The studies described in this chapter investigated the research question identified in Chapter 2: “Can the existence of moral dumbfounding be adequately explained by (2.1.2) dual-process approaches or (2.1.3) Model theory. Initial analysis indicates that the studies described in this chapter failed to identify reliable influences on moral dumbfounding. On aggregate, in line with dual-process predictions, the providing of reasons for a judgement appears to be weakly facilitated by distancing. It also appears that a combination of distancing and a reason prompt appear to facilitate the providing of reasons, in line with mental model theory.

Two findings of particular interest are (a) the robustness of responses in the dumbfounding paradigm; and (b) the unpredictability of responses in the dumbfounding paradigm. Taking each in turn, firstly, the rates of providing each response to the critical slide remains remarkably stable across different studies and different manipulations. The majority of people provide reasons for their judgements, and, generally the dumbfounded participants are in the minority. The rates of participants rating the behaviour as "not wrong" is generally lower than the rates of providing reasons and higher than the rates of dumbfounding. Secondly, though the relative proportions remain relatively stable, the specific proportions can fluctuate unpredictably. The results of Study 10 illustrate this. It is apparent from the studies in both Chapters 5 and 6, that dual-processes and mental models may be related to dumbfounding in some way, however these studies did not provide conclusive evidence in support of these explanations, in that there was variability observed in the direction predicted, but the this variability was not always statistically significant.

Two explanations of moral dumbfounding were tested in this Chapter, and conclusive evidence for or against either explanation was not found. Limitations in the methods have been identified, however the results of the studies in this chapter have highlighted limitations in the explanatory power of existing theories of moral judgement. The studies in the current chapter and Chapter 5 investigated related claims and as such the results should be analysed together. Chapter 7 will provide a discussion of the implications of all the studies conducted, in Chapters 5 and 6, but also Chapters 3 and 4, in an attempt to identify what moral dumbfounding really is, and what it can tell us about how we make moral judgements.

7 Chapter 7 – General Discussion

A series of 12 studies investigated the strength of evidence for moral dumbfounding, and three possible explanations of dumbfounding. This first area of investigation is of particular importance because, as identified in Chapter 1, dumbfounding is widely-cited in support of theories of moral judgement. However, evidence for the phenomenon has been limited to a single study (containing a final sample of 30), that is unpublished in peer-reviewed form, and has not been directly replicated (Haidt et al., 2000). Specific methodological limitations of this original study were identified in Chapter 2.

Three studies in Chapter 3 provided evidence that dumbfounding is a real phenomenon that can be elicited in both an interview setting and using a computerised task. Chapter 3 also served to develop materials for eliciting and studying dumbfounding. Having identified dumbfounding as a genuine phenomenon, various explanations of dumbfounding were tested.

Two studies in Chapter 4 examined the claim that participants presenting as dumbfounded do have reasons for their judgements and that dumbfounding occurs due to social pressure to not appear stubborn (Royzman et al., 2015). It was found (a) that people do not reliably articulate the reasons that were purported to be guiding judgements; and (b) that people do not consistently apply the principle underlying one of these reasons, casting further doubt on the claim that this principle guides judgements. When the articulation and applying of these reasons was accounted for, dumbfounded responding was found.

Four studies in Chapter 5 and two Studies in Chapter 6 assessed two predictions of a dual-process explanation of moral dumbfounding. The results of these studies is inconclusive, the null hypothesis can neither be rejected nor accepted with any confidence. A final study in Chapter 6 investigated dumbfounding from a mental models perspective. Again, the results were inconclusive.

The remainder of this chapter will discuss the implications of these studies for our understanding of moral judgements more generally. Two features of dumbfounding that have been identified will be discussed: robustness and variability. The theoretical implications of these studies will then be discussed. Firstly, the general implications of the existence of moral dumbfounding for the broader moral psychology literature will be discussed. The possible explanations tested in Chapters 5 and 6 will then be examined. Finally, the role that moral dumbfounding may have in furthering our understanding of the making of moral judgements will be discussed.

7.1 Robustness of Moral Dumbfounding

One of the most striking findings of the research conducted in the preceding chapters is that moral dumbfounding is remarkably robust. Evidence for dumbfounding was found across all studies conducted. Twelve studies, with a total (analysable) sample of $N = 1180$, illustrated that when pressed to justify their judgements with reasons, some people fail and admit to not having reasons. Using the calculation employed by Royzman et al. (2015), I assessed whether observed rates of dumbfounding in each study were significantly greater than zero. The results are displayed in Table 7.1. A meta-analysis on these results was conducted (weighted using the square root of N; e.g., Zaykin, 2011). As expected, rates of dumbfounding across 12 studies were significantly greater than zero, $z = 22.31$, $p < .001$ (Stouffer's Z-score method), indicating that dumbfounded responding is a phenomenon that can be reliably evoked in a laboratory setting.

Table 7.1: Results of z tests for each study (and each scenario), testing if rates of dumbfounding (D) were significantly greater than 0.

Study	Scenario	D = rates of dumbfounding	*N*	P(D>0)
Study 1	Heinz	0	31	$p > .999$
	Trolley	3	31	$p = .076$
	Incest	18	31	$p < .001$
	Cannibal	11	31	$p < .001$
Study 2	Heinz	45	72	$p < .001$
	Trolley	45	72	$p < .001$
	Incest	54	72	$p < .001$
	Cannibal	46	72	$p < .001$
Study 3a	Heinz	13	72	$p < .001$
	Trolley	14	72	$p < .001$
	Incest	18	72	$p < .001$
	Cannibal	14	72	$p < .001$
Study 3b	Heinz	12	101	$p < .001$
	Trolley	16	101	$p < .001$
	Incest	16	101	$p < .001$
	Cannibal	19	101	$p < .001$
Study 4	Incest	20	110	$p < .001$
Study 5	Incest	21	111	$p < .001$
Study 6	Incest	13	66	$p < .001$
Study 7	Incest	20	100	$p < .001$
Study 8	Incest	22	163	$p < .001$
Study 9	Incest	20	156	$p < .001$
Study 10	Incest	20	120	$p < .001$
Study 11	Incest	6	67	$p = .012$
Study 12	Incest	17	118	$p < .001$

Dumbfounded responding also appears to be remarkably resistant to manipulation. Studies 4 and 5, in line with rationalist explanations, attempted to eliminate dumbfounding by assessing the degree to which people’s judgements could legitimately be attributed to reasons, however dumbfounded responding persisted.

Studies 10, 11, and 12 attempted to eliminate dumbfounding by facilitating analytical thinking (distance manipulation – Studies 10 and 11) and even providing a prompt for reason that was immune to the counter-arguments (Study 12) and again, dumbfounded responding persisted. Prompting people with a reason failed to prevent people from selecting an admission of not having reasons for their judgements. In Studies 6-9 it was attempted to inhibit the identification of reasons for judgements and potentially increase the frequency of dumbfounding. Patterns of responses appeared to vary as predicted however this finding is not conclusive or convincing, with rates of dumbfounding remaining more stable than expected across Studies 6-9, given the experimental manipulation.

With some notable exceptions (e.g., Study 2, Study 3a, and Study 7), the relative proportions of responses in the dumbfounding paradigm appears to follow a reasonably stable pattern. Generally, the majority of people provide reasons for their judgements and dumbfounded participants are in the minority, with the remaining participants selecting “There is nothing wrong”. The total rates of selecting/providing each type of response across Studies 1-12 are shown in Figure 7.1 (the total number of cases displayed is higher than the total number of participants, as Studies 1 to 3 included 4 scenarios). The high rates of dumbfounding in Study 2 bring the total rates of dumbfounded responding above the rates of selecting “There is nothing wrong”. Given that this was identified as an irregular pattern of responding in Chapter 3, Figure 7.1 includes the rates of selecting/providing each type of response for all studies with the data from Study 2 removed (*N* = 1108).

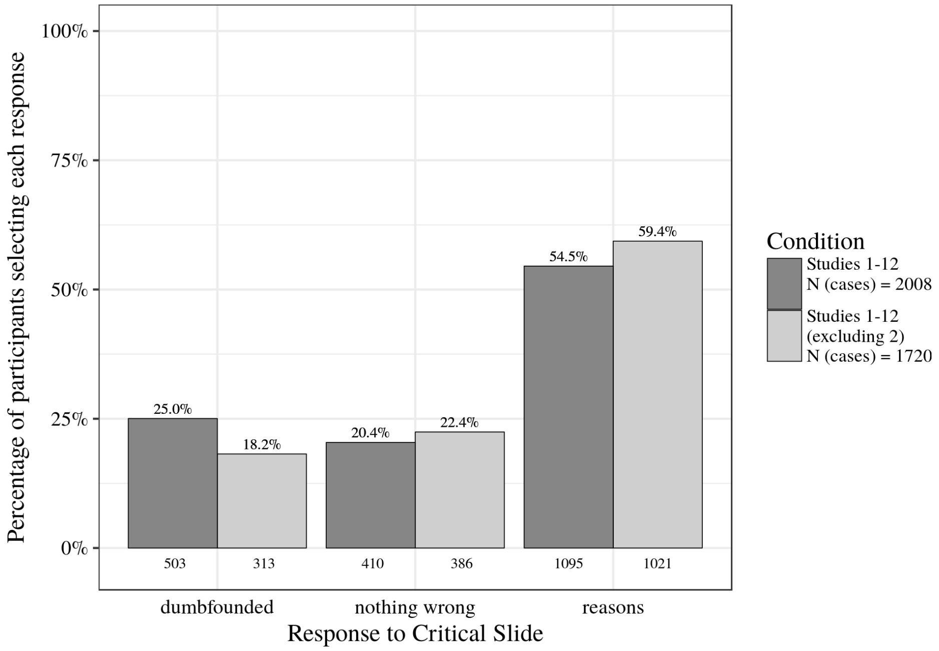

Figure 7.1: Rates of selecting/providing each type of response to the critical slide across studies 1-12

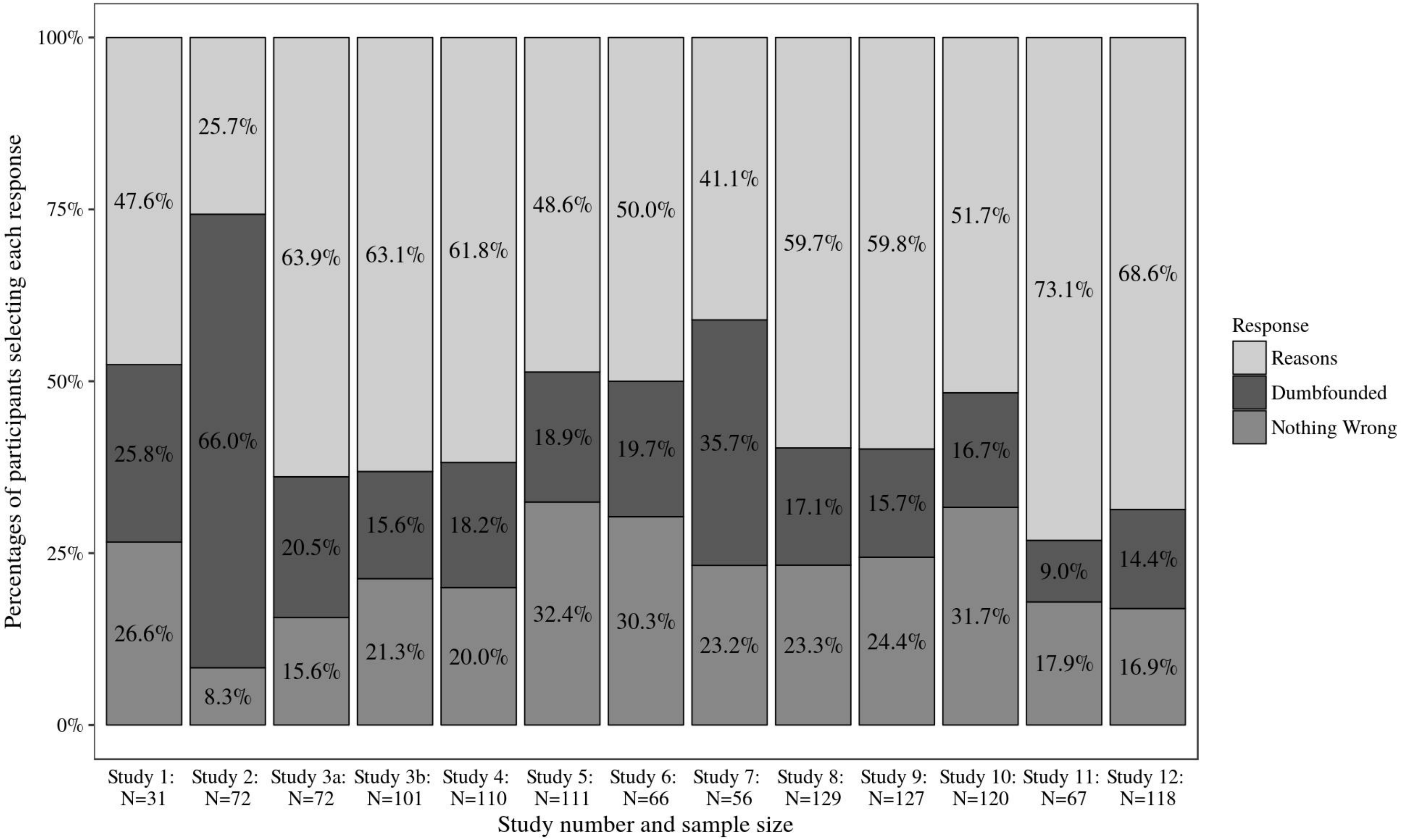

Figure 7.2: Rates of selecting/providing each type of response to the Incest dilemma across Studies 1-12.

The pattern of responses described above identifies the providing of reasons as the most frequent response, this is clearly visible in Figure 7.1. Dumbfounded responding is identified as the least frequent response, and rates of selecting “There is nothing wrong” are slightly higher than rates of dumbfounding. Though this is reversed when the data from Study 2 are included. Any difference between rates of dumbfounded responding and selecting “There is nothing wrong” appears to be negligible. However, it was relatively consistently observed across the vast majority of studies conducted. Figure 7.2 shows the rates of selecting/providing each response for each study individually. To ensure consistency, for Studies 1-3, Figure 7.2 includes rates of responses to the *Incest* dilemma only.

The reasons provided in the *Incest* dilemma across Studies 6 to 9 ($N = 485$) were analysed and coded. Consistent with what was found in Chapter 3, participants did not provide reasons that directly contradicted the information provided in the scenario and counter-arguments, though some participants challenged the validity of the facts presented. The most common reason provided was potential consequences (60%), these consequences were unnamed (4%), related to considerations of psychological harm to Julie and Mark, or their family/friends/future spouses (37%), or the possibility of pregnancy (18%). Religion was the next largest reason provided (14%), followed by social norms (13%), and emotion/disgust (8%). Some participants cited the secrecy/deceit as a reason (6%), while others referred to the act as unnatural (4%), or illegal (4%).

For 9 of the 13 studies displayed the rates of selecting “There is nothing wrong” on the *Incest* scenario are higher than rates of dumbfounding. One of the remaining 4 studies was Study 2, for which unusually high rates of dumbfounding have already been identified. Study 1 also showed higher rates of dumbfounding than selecting “There is nothing wrong”. This was an interview task, and social pressure may have inhibited participants from changing their judgement.

Interestingly Study 7 also demonstrated higher rates of dumbfounding than selecting “There is nothing wrong”. Recall that Study 7 was investigating cognitive load, and that both the control group and the manipulation group completed a memory task (manipulation was task difficulty). This may be indicative of dumbfounding being linked to cognitive capacity and cognitive load, though the studies in Chapter 5 did not clearly show this to be the case.

The final study that does not follow the pattern of responses as in other studies is Study 3a. It is unclear why this is the case. One possibility is the presence of other scenarios in the study. The *Incest* scenario was one of four scenarios that people engaged with. Given that framing and order have been identified as influencing moral judgements (Lanteri et al., 2008; Lombrozo, 2009; Nichols & Mallon, 2006; Petrinovich & O’Neill, 1996), it is possible that the presence of other scenarios influenced responses to the *Incest* scenario. Another possible reason is the homogeneity of the sample, consisting of undergraduate and postgraduate students from MIC, though Study 6 included a similar sample. This unexplained inconsistency in responding present in Study 3a highlights another feature of moral dumbfounding that the present work has identified. Despite being robust and reliably elicited, dumbfounded responding can vary unpredictably.

7.2 Variability and Unpredictability of Dumbfounding

Dumbfounded responding was reliably elicited in all 12 studies. There appeared to be a generally stable pattern of relative frequencies in patterns of responses, however this stability did not hold in all cases. Across 12 studies, a number of instances of unexplained variation were observed.

Firstly, the pattern of responses in Study 1 is considerably different from what has been observed in the other studies in two ways: (a) in Study 1, rates of dumbfounding varied depending on scenario, a result not found in Studies 2 or 3 (it was not possible to observe this variation in Studies 4-12 as they contained a single scenario); (b) rates of dumbfounded responding for the *Incest* scenario are substantially higher in Study 1 than in the majority of other studies. A speculative explanation for this was outlined in Chapter 3. Study 1 was an interview whereas all other studies were computer-based tasks. An interview is qualitatively different from a computer-based task, particularly regarding the relative social demands associated with each (e.g., Royzman et al., 2015). It is hypothesised that the variation in responses in Study 1 when compared to the other studies emerged as a result of an interaction between the relative difficulty of the various scenarios and the demands of the task. It is hypothesised that *Incest* is the most difficult scenario to justify, followed by *Cannibal* and *Trolley*, with *Heinz* as the easiest scenario to justify judgements for.

It is further hypothesised that two key features of an interview interact with the relative difficulty of the various scenarios to produce the pattern of responses observed. In an interview it is likely that participants experience a greater demand to (a) provide a justification for their judgement, and (b) to be accurate in providing justification. That participants experience a greater demand to provide a justification for their judgement may lead to lower rates of dumbfounding for the easier scenarios than in a computer-based task. In a computer-based task, participants may choose neglect to provide a reason and move on, while in an interview participants are pushed to a greater extent to provide a reason. For an easier scenario, participants are largely successful in identifying reasons for their judgement.

The second feature of an interview identified above may lead to higher rates of dumbfounding for more difficult scenarios than in a computer task. In a computer task, participants may provide a reason that is inconsistent or easily refuted without consequence. However, in an interview, inconsistencies in reasoning are pointed out to the participants in real time. This greater demand for accuracy in an interview may lead to higher rates of dumbfounding to be observed in an interview task for scenarios that are more difficult to justify. To test this possibility a second analysis was conducted whereby the coded string responses were included. This was conducted for Studies 3a and 3b separately (recall that the sample in 3b appeared to be more permissible of *Incest* than other samples). Study 2 included an unsupported declaration on the critical slide leading to high rates of dumbfounded responding, as such was not included in this analysis. A chi-squared test for independence revealed a significant association between scenario and response to the critical slide in Study 3a, $\chi^2(6, N = 288) = 14.850, p = .021, V = ..23$ (with an observed power of .83), and Study 3b, $\chi^2(6, N = 404) = 20.786, p = .002, V = .23$ (with an observed power of .95). The observed counts, expected counts, and standardised residuals are displayed in Table 7.2.

Table 7.2: Observed counts, Expected counts, and Standardised residuals for each response depending on Scenario in Studies 3a and 3b.

Study	Response		Heinz	Trolley	Incest	Cannibal
Study 3a	Observed count	Nothing wrong	14	15	12	4
		Dumbfounded	19	22	31	21
		Reasons	39	35	29	47
	Expected count	Nothing wrong	11.25	11.25	11.25	11.25
		Dumbfounded	23.25	23.25	23.25	23.25
		Reasons	37.50	37.50	37.50	37.50
	Standardised residuals	Nothing wrong	1.03	1.41	0.28	-2.72*
		Dumbfounded	-1.24	-0.36	-2.26*	-0.65
		Reasons	0.41	-0.68	-2.32*	2.59*
Study 3b	Observed count	Nothing wrong	21	24	31	10
		Dumbfounded	16	22	28	30
		Reasons	64	55	42	61
	Expected count	Nothing wrong	21.50	21.50	21.50	21.50
		Dumbfounded	24.00	24.00	24.00	24.00
		Reasons	55.50	55.50	55.50	55.50
	Standardised residuals	Nothing wrong	-0.14	0.70	2.67*	-3.23**
		Dumbfounded	-2.16*	-0.54	1.08	1.61
		Reasons	1.96*	-0.11	-3.12**	1.27

*Note: * = sig. at p < .05 (|z| > 1.96); ** = sig. at p < .001 (|z| > 3.11)*

When the coded open-ended responses were included in the analysis, responses in the dumbfounding paradigm appeared to vary with scenario in a manner consistent with observed in Study 1. In both Studies 3a and 3b, rates of providing reasons for *Incest* were significantly lower than the expected count. In Study 3a this translated into significantly higher rates of dumbfounded responding, in Study 3b however this was associated with higher rates of selecting “There is nothing wrong”.

In both Studies 3a and 3b rates of selecting “There is nothing wrong” for *Cannibal* were significantly lower than the expected count. In Study 3a this led to significantly higher rates of providing reasons, in Study 3b this did not lead to significantly higher rates of providing either of the other responses. Finally, in Study 3b rates of providing reasons for *Heinz* were significantly higher than the expected count, and rates of dumbfounded responding were significantly lower than the expected count. The inclusion of the coded open-ended responses in the analysis appears to align the patter of results of the computer-based task more closely with the results of the interview. However, in view of the concerns raised by Royzman et al. (2015) and discussed in Chapter 4, caution is advised in taking unsupported declarations or tautological responses as indicators of dumbfounding. The remaining discussion relates to the selecting of admissions of having no reasons only.

The results of study 1 differed from the other studies in two ways: variation in rates of dumbfounding with scenario, and substantially higher rates of dumbfounding for *Incest* than in other studies. The interaction between demands of the interview and the scenario difficulty provides a possible explanation for both of these. This explanation is speculative and has not been tested. The varying rates of dumbfounding, providing reasons, and selecting “There is nothing wrong”, depending on scenario and study are shown in Figure 7.3.

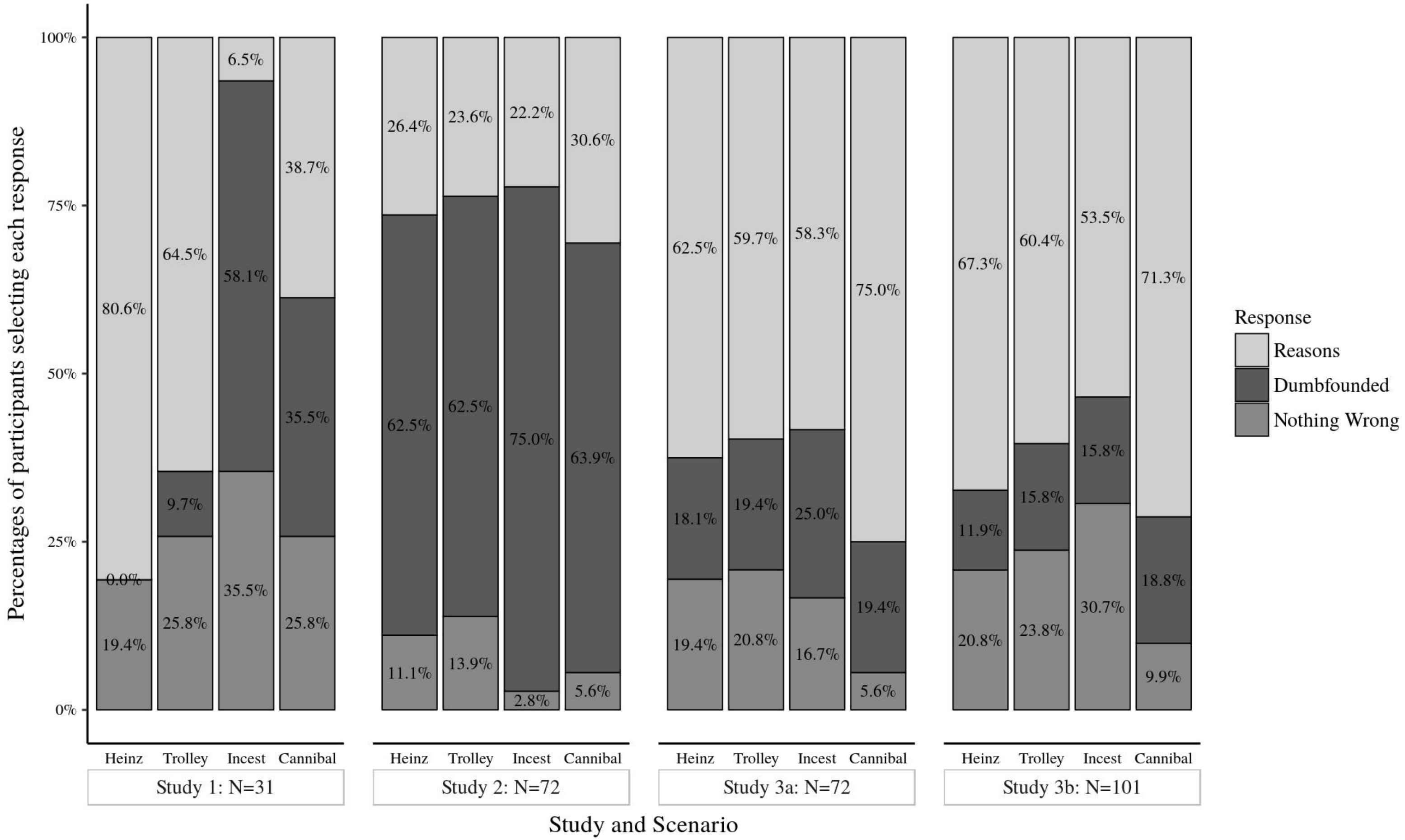

Figure 7.3: Rates of dumbfounding, providing reasons, and selecting "There is nothing wrong", for each scenario, for each Studies 1-3

Another unexplained instance of variability (identified in the previous section) can be seen in Figure 7.2 whereby the pattern of responses to the critical slide in Study 3a differs from the other studies. Under normal conditions (computer-based task; dumbfounding measured using the selection of an admission of not having reasons; a control group with no cognitive load manipulation) the rates of selecting "There is nothing wrong" are generally higher than rates of dumbfounding. It is unclear why the pattern of responses observed in Study 3a is different from the other studies. It is possible that the participants in this sample were generally less permissive of the behaviours they read than participants in other samples. However reasons for this are unclear, no relationship between religiosity (as measured by the Centrality of Religiosity Scale, Huber & Huber, 2012), and responses to the critical slide was found. As such, reasons for variability in the pattern of responses in the dumbfounding paradigm are unclear, though it seems likely that they are linked to people's judgements of the behaviours described.

A further instance of unpredictability of responses in the dumbfounding paradigm was identified in Chapter 6. Participants in Study 10 provided reasons for their judgements at a noticeably lower rate than in other studies; rates of selecting "There is nothing wrong" were higher than expected (see Figure 7.2). There was nothing unusual about the materials in Study 10 and as such this surprisingly low rates of providing reasons (and higher rates of selecting "There is nothing wrong") appears to be due to an unusual sample. On the other hand, the rate of providing reasons in Study 11 was unexpectedly high. Though, the smaller sample size following the exclusion of large numbers of participants may provide a reason to doubt the accuracy of this. Again it seems likely that this variability may be attributed to the relative permissiveness of the participants in the sample.

Responses to the dumbfounding paradigm are not systematically affected by any of the manipulations used in Studies 6-12. Responses in the dumbfounding paradigm appear to be stable. Attempts to influence the rates of various types of responses by three different experimental manipulations did not yield convincing results. Despite this apparent stability, across 12 studies there were at least three instances of variability in responding that has not been fully explained (Study 1, Study 3a, and Study 10).

7.3 Implications

Chapter 1 detailed the significant influence the existence of moral dumbfounding had on the moral judgement literature. The discovery of moral dumbfounding coincided with the emergence of intuitionist and dual-process theories of moral judgement (e.g., Greene, 2008; Greene et al., 2001; Haidt, 2001; Haidt & Björklund, 2008; Prinz, 2005). Haidt (2001; see also Haidt & Björklund, 2008) makes explicit reference to moral dumbfounding in outlining and defending his social intuitionist model of moral judgement. Though later theorists do not draw as heavily on moral dumbfounding, Haidt's work more generally has had a considerable influence on the development of theories of moral judgement over the past decade and a half (e.g., Brand, 2016; Bucciarelli et al., 2008; Crockett, 2013; Cushman, 2013; Cushman et al., 2010; Dwyer, 2009; Greene, 2008). That moral judgements are widely accepted to be, at least to some degree, grounded in intuition (or emotion, e.g., Prinz, 2005) may be attributed to the influence of Haidt (and moral dumbfounding).

7.3.1 Rationalism, intuitionism and moral dumbfounding. Moral dumbfounding provides a clear illustration of the intuitive nature of moral judgements and poses a significant challenge to rationalist approaches that identify moral judgements as grounded in principles. There is limited evidence that moral dumbfounding is a genuine phenomenon. It is surprising that the phenomenon that illustrates such a key theoretical claim has not been tested empirically in peer reviewed work. Indeed, sceptics of intuitionism have challenged the existence of moral dumbfounding (Jacobson, 2012; Kitcher, 2011; Royzman et al., 2015). The studies described in this thesis demonstrated moral dumbfounding as a real and robust phenomenon. In demonstrating that moral dumbfounding can be elicited, the studies described in this thesis have provided new evidence for the intuitive (as opposed to rationalist) nature of moral judgement. The studies described in Chapter 4 specifically addressed the challenge and associated rationalist explanation of moral dumbfounding from by Royzman et al. (2015) providing evidence against their explanation. Beyond a demonstration of the intuitive nature of dumbfounding, and a renewed refutation of a rationalist perspective, the broader theoretical implications of the studies described in this thesis are less clear. Indeed, the apparent evidence against a rationalist perspective and for the claim that moral judgements, are at least to some degree intuitive in nature of moral judgements does not provide support for any particular intuitionist theory of moral judgement, and such theories also fail to adequately explain moral dumbfounding.

7.3.2 Dual-processes and moral dumbfounding. Drawing on dual-process theories of moral judgement (e.g., Cushman, 2013), and on dual-process theories of psychology more generally (De Neys, 2006; Evans, 2010), a possible explanation for moral dumbfounding was described. According to this explanation, providing reasons for a judgement involves successful deliberation, and dumbfounded responding identified as relying on intuition or habitual responding following failed deliberation. The degree to which revising a judgement in light of a failure to justify it is grounded in deliberation or habitual responding is unclear. This explanation was tested through a series of studies that attempted to manipulate dumbfounded responding across Chapters 5 and 6. The results of these studies did not provide conclusive evidence in support or against this explanation. It seems likely that a dual-process explanation of moral dumbfounding may be partly right, but that the phenomenon is more complex than predicted by this explanation.

It was hypothesised that cognitive load would inhibit the identification of reasons, leading to increased rates of dumbfounding (or selecting "There is nothing wrong"). Conversely, it was hypothesised that psychological distancing would facilitate the identification of reasons and reduce rates of dumbfounding. Drawing on model theory (Bucciarelli et al., 2008), a further prediction was made, that providing a prompt for a reason would reduce rates of dumbfounding. A number of individual difference variables were also recorded. A brief summary of the key findings of each study can be found in Table 7.3.

Table 7.3: Summary of key findings in Chapters 5 & 6

Study		Control	Manipulation	Sig.	Individual Difference	Sig.
Study 6	Cognitive Load	No Manipulation	Memory task (Number String)	Yes	Need for Cognition	No
Study 7	Cognitive Load	Easy dot pattern	Difficult dot pattern	No Yes: (engaged)	Need for Cognition	No
Study 8	Cognitive Load	No Manipulation	Difficult dot pattern (and engagement)	Yes: engaged	Need for Cognition	Yes
Study 9	Cognitive Load	No Manipulation	Difficult dot pattern (and engagement)	No	Need for Cognition	No
Studies 6 – 9	Cognitive Load	-	-	Yes	Need for Cognition	Yes
Study 10	Distancing	No Manipulation	Anne Vignette	No	Social desirability	No
Study 11	Distancing	No Manipulation	Anne vignette (perspective)	No	Social desirability	No
					Need for Closure	No
					CRT	No
Studies 10 & 11	Distancing	-	-	Yes	Social desirability	No
Study 12	Models	Anne vignette	Anne vignette	No	Social desirability	No
		(perspective)	(reason prompt)		Need for Closure	No
					CRT	Yes
Studies 11 & 12	-	-	-	-	Social desirability	No
(& 10)					Need for Closure	No
					CRT	Yes

As noted previously, for each study, the variation in responding appeared to be consistent with that predicted by the relevant manipulation, however this variation did not reach statistical significance. The combined analyses in of Studies 6-9 in Chapter 5 found significant variation in responses to the critical slide depending on cognitive load. Furthermore, a mini meta-analysis was conducted and found that cognitive load significantly influenced responding across all studies in Chapter 5,

$\chi^2(8) = 23.81$, $p = .002$ (Fisher's method); or when weighting for sample size, $z = 2.95$, $p = .002$ (Stouffer's Z-score method). Similarly the combined analysis in Chapter 6 found significant variation in responses to the critical slide depending on the distancing manipulation. The mini meta-analysis did not find this effect in this case, $\chi^2(4) = 7.01$, $p = .135$ (Fisher's method); or when weighting for sample size, $z = 1.20$, $p = .116$ (Stouffer's Z-score method).

The data from all the studies that used the *Incest* scenario only were aggregated for combined analysis. Each participant in this aggregated dataset falls into one of the following four conditions: (1) cognitive load manipulation, (2) no manipulation, (3) distancing manipulation, (4) model prompt manipulation. The cognitive load manipulation was designed to inhibit deliberation, and as such level of deliberation of the participants in the cognitive load group (1) should be lower than the level of deliberation in the no manipulation group (2). The distancing manipulation (3) was designed to facilitate deliberation. The model prompt manipulation (4) was also designed to facilitate deliberation, but to a greater extent than the distancing manipulation in isolation. Given the relative levels of deliberation associated with each condition, and the degree to which deliberation is hypothesised to be required for the successful identification of reasons, it is hypothesised that if the four conditions are listed in order from 1 (cognitive load, lowest deliberation) to 4 (model prompt, highest deliberation) we should observe incremental increases in rates of providing reasons. The analyses of the individual studies demonstrated that any differences between incrementally adjacent groups may not necessarily be statistically significant, however, it is possible that significant differences may exist between the groups at either extreme. It is hypothesised that the lowest rate of providing reasons should be observed in the cognitive load manipulation group (1), while the model prompt group should display the highest rate of providing reasons (4). The distance manipulation (3), was designed to facilitate deliberation, as such it hypothesised that this rates of providing reasons should be higher for this group than no manipulation group (2). To recap, hypothesised rates of providing reasons, from lowest to highest are: (1) cognitive load manipulation, (2) no manipulation, (3) distancing manipulation, (4) model prompt manipulation. The aggregated responses to the critical slide are shown in Figure 7.4.

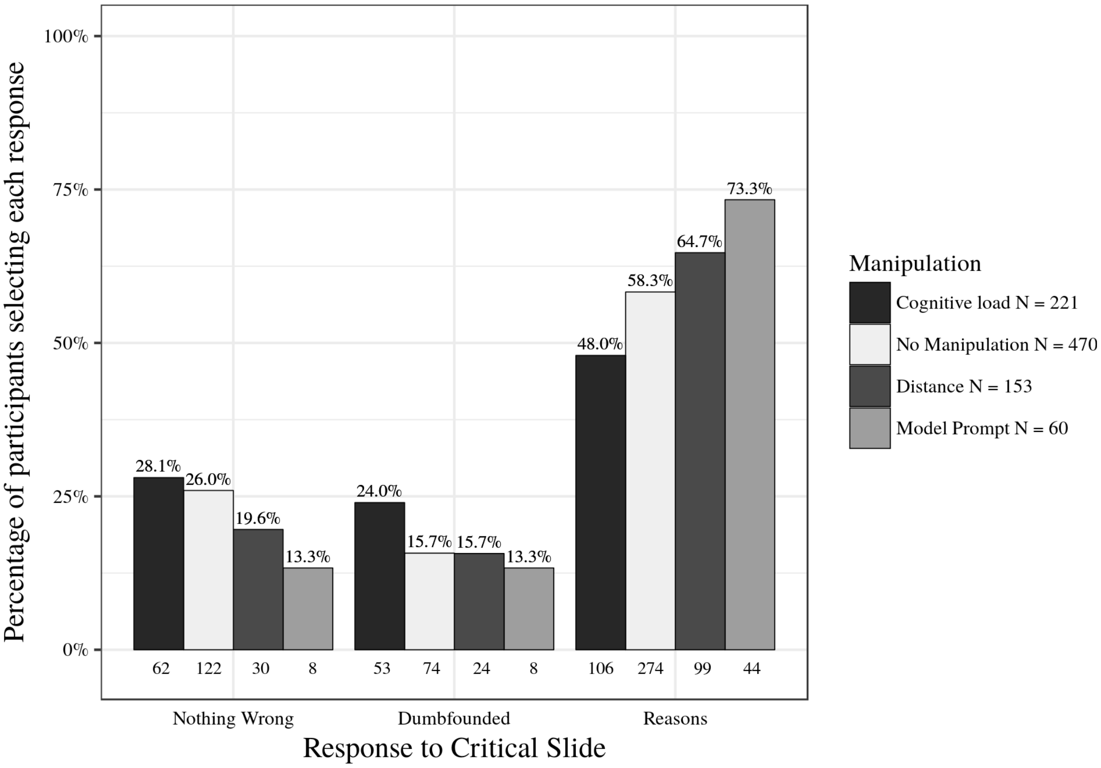

Figure 7.4: Responses to the critical slide for the Incest dilemma across studies 4-12

Table 7.4: Observed counts, Expected counts, and Standardised residuals for each response depending on manipulation across Studies 4 – 12

Response		Cognitive Load	No Manipulation	Distance	Models
Observed count	Nothing wrong	62	122	30	8
	Dumbfounded	53	74	24	8
	Reasons	106	274	99	44
Expected count	Nothing wrong	54.27	115.42	37.57	14.73
	Dumbfounded	38.87	82.67	26.91	10.55
	Reasons	127.86	271.91	88.52	34.71
Standardised residuals	Nothing wrong	1.39	1.02	-1.56	-2.09*
	Dumbfounded	2.87*	-1.52	-0.68	-0.9
	Reasons	-3.43**	0.28	1.88	2.51*

*Note: * = sig. at p < .05 (|z| > 1.96); ** = sig. at p < .001 (|z| > 3.11)*

A chi-squared test for independence revealed a significant association between manipulation and response to the critical slide, $\chi^2(6, N = 904) = 20.536$, $p = .002$, $V = .15$, the observed power was .94. Figure 7.4 shows the variation in responses to the critical slide depending on manipulation across all studies investigating only the *Incest* dilemma. Table 7.4 shows the observed counts, expected counts and standardised residuals for each response depending on manipulation. According to this analysis, a cognitive load manipulation led to significantly lower rates of providing reasons, and significantly higher rates of dumbfounded responding. Distancing did not appear to influence response to the critical slide. The model prompt led to significantly higher rates of providing reasons and significantly lower rates of selecting “There is nothing wrong”.

The trend in responding depending on manipulation is easily visible in Figure 7.4. However, this trend can only be reliably observed when the results of a large number of studies are combined. In truth, responding is much more variable and moral dumbfounding appears to be more complex than the dual-process or model theory explanations predict. A more detailed illustration of the rates of selecting each response depending on experimental manipulation can be seen in Figure 7.5. Rates of dumbfounding are represented by the darkest grey (middle), rates of providing reasons are represented by the lightest grey (top), and rates of selecting "There is nothing wrong are represented by the grey at the bottom. The manipulation type and study number are on the X axis. Data for each study are paired by colour. The trend that was easily observable in Figure 7.4 is less clear, with a great deal of fluctuation within manipulation types. Looking at each study individually (coloured pairs) the trend still holds – the spike in providing reasons under cognitive load in Study 8 is matched by a spike in the Control; similarly, the surprisingly low rates of providing reasons in the control Study 10 (distancing) appears to be matched by a lower rate of providing reasons in the manipulation group. However, the high degree of fluctuation seems to indicate that providing of reasons is moderated by more than just the manipulations tested here.

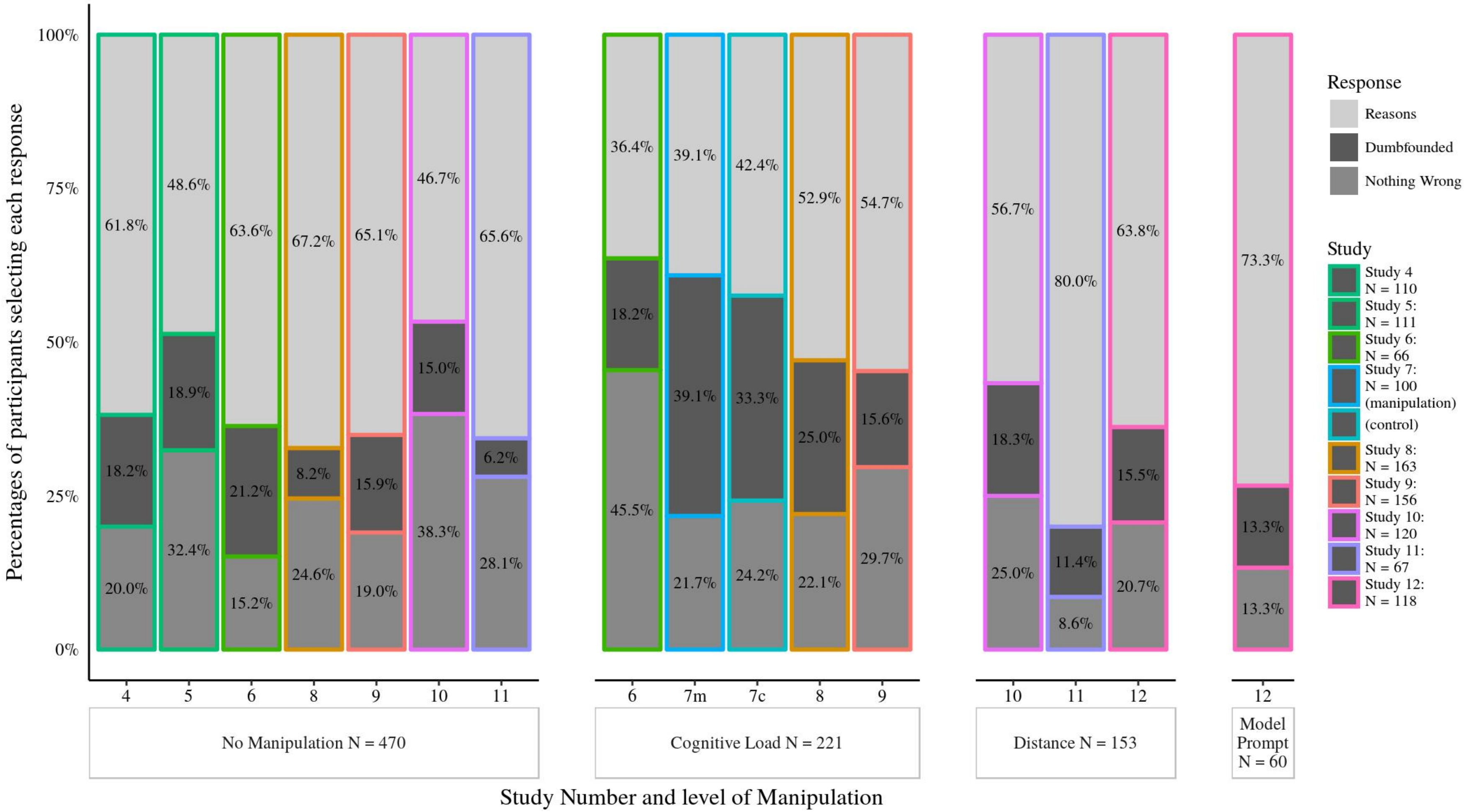

Figure 7.5: Percentage of participants selecting each response for each type of manipulation for each study across Studies 4-12

7.3.3 Practical implications. The studies described in this thesis provide a practical demonstration that in some cases people maintain a moral judgement even thought they cannot justify it through reasoned argument. This has implications regarding the degree to which any moral judgement can be justified through reasoned argument. These implications extend to public discourse regarding to moral issues, particularly in relation to controversial issues (e.g., abortion, euthanasia, US gun laws). Invariably, in discussions of such issues, people on both sides are accused by the other side of not having reasons for their judgement. The existence of moral dumbfounding suggests that these accusations may be grounded in truth. In effect, these debates could descend into an argument between (two) people who are simply morally dumbfounded. Clearly, such a discussion is of limited value, particularly in situations where the purpose of engaging in discussion is to identify common ground and seek solutions (e.g., government ministers debating contentious policy). Given our current understanding of moral dumbfounding such a situation would be little more than an interesting example of moral dumbfounding in the real world. However, a better understanding of the mechanisms that lead to dumbfounding may provide insights into ways to reduce dumbfounding. It may be possible to devise a set of strategies that a person chairing or hosting such a discussion (e.g., radio DJ, committee/task-force chairperson) may be able to use in order to prevent people from getting dumbfounded.

7.4 Limitations and Future Directions.

There are a number of limitations of the studies described in this thesis, both practical and theoretical. Firstly, it is not clear that the dumbfounded responding we observed is necessarily evidence of moral dumbfounding as widely understood. Two responses were taken to be evidence of the “stubborn” and “puzzled” maintenance of a judgement in the absence of supporting reasons (Haidt et al., 2000, p. 2). These responses were an admission of not having reasons, and the use of unsupported declarations as justifications for judgement.

These responses are taken as evidence for “true moral dumbfounding” a failure of deliberation in attempting to support an intuitive moral judgement (reasoning failure). It may possible to provide alternative explanations of dumbfounded responding. For example, a person may (as argued by Royzman et al., 2015) succumb to the pressure to accept that the reasons for their judgements they provided are inadequate and outwardly acknowledge the inadequacy of these reasons, however, this does not mean that the person believes the reasons are inadequate. They may have just given up as a result of social pressure (compliance). Alternatively a person may simply regard defending their judgement as too effortful and resort to providing dumbfounded responses as a result of laziness or neglect (laziness). Beyond compliance and laziness, it is also possible that a person may become frustrated with the experiment and choose not to cooperate with the researcher and simply refuse to provide reasons for their judgement (obstinance).

Four possible reasons for providing a dumbfounded response have been provided above. This list is not exhaustive, a person may provide a dumbfounded response for any number of reasons. However, of particular importance for the current discussion is that only “reasoning failure” may be viewed as consistent with how moral dumbfounding is generally presented in the literature. It was presented by Haidt (2001, Haidt et al., 2000) as evidence that the making of moral judgement is not grounded in reasons or principles because people fail to provide relevant reasons when requested. This interpretation neglects the other possible reasons for providing a dumbfounded response identified above.

Throughout this thesis, I have followed the practice of the majority of authors discussing moral dumbfounding and viewed dumbfounded responding as evidence of a failure to provide reasons, and evidence of an absence of reasons. This decision was informed by the analysis of the video-recorded interviews in Study 1. The responses adopted as evidence of dumbfounding were accompanied by distinctive patterns of behaviour such that it was clear to an observer that participants were struggling to identify reasons. Quantifying such a pattern of behaviour is challenging, individuals' overall behaviour in the interview varied considerably (e.g., some participants sat rigid and still for most of the interview while others made extensive use of gestures and changes of posture). I attempted to break down the implicitly recognisable signs that a person may be struggling to identify a reason for their judgement into objectively measurable behavioural variables (see DeLancey, 2001 for discussion on the practical limitations of describing behaviour in terms of objectively measurable micro-behaviours). Three variables that appeared to capture this struggle were time spent in silence, frequency of smiling, and frequency of laughing. There were significant differences across each of these variables between participants who provided a dumbfounded response and participants who did not provide a dumbfounded response. It is also possible apparent rates dumbfounding based on these responses (an admission of not having reasons/unsupported declaration) provide a conservative estimate of the prevalence of dumbfounding. There were at least two participants who appeared to be dumbfounded based on their pattern of behaviour but they did not provide an admission or an unsupported declaration (while the video was recording) and could not be identified as dumbfounded.

Despite the confidence that these responses provided evidence for "true moral dumbfounding" in the interview in Study 1, it is possible that the providing of these responses in a computerised occur due to some reason other than a failure to identify reasons (compliance, obstinance, laziness). Indeed the high rates of dumbfounding observed for *Heinz* and *Cannibal* may provide evidence for reasons other than a failure to identify reasons leading to dumbfounded responding.

The possibility that the dumbfounded responding observed in a computerised task may not reflect "true moral dumbfounding" has significant implications for the interpretation of the results of Studies 6-12. In each of these studies the manipulations employed were designed to influence "true" cases of dumbfounded responding, either through inhibiting the identification of reasons (Cognitive Load – Studies 6-9), or facilitating the identification of reasons (Distancing – Studies 10 and 11; Model Prompt – Study 12). These manipulations can only be successful when dumbfounding is viewed as a failure to provide reasons (as opposed to laziness/disengagement with the task). If participants provide a dumbfounded response due to laziness or lack of engagement with the task then it is unlikely that these responses will be influenced by the manipulations employed. This is particularly true for attempts to facilitate the identification of reasons, as in such cases participants must engage with both the perspective taking task and the task of identifying reasons for their judgement. Providing a dumbfounded response as a consequence of disengagement is unlikely to be influenced by the introduction of an additional task that requires further engagement. Recall participants in the manipulation group in Study 12. These participants were provided with a reasons to judge the behaviour as wrong. Despite being provided with a reason, 8 participants (13%) selected the dumbfounded response. It is possible that these 8 participants were not engaging with the task. It is only through further research that this possibility may be explored further. The methods and materials developed in this thesis will provide a valuable resource in addressing this question.

Incorporating research on meaning maintenance (Heine et al., 2006, 2006; Proulx & Inzlicht, 2012) into the study of moral dumbfounding may prove useful in this. One hypothesis is that "true moral dumbfounding" presents a threat to meaning. A threat to meaning is accompanied by a range of compensatory behaviours, such that if a person engages in one of these behaviour it provides evidence that they experienced a threat to meaning (Proulx & Inzlicht, 2012). The hypothesised relationship between meaning maintenance and moral dumbfounding could be investigated in an interview setting to ensure that participants providing dumbfounded responses can truly be identified as dumbfounded (based on their behavioural responses). If such a relationship exists, a meaning compensation task may be included in the computerised version of the dumbfounding paradigm in order to differentiate the "truly dumbfounded" from participants providing dumbfounded responses for other reasons.

A second key limitation was identified while characterising the possible responses in the dumbfounding paradigm according to intuition (habitual responding) and deliberation. It was initially hypothesised that identifying reasons relied on deliberation, dumbfounded responding relied on intuition, and changing a judgement required deliberation (though not as much deliberation as identifying reasons). This characterisation would provided clear simple testable hypotheses with clear (hopefully) results. On closer inspection it became apparent that this initial characterisation was incorrect, that intuition (habitual responding) may play a greater role in other responses, including the identification of reasons. It also became apparent that there may be more than one intuition at play, and that these intuitions may come into conflict.

The extra layers of complexity in the dumbfounding paradigm were apparent again in the results of the studies conducted in Chapters 5 and 6. These results appeared to indicate patterns of responses in line with the predicted trends. These results were inconclusive. This may have been due to the limited power of the studies conducted. Future research may investigate this possibility, by recruiting much larger samples. Given the (at times unpredictable) variability of responses, it is likely that the dumbfounding paradigm is more complex than predicted by these theories.

Haidt (2001) proposed SIM in opposition to rationalism in order to provide an explanation of moral dumbfounding. Since then various aspects of SIM have been developed and incorporated into more recent theories of moral judgement. The limited empirical work investigating moral dumbfounding has meant that explaining moral dumbfounding has not been a central tenet of more recent theories. Two hypothesised explanations of dumbfounding drawing on dual-process approaches more generally, and model theory specifically were tested in Chapters 5 and 6 of this thesis. Moral dumbfounding is more complex than predicted by these explanations, such that currently there is no theory of moral judgement that provides a clear explanation of moral dumbfounding that addresses the complexities of dumbfounded responding. In view of this limitation of the existing moral judgement literature, developing an explanation for moral dumbfounding is a logical next step in progressing theories of moral judgement. Given the limited explanations of moral dumbfounding extant in the literature, placing an understanding of moral dumbfounding at the centre of the development of theories of moral judgement may serve to push the morality literature in a new direction, as happened in the early 2000s when Haidt (2001) proposed SIM in opposition to rationalism.

The final chapter in this thesis explores one potential such theory. Moral dumbfounding is taken as evidence of the intuitive nature of moral judgement, however few theories provide an account of the emergence of moral intuitions. The theoretical position explored in the final chapter draws on the categorisation literature in an attempt to identify the processes that lead to the emergence of moral intuitions. The occurrence of moral dumbfounding is attributed to moral intuitions, as such by identifying the processes that give rise to the emergence of moral intuitions; the theoretical position outlined in the next chapter provides a possible explanation of moral dumbfounding.

7.5 Conclusion

The aims of the studies described in this thesis were to assess if moral dumbfounding is a real phenomenon, and assess the implications of the existence (or absence) of moral dumbfounding for theories of moral judgement. Evidence that moral dumbfounding is a genuine phenomenon was found across 12 studies. The second aim, assessing the implications of the existence of moral moral dumbfounding for theories of moral judgement did not yield a clear answer, that is, the existence of moral dumbfounding highlighted a limitation of extant theories of moral judgement in terms of their ability to explain dumbfounding.

Dumbfounded responding can be reliably evoked using the materials developed in the studies in this thesis. With some notable exceptions, the pattern of responses in the dumbfounding paradigm appears to be relatively stable, most people provide reasons, and dumbfounded responding is generally the least frequent response. Furthermore dumbfounded responding appears to be more resistant to manipulation than a dual-process explanations would predict. Similarly, the observed variability was not as reliable as model theory would predict. Any effects observed in the studies described in Chapters 5 and 6 are either too fragile or too weak to be reliable. Despite the apparent stability of moral dumbfounding, there are some notable instances of variability that is still unexplained. Dumbfounded responding is both variable and resistant to manipulation. Given that moral dumbfounding cannot be adequately explained by the existing theories of moral judgement adopted here, the following chapter will explore a possible alternative theory of moral judgement that may provide an explanation of moral dumbfounding.

8 Chapter 8 (Epilogue) – Explaining Moral Judgement by Attempting to Understand Moral Dumbfounding

Three possible explanations of moral dumbfounding were examined in the studies described in Chapters 4, 5, and 6. These were drawn from three theoretical approaches: rationalism, dual-process approaches, and model theory. In Studies 4 and 5 (Chapter 4) we presented evidence against the rationalist perspective proposed by Royzman et al. (2015). We failed to provide strong support for a dual-process explanation of moral dumbfounding in Studies 6-9 (Chapter 5) and Studies 10 and 11 (Chapter 6). Study 12 (Chapter 6) did not offer support for model theory. The critical literature review and new studies presented in this thesis have highlighted the weaknesses of existing theories of moral judgement in explaining the phenomenon of moral dumbfounding. Thus, the aim of this chapter is to explore a possible theory of moral judgement that is developed around providing an explanation of moral dumbfounding (Research question 2.2).

Moral dumbfounding provides an illustration of the intuitive nature of moral judgements and thus. In doing so, moral dumbfounding has played a key role in the development of our understanding of moral judgements more broadly. It is now widely accepted that moral judgements are grounded (at least to some degree) in intuition (see Cameron et al., 2013). Despite this, the emergence of moral intuitions remains poorly explained in the morality literature. Haidt's (2001) claim that moral intuitions are innate has been widely rejected (Machery & Mallon, 2010; Mallon, 2008; Prinz, 2008a, 2008b). Cushman (2013) described two types of learning (model-free and model based) that may lead to the emergence of moral intuitions, however Cushman's overall approach is limited by reliance on the alignment of the model-free/model-based distinction with the untenable action/outcome distinction (see Chapter 1, section 1.5.2.1).

In the remainder of this chapter, we argue that explaining the emergence of moral intuitions requires looking beyond the morality literature, toward the cognitive psychology literature more generally. Parallels have previously been drawn between the process of categorisation and the process of making moral judgement (Harman et al., 2010; Prinz, 2005; Roedder & Harman, 2010; Stich, 1993). According to a categorisation approach to moral judgement, making a moral judgement is simply a categorisation task: behaviours are *categorised* as either RIGHT or WRONG. Building on this link between morality and categorisation, we draw on the categorisation literature to provide an account for the emergence of moral intuitions. By accounting for the emergence of moral intuitions, this approach offers an explanation of moral dumbfounding, where dumbfounding occurs as a consequence of the way in which specific intuitions are acquired.

A brief rationale for this approach is outlined below. Next, processes that give rise to the emergence of stability categorisation are described in detail (discussions of moral judgement will be notably absent in this section). The processes of categorisation are then applied to the moral domain, providing an account for the emergence of moral intuitions, and a potential explanation of moral dumbfounding. Finally, evidence in support of this categorisation approach to understanding moral dumbfounding is presented.

8.1 Rationale for a Categorisation Approach

The key motivation for exploring the categorisation approach to moral judgement proposed in this chapter is that it provides a coherent account for the emergence of moral intuitions and in doing so provides a possible explanation of moral dumbfounding currently absent in the morality literature. The idea that moral judgements may be studied as categorisations is not unprecedented, having been proposed independently by Stich (see also Harman et al., 2010; Roedder & Harman, 2010; Stich, 1993) and by Prinz (2005). However as noted in Chapter 1, the view of categorisation adopted by these approaches does not reflect developments in the categorisation literature.

Stich (1993) rejected a view of moral judgement that is grounded in principles (i.e. a rationalist approach) almost a decade before Haidt published his SIM (2001). Stich (1993) discusses the pervasive influence of non-moral features of a situation on moral judgements – recall that subtle differences in trolley dilemmas lead to different judgements. Stich does not cite trolley problems to illustrate this point, rather discusses the difficulty people have in condemning the raising of human babies for meat while simultaneously defending the raising of farm animals (e.g., pigs) for meat. Stich describes the various difficulties people encounter when attempting this, a description that appears to include various features of moral dumbfounding.

Stich notes limitations in both exemplar (where an exemplar is a specific instance of a category member) and prototype (where a prototype is a "typical" member of a particular category) theories of categorisation. In rejecting rationalism, and recognising the inadequacy of existing theories of categorisation, Stich turns to the linguistic analogy as a more promising possibility. The limitations with the linguistic analogy have been outlined in Chapter 1, while developments in the categorisation literature have resulted in prototype and exemplar approaches being replaced by better theories. As such this chapter will describe a categorisation approach to moral judgement that is consistent with Stich's original claims and reflects developments in the categorisation literature.

8.2 The Emergence of Stability in Categorisation.

The idea that there are stable categories within the human conceptual system has been challenged by various authors (Barsalou, 1987; e.g., McCloskey & Glucksberg, 1978; Rosch & Mervis, 1975), and there is now a substantial body of evidence suggesting that categorisation is a dynamical process (for review see Barsalou, 2003). Despite this demonstrated lack of stability, it is also clear that people do demonstrate some stability in the making of a wide range of categories. The premise of the account of categorisation discussed here is that stability in categorisation emerges through the acquisition of skill in making relevant categorisations (Barsalou, 1999, 2003, 2008, 2009). Applying a skill formation account of categorisation to the moral domain is consistent with extant skill based accounts of moral judgement (e.g., Dreyfus & Dreyfus, 1990; Hulsey & Hampson, 2014; Narvaez, 2005), and provides an account for the underlying mechanisms that give rise to moral intuitions discussed in intuitionist and dual-process theories (Crockett, 2013; Cushman, 2013; Prinz, 2005). According to this account, everyday categorisation, including the categorisation of behaviours, is an implicitly acquired skill. It emerges, and is maintained through practice; this practice occurs within specific contexts such that subsequent performance is context dependent. Barsalou (1999; 2003) proposed that this practice leading to stability is grounded in the process he calls type-token interpretation.

8.2.1 Type-token interpretation. According to Barsalou (2003) type-token interpretation is the process that underlies the emergence of stability in categorisation. Barsalou (1999) describes type-token interpretation as the binding of specific tokens to general types, thus allowing relevant inferences to be made. Put simply, type-token interpretation is the identification of an item as a member of a particular category. This does not necessarily imply the explicit naming of categories or category members, it simply requires the treating of items as members of a particular category.

To illustrate this point, consider the development of the goal-derived, ad-hoc category THINGS TO PACK INTO A SUITCASE as described by Barsalou (1991). Items that fall into this category (toothbrush, spare clothes etc.) are not generally categorised as such on a day to day basis. The category emerges as required; i.e., when a person needs to pack things into a suitcase. In this case, type-token interpretation, as the identification of an item (token) as a member of a particular category (type), is the identification of a given item as something that you pack or do not pack into a suitcase. Type-token does not necessitate the explicit naming of tokens and types, it simply requires the treating of tokens as types, the treating of an item as a member or not a member of a particular category, in this case, packing it or not packing it.

A person who travels frequently will be able to form the category THINGS TO PACK INTO A SUITCASE more readily than a person who does not travel as regularly. Through repetition and rehearsal, a person who travels regularly develops a greater level of skill at forming the category THINGS TO PACK INTO A SUITCASE than a person who does not travel as regularly. The development and execution of this skill is grounded in the development of skill or automaticity in type-token interpretation. It is this skilled, automatic, context-driven, and habitual type-token interpretation that underlies all categorisations.

The above discussion has described the role of type-token interpretation in the emergence of an ad-hoc goal derived category. It is this same process that underlies the formation of categories more generally. Barsalou (2003) refers to the extensive contextual influences on categorisation to support this position (e.g., the effect that manipulations of perceptual variables such as occlusion, size, shape, orientation, and modality has on performance 2003, pp. 530–535). Our continued interaction with the environment allows for the encountering of an infinite number of categories and category members. As category members are encountered, they are identified as members of their particular categories through type-token interpretation. The emergence of stability in categorisation results from continued and consistent type-token interpretation. In this view, "stable categories" are categories that the agent is highly skilled at identifying. Typically, these "stable categories" mirror real world categories or classes and social norms. This reflects the use of these categories in (a) interacting effectively with the world and (b) communicating with others effectively. In this way, categories such as natural kinds and social norms emerge as categories that may be identified with a degree of stability.

Type-token interpretation occurs every time a given token is encountered. Objects are not encountered independent of context, rather they are encountered in an on-going stream of goal-directed behaviour. As such, every categorisation of a given token (object/item/event) is subject to contextual influences of the current situation. This means that the properties of an object that are relevant to a particular interaction with it become salient during that interaction. These properties are learned and their identification or recognition may become a part of the subsequent interactions with the object, through type-token interpretation. The properties that are learned and the properties that become salient during a given interaction with a given type depend both the current interaction with the given type and also on previous interactions with the given type.

To illustrate this, consider a study by Barsalou (1982 as cited in, 2003). In this study participants were presented with a series of sentences involving particular items. For example: The basketball was used when the boat sank; or The basketball was well worn from much use (Barsalou, 1982, 2003, p. 537). Following each sentence, participants were asked to verify whether particular properties were true for the item; for example whether or not “floats” is true for “basketball” following reading either of the above sentences. The fact that basketballs float is relevant to the first sentence and thus this property is inferred from reading this sentence. In the second sentence, this property (while still true for basketball) is irrelevant, and does not become salient by reading the sentence. Thus, while what is true for basketball does not change depending in the situation, the properties that are inferred in a given instance do. This is evident in that participants were faster at verifying “floats” as true for basketball following reading the first sentence than the second (Barsalou, 1982, 2003). Other studies have yielded similar results, demonstrating that different sentences cause different properties to become salient depending on these properties' relevance to the given sentence; for example the ease with which “yellow” and “malleable” are identified as properties of “gold” when given the sentence “In the light, the blond hair of the little girl had the luster of gold” or “In the shop, the artisan shaped with ease the bar of gold” (Greenspan, 1986; Tabossi, 1988; Yeh & Barsalou, 2006).

8.2.2 Expertise in categorisation. Further evidence for the role of rehearsal in the emergence of skill in making particular categorisations can be found in research on the categorisation of emotions in the faces of others, and in research on absolute pitch/perfect pitch. Furthermore both areas of research also demonstrate the importance of context in performance, specifically, the importance of coherence between learned context and performance context.

Research on specific contextual influences on emotion recognition in faces suggests that this is a context dependent acquired skill. It is established that recognising emotion in the faces of others is better for “own race” faces (Anthony, Copper, & Mullen, 1992; Elfenbein & Ambady, 2002; Meissner & Brigham, 2001). A recent study by Yankouskaya, Humphreys and Rotshtein (2014) suggests that the skill emerges as a result of frequency of exposure. Using a British sample that included Europeans, Asians, and Africans, they found that the improved performance for “own race” was moderated by social contact, such that people who had high levels of social contact with members of another race (other than their own) displayed better performance for this race. Thus it was not simply a case of better performance for “own race”. The performance on facial recognition was related to the exposure to particular race. Increased exposure improved performance. This example demonstrates that particular tasks, traditionally regarded as categorisation, demonstrate clear context dependent skill effects. The categorisation of particular emotions in the faces of others is a skill that develops with practice. Rehearsal improves performance, but so too does the context of the rehearsal. The context in this case is the race of the person whose face displays the emotion being identified.

Further evidence of a context dependent skill view of categorisation can be seen in research on absolute pitch. Absolute pitch is defined as the “ability to identify the pitch of a musical tone or to produce a musical tone at a given pitch without the use of an external reference pitch” (Takeuchi & Hulse, 1993, p. 345). It is widely believed to be quite rare, present in less than .01% of the population (Takeuchi & Hulse, 1993). However this statistic refers to the generation or recognition of the pitch of tones in isolation, free of context. Recent research has found that if the phenomenon is investigated in the context of everyday experiences with specific tones then absolute pitch or pitch memory appears to be much more widespread than previously thought.

Levitin (1994) showed that when asked to sing a well known song in the correct key, people perform well above chance. The suggestion that this may be due to the development of "muscle memory" as opposed to evidence for absolute pitch has been rejected (Schellenberg & Trehub, 2003). In another study by Schellenberg and Trehub (2003), participants listened to instrumental recordings of the theme tunes from well known television shows. These recordings were either in the original key or shifted up or down by 1 or 2 semitones. Again, successful performance at this task was well above chance (Schellenberg & Trehub, 2003). Wong and Wong (2014) found that when a pitch naming task was replaced with "match/mismatch" identification task accuracy improved. Similarly incrementally introduced contextual factors (timbre, visual cues, sensory-motor cues) increased accuracy of pitch verification (Wong & Wong, 2014).

Clearly absolute pitch is a skill that is developed within a particular context. Removing the context in which the skill usually occurs hinders performance and removes evidence for the skill. For most people the skill of absolute pitch only exists within a particular context. There is a small minority who have developed the skill to such a degree that their ability to identify the pitch of a tone is not dependent on a particular context. This means that absolute pitch is not an ability reserved for a gifted minority as traditionally assumed, rather it is a skill that can develop in anyone under the right conditions. In order to identify the skill of absolute pitch in people, the right conditions need to be present.

The two examples above may be regarded as context specific expertise in a categorisation task in a given domain. Recognising (and correctly categorising) emotion in the faces of others is a skill that improves with rehearsal. Performance at a later stage depends on the context of rehearsal; performance is better for frequently encountered contexts than for less frequently encountered contexts. Similarly, correctly recognising or generating a tone of a particular pitch can be regarded as categorisation tasks, the performance of which improves with practice. People display a greater level of skill if performing these tasks in familiar contexts (e.g., for the theme tune of a common TV show).

8.2.3 Categorisation and rules. That stability in categorisation emerges through repetition and rehearsal means that these categorisations are not directly governed by an explicitly represented stable rule set. It emerges through, is governed by, and maintained by repeated type-token interpretation; thus in everyday use of categories there are no set rules that necessarily govern category membership. In reality, categories that can be referenced to natural kinds may take on the causal rules that distinguish natural kinds. For example, fruit are distinct from vegetables in that the agreed scientific classification of fruit (in our culture) is as containing the seeds. This causal rule is not necessarily operationalised in everyday interactions with fruit and vegetables, however in certain situations it may be referenced in order to aid in the classification of ambiguous items.

More abstract categories are more difficult to define because there may not be a set of causal rules governing membership to reference. Consider emotion categories; there is a large body of literature documenting the search for causal rules or specific identifying characteristics of particular emotion categories but there is, as yet, no approach that has fully answered this question (Griffiths, 1997). Barrett et al.

(2014) suggest that the only truly accurate way to define emotion categories is as populations of instances. In this view, identifiable emotion categories “exist” simply as a result of emergent stability in category formation. There is no universal physiological or behavioural reaction that holds for all instances of any particular emotion, similarly there is no situation that universally elicits a particular emotion (see Mesquita, Barrett, & Smith, 2010).

Barsalou and Wiemer-Hastings (2005) provide an insight into the study of more abstract concepts. According to their account, the content of increasingly abstract concepts contains increasing situational and introspective focus. In other words the degree to which situational and introspection inferences are implicated in the categorisation of abstract concepts is greater than for concrete concepts. Consider the possible inferences associated with the categorisation of SOFA versus FREEDOM. Various properties of SOFA will remain relatively stable across contexts (e.g., it has cushions, it has arm rests, it is for sitting on), such that object level inferences can be made independent of context. Identifying inferences regarding FREEDOM that remain similarly stable across contexts is more difficult. Instead, such inferences are generally linked to specific situational contexts or introspections (e.g., freedom *from oppression*, freedom *of speech*, to *feel* like one’s freedom has been infringed upon). As concepts become increasingly abstract, the associated inferences become increasingly situational and/or introspective.

8.3 Moral Judgement as Categorisation

It is proposed here, in line with Stich (1993), that the making of a moral judgement is grounded in the same processes that underlie categorisation. The current discussion expands on Stich's (1993) work to incorporate significant developments in the categorisation literature over the past 20 years. The learning of moral categories is identified as being governed by the process of type-token interpretation. There is a significant body of evidence to show that moral judgements are subject to the same types of contextual influences that affect categorisation more generally. Pertinent to the current discussion is that notion that there is not an explicitly represented set of rules that govern category membership. In other words, moral judgements are not grounded in explicit moral principles. This view of moral judgement would lead to the emergence of stability in the categorisation of behaviours as morally right or wrong.

When we encounter a certain behaviour we may learn that it is morally right. Each subsequent time this behaviour is encountered it is associated with moral rightness. As our exposure to this behaviour increases and, as it is identified as morally right more and more frequently (continued type-token interpretation), it emerges as a “good example” of morally right behaviour; or an exemplar for MORALLY RIGHT. Over time, we develop a range of “exemplars” that constitute MORALLY RIGHT; by the same process we develop a range of “exemplars” that constitute MORALLY WRONG.

As we develop a larger number of exemplars for MORALLY RIGHT and MORALLY WRONG we may begin to make links between specific exemplars and develop more generalised exemplars. Similar behaviours may become categorised together, for example continued identification of “hitting people” as WRONG, and “kicking people” as WRONG may lead a person to form a parent category CAUSING HARM TO PEOPLE which is consistently identified as WRONG. This may then be taken a step further and “don't harm people” and “don't harm animals” may merge to form INFLICTING HARM which is consistently identified as WRONG.

It is proposed here that the emergence of highly generalised morally grounded exemplars may form the basis of what we call values. Furthermore, as more and more exemplars are developed and become increasingly generalised, these generalised exemplars become arranged hierarchically in terms of severity. This essentially becomes our “moral code”. It must be stressed that there is not necessarily an underlying set of rules governing this moral code, it is based on a large collection of exemplars. It is even likely that some of the generalised exemplars (values) may appear to exhibit sufficient powers of “governance” to constitute rules. However, these are not true rules, they are simply coherent sets of exemplars, and even at this complex level of abstraction people continue to categorise by type-token interpretation. As with the mapping of stable categorisations onto natural kinds, it may be possible to construct plausible (and often true) causes for the associations that define many categories, however the process of categorisation is grounded in type-token interpretation as opposed to the rules that can be inferred from referencing observable categories.

Recall that Barsalou and Wiemer-Hastings (2005) identified abstract concepts as grounded in situational and introspection inferences. This would imply that moral categories are rich in situational and introspection inferences. This appears to be the case. Whether a particular behaviour is viewed as right or wrong varies depending on the situation (consider the array of variants on the trolley problem that produce different judgements). Similarly, with regards introspection, the tight coupling of moral judgements and emotions has been widely discussed in the literature. Prinz's claim that emotions cause moral judgements (Prinz, 2005) may be too strong. However, that they play some role is predicted by adopting Barsalou's categorisation framework, and widely supported by data (e.g., Cameron et al., 2013; Huebner et al., 2009; Royzman, Atanasov, et al., 2014; Rozin et al., 2009, 1999; Valdesolo & DeSteno, 2006; Wheatley & Haidt, 2005).

The categorisation account of moral judgement provides a coherent explanation of the emergence of moral dumbfounding. Recall that dumbfounding typically occurs for harmless taboo behaviours. Consider learning of taboo behaviours as wrong through type-token interpretation and a typical interaction with such a behaviour. The taboo nature of these topics means that they are consistently identified as morally wrong, without much discussion. This leads to a high degree of stability in categorising them as WRONG. However, while other behaviours may be discussed or disputed, generating a deeper knowledge surrounding the rationale for identifying as right or wrong, the taboo nature of these behaviours prevents them from being discussed. Recall that the petition to legalise incest in Scotland was dismissed without discussion, and that the media appeared to place more focus on the fact that the petition was required to be considered rather than on the content of the petition. This means that a typical encounter with such a behaviour involves little more than identifying it as wrong, possibly with an expression of disgust, and changing the subject. Identifying causal rules that govern the behaviour's membership of the category MORALLY WRONG is likely problematic, in that a person would have limited experience at attempting to do so. Deliberate online external referencing simply reveals “taboo” as a reason for the behaviour to be morally wrong and there is not necessarily a lot of detail that comes to mind easily. In this view, type-token interpretation of taboo behaviours logically leads to moral dumbfounding.

The categorisation approach to moral judgement proposed here provides an account for the emergence of moral intuitions that also explains how moral dumbfounding may arise. This approach is largely consistent with the approaches of both Stich (1993) and Prinz (2005), reflecting developments in the categorisation literature. At present, there is no direct evidence in support of this approach, as it has not been tested. Despite this, there is indirect evidence for this approach in the form of extensive parallels between the categorisation literature and the morality literature. The following section will provide an overview of these parallels.

8.4 Evidence for a Categorisation Approach to Moral Judgement

A range of parallels exist between morality and categorisation. Variability in morality occurs both within people (e.g., Narvaez, 2005; Rest, 1979a) and between people (e.g., Haidt et al., 1993; Petrinovich & O'Neill, 1996), and categorisation (e.g., Barsalou, 1987; McCloskey & Glucksberg, 1978). Patterns in the variability of both moral judgements and categorisation have been identified and in many cases specific instances of variability can be attributed to contextual factors. Many contextual factors have been shown to influence both moral judgements and categorisation. The contextual influences identified as common to both are: (a) order effects (priming); (b) Wording, language, and framing effects; (c) emotion effects; (d) developmental influences; (e) social influences; and (f) cultural influences. Beyond these contextual influences, theories of skill development have been proposed for both categorisation and moral judgement. Finally, two other phenomena have been identified as potentially being common to both categorisation, typicality and dumbfounding.

8.1.2 Contextual influences common to both morality and categorisation. Order effects in moral judgement were discussed in Chapter 1, in summary, responses to different dilemmas varied depending on what order they were presented (Lanteri et al., 2008; Liao et al., 2012; Lombrozo, 2009; Nichols & Mallon, 2006;

Petrinovich & O’Neill, 1996; Schwitzgebel & Cushman, 2012; Wiegmann et al., 2012). One explanation of these order effects in moral judgement is that they occur as a result of priming, that is that the scenario that is presented first causes some features of the second scenario to become more salient. The salience of these features led to a different judgement than if the initial scenario was not presented. The effect of this type of priming in categorisation is primarily studied in relation to reaction times. For example, a study by Barsalou (described above 1982, 2003) showed that reading sentences that made particular features of a given object salient influenced the speed at which participants verified related propertied of the given object. Similar effects have been identified by Tabossi and Johnson-Laird (Tabossi, 1988; Tabossi & Johnson-Laird, 1980). There is also evidence that priming people with particular concepts can influence their subsequent categorisations. In a study by Higgins, Bargh, and Lombardi (1985), participants completed a task in which they were required to create sentences from a selection of words. Some of the words presented were selected in order to prime a particular concept, e.g., “bold”, “courageous”, and “brave” primed “Adventurous”; “careless”, “foolhardy”, and “rash” primed “Reckless” (Higgins et al., 1985, p. 63). Participants were later presented with a description of an ambiguous behaviour. It was found that the categorisations of these behaviours were influenced by the concept that was primed. A similar study by Srull and Wyer demonstrated the same effect (Srull & Wyer, 1979).

Recall that changing the wording of the question relating to the trolley problem influenced the judgements made. Participants' willingness to agree or disagree with a statement advocating action or inaction varied depending on whether the statements included the word “death” or “saved” (Petrinovich & O’Neill, 1996).

Similar effects have been found in the non-moral domain whereby framing gambling decisions in terms of losses vs gains or certainty vs ambiguity influences the behaviour of participants (Kahneman, 2011). Referring to the 2006 soccer World Cup final, Kahneman highlights the difference in the meanings of the two sentences “Italy won” and “France lost” (Kahneman, 2011, p. 354). The first would lead people to view the match in terms of the merits of Italy's performance, while the second draws attention to mistakes that France may have made.

As in moral judgements, language influences (e.g., the foreign language effect Costa et al., 2014; Geipel et al., 2016; Hayakawa et al., 2017), have also been identified in categorisation. Consider a study by Boroditsky, Schmidt, and Phillips (2003) in which the gender of a noun in a person's native language was found to influence the adjectives generated in English to describe the noun. The study was conducted through English, with participants for whom English was not their native language. Participants were asked to generate adjectives to describe particular nouns such as “key” or “bridge”. In German the word for “key” is masculine while it is feminine in Spanish. “German speakers described keys as hard, heavy, jagged, metal, serrated, and useful, while Spanish speakers said they were golden, intricate, little, lovely, shiny, and tiny. The word for “bridge”, on the other hand, is feminine in German and masculine in Spanish. German speakers described bridges as beautiful, elegant, fragile, peaceful, pretty, and slender, while Spanish speakers said they were big, dangerous, long, strong, sturdy, and towering” (Boroditsky et al., 2003, p. 70). These differing adjectives would influence the categories that particular items would fall into (e.g., a key falling into HEAVY THINGS as opposed to LITTLE THINGS).

A language effect that is perhaps more similar to the foreign language effect identified in the moral domain has been described by Harris, Ayçiçeĝi, and Gleason (2003). They measured skin conductance of English speakers and Turkish speakers when rating different types of words in their first language and in their second language. It was found that (non-moral) taboo words led to greater arousal when presented in participants' first language than when presented in a second language.

Emotion, the most widely cited contextual influence on moral judgement (e.g., Cameron et al., 2013; Huebner et al., 2009; Royzman, Atanasov, et al., 2014; Rozin et al., 2009, 1999; Valdesolo & DeSteno, 2006; Wheatley & Haidt, 2005), has also been shown to influence judgements and categorisations in the non-moral domain. Various effects of emotion and mood on categorisation have been identified. Isen and Daubman (1984) reported that positive affect influenced the categorisation of words and colours. Positive affect was induced in participants using either an amusing clip or a free gift. Control groups received either no manipulation or viewed a short clip about mathematics. They then rated items on a 10 point scale as to whether they belonged to particular categories or not (where the midpoint was between 5 and 6, with 5 as not a member but resembling members and 6 as a member but not a typical member). It was found that atypical exemplars of categories were more often rated as members when positive affect had been induced than in the control condition. Similar findings from Murray, Sujan, Hirt, and Sujan (1990) showed that in categorisation tasks positive mood resulted in fewer broader categories. There is possibility for further study of the role of emotion in general categorisation tasks.

Other studies document the effect of emotion on categorisation of emotion. Wallbott (1991) videotaped participants as identified emotions on faces on a screen.

Analysis of the tapes (as described by Barsalou, 2003) indicated that participants partially simulated the emotions they were categorising. This partial simulation improved performance in the identification task. Two weeks after the task participants were able to correctly identify which emotion they were categorising by viewing the tape of their own face. Further evidence supporting the notion that correctly identifying emotion in others is aided by simulating their facial expression comes from Niedenthal et al. (2001) whereby preventing participants from simulating they expressions of others impaired emotion recognition. There are other findings that similarly support this idea (e.g., Adolphs, Damasio, Tranel, Cooper, & Damasio, 2000).

Both morality and categorisation display variability with development. Haidt, Koller, and Dias (1993) describe variability in the making of moral judgements as due to development. Overall, according to Haidt et al. (1993), when compared with adults (ages 19-26), children (aged 10-12) were more severe in their moral judgements, more likely to advocate punishment, and also more likely to condemn an act as universally wrong even if it is supposedly permissible in another culture than adults. The defining issues test (DIT, Rest, 1979a, 1979b, 1986), is an instrument for objectively measuring moral judgement. Participants are presented with several moral dilemmas and some specific considerations which they are asked to rate and rank in relation to the dilemma. It has been extensively used in the study of moral development. Drawing on 25 years worth of data from the DIT, Narvaez (2005) reports that a significant amount of variation observed in responses can be attributed to development. Developmental variability in categorisation may be seen as children acquire particular concepts; for example imagine a young child playing with their sea creature toy set, all of them look like fish and all are regarded as fish.

As the child grows older some of the toys become identified as whales as opposed to fish.

The extent to which social groups within society influence moral judgements can be readily seen in the treatment of divisive issues such as abortion or euthanasia. Similarly, it is likely that social factors (e.g., parenting, religious views, peer views) influence the categorisation of particular objects or events. Consider the emergence of variation in the categorisation of items depending on how a person interacts with the item as part of their life (e.g., a carpenter's view of wood compared with that of a firewood merchant).

One social influence present in both morality and categorisation is socio-economic status (SES). Haidt et al. (1993) showed that people of similar socio-economic status but from different countries, have more in common with each other than people from the same country but from a lower socio-economic background; i.e., that American college students had more in common (on moral grounds) with Brazilian college students than with fellow Americans from lower socio-economic status. Similarly, variability in non-moral categorisation has been found to be linked to socio-economic status. A study by Nelson and Klausmeier (1974) demonstrated that lower SES children classified objects according to observable likenesses and differences. Similarly, Björklund and Weiss (1985) found that children from higher SES showed a greater tendency to sort items according to taxonomic categories than their lower SES counterparts.

A detailed account of the type of variation between cultures in the making of moral judgements can be found in Haidt, Koller, and Dias (1993). Participants of varying age and socio-economic status from Brazil and America were presented with harmless morally questionable actions (e.g., using a national flag to clean the bathroom, or (used on adults only) a man masturbating with a chicken and then cooking and eating it). Participants were asked if the actions were wrong and then presented with a series of related questions. Cross-cultural variation was evident in that Americans were generally more permissible than Brazilians. This "westernised" approach to moral dilemmas focused more on harm than on whether or not the behaviour would cause offence if observed. Overall, the answer to the question "would it bother you?" was a better predictor of whether an action was rated as wrong or not than whether or not the action was perceived to cause harm. It is suggested that "westernised" morality is a more harm-based morality with focused deliberation in order to reach an objective judgement. Non-westernised morality relies more on affect towards the actions (Greene, 2008; Haidt et al., 1993).

Cross-cultural variation in categorisation, while not particularly widely documented is to be expected, emerging through differing cultural practices in using and naming particular category members (e.g., a knife as a member of the category CUTLERY in countries where chopsticks are used compared with in countries where knife and fork are used). This implicit expectation for cultural variation in categorisation has been identified in a study by Barsalou and Sewell (1984). In this study, American participants identified a robin as a typical example of the category BIRD in America, however they suggested that in China a peacock may be a more typical example. A more concrete example of cultural variation in categorisation, may be inferred from the way in which language shapes the categorisation of objects. Recall that the gender of a noun in a person's native language influences the adjectives used to describe that noun in English (Boroditsky et al., 2003).

The final parallel identified here is that skill based explanations have been proposed for both moral judgement (e.g., Dreyfus & Dreyfus, 1990; Hulsey &

Hampson, 2014; Narvaez, 2005) and categorisation (e.g., Barsalou, 2003; Barsalou, Breazeal, & Smith, 2007). This parallel provides the strongest argument for the alignment of the morality literature and the categorisation literature. It suggests parity between the underlying mechanisms of moral judgement and categorisation. Building on this, two further parallels are hypothesised here. Variation in the typicality of category membership is well documented (McCloskey & Glucksberg, 1978; Mervis & Rosch, 1981; Oden, 1977; Rosch, 1973b, 1973a; Rosch & Mervis, 1975). It is hypothesised here that behaviours can vary in their typicality in being identified as right or wrong; e.g., cold blooded murder versus violence in pursuit of a cause. Furthermore, it is possible that dumbfounding may occur for categories other than MORALLY WRONG. The work of Boyd and Keil (Boyd, 1989, 1991; as described by Griffiths, 1997; Keil, 1989) offers some suggestive evidence to this effect whereby children struggled to explain their reasons for categorising an imagined creature as A CAT or NOT A CAT. A summary of the parallels between categorisation and morality can be found in Table 8.1.

Table 8.1: Parallels between morality and categorisation

Phenomenon		Categorisation	Morality
Variability	Interpersonal	✔	✔
	Intrapersonal	✔	✔
Context	Culture	✔	✔
	Social	✔	✔
	Development	✔	✔
	Emotion	✔	✔
	Framing	✔	✔
	Language	✔	✔
	Order/recency	✔	✔
Other	Skill	✔	✔
	Typicality	✔	Hypothesised
	Dumbfounding	Hypothesised	✔

8.3.2 Limitations of a categorisation account of moral judgement. The categorisation account of moral judgement is consistent with the intuitionist approaches described in Chapter 1. Drawing on the categorisation literature provides an account of the mechanisms that give rise to the learning and making of moral judgements, or, the mechanisms that lead to the emergence and maintenance of moral intuitions, where a moral intuition a well-rehearsed skilled categorisation. Furthermore, it also offers a coherent explanation of the emergence of moral dumbfounding.

There are two general predictions (identified in Table 8.1) associated with the categorisation account of moral judgement adopted here. The first is that moral judgements should vary in typicality ratings. The second is that, under the right conditions, dumbfounding should also be observed in a non-moral domain. Moral dumbfounding is the phenomenon of interest for the current research. Testing either prediction is beyond the scope of the current research. Currently there are no materials to measure typicality ratings of moral behaviours, and the ease with which typicality could be conflated with other variables (e.g., severity) means that devising these materials would be an extensive new project. The work outlined in this thesis identifies the conditions and possible moderating variables that give rise to moral dumbfounding, providing an important foundation for testing the predictions of a categorisation approach to moral dumbfounding and extending the paradigm to a non-moral domain.

Beyond these two generalised predictions, the categorisation account of moral judgement is limited in its predictive power, particularly in relation to the phenomenon of interest, moral dumbfounding. The categorisation account predicts the existence of dumbfounding under the right conditions. However, it also predicts the possibility of contextual influences and variability. This means that while the eliciting of dumbfounding provides strong evidence for the approach, the failure to elicit dumbfounding does not necessarily provide evidence against the approach. Recall that the categorisation approach allows for the possibility of identifying rules that appear to govern category membership. Furthermore, if the conditions are not met, dumbfounding may not be elicited. There is limited empirical work investigating what the conditions necessary to elicit dumbfounding may be though the research presented in this thesis provides some first steps in that direction.

There are two problematic predictions associated with the categorisation account of moral judgement. The first relates to the undefined role of contextual influence: context effects on moral judgement are predicted, but the specific nature of these effects is not. The second relates the role of experience/personal history: identifying type-token interpretation as the underlying mechanism that gives rise to moral judgements means that the only factor that can really be identified as predicting moral judgements is personal history. There is no way to account for the personal history of every individual in such a way as to make the study of moral judgement meaningful. The identification of specific contextual factors that moderate judgements requires identifying contextual factors that presumably were reliably experienced during type-token interpretation (the learning and maintaining of moral categorisations) by the majority of people.

For this reason, by providing an account of the underlying mechanisms that may give rise to the emergence of moral intuitions, the categorisation account of moral judgement serves to complement rather than replace existing intuitionist theories of moral judgement. It provides less in the way of predictive power than the intuitionist theories described in Chapter 1, and at present does not provide any testable hypotheses that are significantly different from these.

8.5 Conclusion

Moral dumbfounding coincided with and, through the influential work of Haidt (e.g., Haidt, 2001), arguably contributed to the growth of intuitionist theories of moral judgement. The initial influence of moral dumbfounding on theories of moral judgement is clear, however in the in the 18 years since the original demonstration, the direct influence of moral dumbfounding on later theories of moral judgement has become less obvious. As the initial influence of moral dumbfounding on theories of moral judgement waned, so too did the ability of these theories to explain moral dumbfounding, and it remains poorly understood.

For many years, evidence for moral dumbfounding was limited to a single study, unpublished in peer-reviewed form, and with a final sample of just 30 participants. The paucity of evidence for moral dumbfounding has meant that very existence of the phenomenon has come under scrutiny with some authors suggesting that moral dumbfounding is not a real phenomenon (e.g., Jacobson, 2012; Kitcher, 2011; Royzman et al., 2015; Sneddon, 2007). The work presented in this thesis demonstrated that moral dumbfounding is indeed a real phenomenon that can be reliably evoked in a laboratory setting. In Chapter 3, I developed methods and materials for eliciting and studying dumbfounded responding. Chapter 4 provided evidence against the rationalist explanation of dumbfounded responding proposed by Royzman et al. (2015), demonstrating that people do not reliably articulate or apply principles that are claimed to be guiding their moral judgements.

Chapters 5 and 6 investigated two related explanations of moral dumbfounding and it was found that neither a general dual-process approach to moral judgement, nor model theory provide an adequate explanation of the complexities of dumbfounded responding. That moral dumbfounding, remains poorly explained by existing theories of moral judgement is a key limitation of the morality literature. The materials and methods developed in this thesis provide a means for further study of moral dumbfounding. Furthermore, a potential theoretical approach that does provide an explanation of moral dumbfounding was also explored in this final chapter. The work presented in this thesis identified a key limitation of the existing morality literature, and provided some of the tools (materials, theoretical framework) that may lead to this limitation being addressed. By providing an explanation of moral dumbfounding we will address limitations in the existing morality literature, and further our understanding of how we make moral judgements.

References

Abelson, R. P. (1988). Conviction. *American Psychologist*, *43*(4), 267–275. https://doi.org/10.1037/0003-066X.43.4.267

Acee, T. W., Kim, H., Kim, H. J., Kim, J.-I., Chu, H.-N. R., Kim, M., … Wicker, F. W. (2010). Academic boredom in under- and over-challenging situations. *Contemporary Educational Psychology*, *35*(1), 17–27. https://doi.org/10.1016/j.cedpsych.2009.08.002

Adolphs, R., Damasio, H., Tranel, D., Cooper, G., & Damasio, A. R. (2000). A Role for Somatosensory Cortices in the Visual Recognition of Emotion as Revealed by Three-Dimensional Lesion Mapping. *The Journal of Neuroscience*, *20*(7), 2683–2690.

Allard, T., Clark, S. A., Jenkins, W. M., & Merzenich, M. M. (1991). Reorganization of somatosensory area 3b representations in adult owl monkeys after digital syndactyly. *Journal of Neurophysiology*, *66*(3), 1048–1058.

Amazon Web Services Inc. (2016). *Amazon Mechanical Turk*.

Anthony, T., Copper, C., & Mullen, B. (1992). Cross-Racial Facial Identification: A Social Cognitive Integration. *Personality and Social Psychology Bulletin*, *18*(3), 296–301. https://doi.org/10.1177/0146167292183005

Asch, S. E. (1956). Studies of independence and submission to group pressures. *Psychological Monographs*, *70*, 416.

Aust, F. (2016). *citr: ‘RStudio’ Add-in to Insert Markdown Citations*. Retrieved from https://CRAN.R-project.org/package=citr

Aust, F., & Barth, M. (2017). *papaja: Create APA manuscripts with R Markdown*. Retrieved from https://github.com/crsh/papaja

Ballard, R. (1992). Short Forms of the Marlowe-Crowne Social Desirability Scale. *Psychological Reports*, *71*(3_suppl), 1155–1160. https://doi.org/10.2466/pr0.1992.71.3f.1155

Barrett, L. F., Wilson-Mendenhall, C. D., & Barsalou, L. W. (2014). A Psychological

Construction Account of Emotion Regulation and Dysregulation: The Role of Situated Conceptualizations. In J. J. Gross (Ed.), *Handbook of Emotion Regulation*. New York: Guilford Press.

Barsalou, L. W. (1982). Context-independent and context-dependent information in concepts. *Memory & Cognition*, *10*(1), 82–93. https://doi.org/10.3758/BF03197629

Barsalou, L. W. (1987). The instability of graded structure: Implications for the nature of concepts. In U. Neisser (Ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization* (pp. 101–140). Cambridge University Press.

Barsalou, L. W. (1991). Deriving categories to achieve goals. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 27, pp. 76–121). San Diego: Academic Press.

Barsalou, L. W. (1999). Perceptions of perceptual symbols. *Behavioral and Brain Sciences*, *22*(04), 637–660. https://doi.org/10.1017/S0140525X99532147

Barsalou, L. W. (2003). Situated simulation in the human conceptual system. *Language and Cognitive Processes*, *18*(5–6), 513–562. https://doi.org/10.1080/01690960344000026

Barsalou, L. W. (2005). Abstraction as Dynamic Interpretation in Perceptual Symbol Systems. In L. Gershkoff-Stowe & D. H. Rakison (Eds.), *Building object categories in developmental time*. Mahwah, N.J.: L. Erlbaum Associates.

Barsalou, L. W. (2008). Grounded Cognition. *Annual Review of Psychology*, *59*(1), 617–645. https://doi.org/10.1146/annurev.psych.59.103006.093639

Barsalou, L. W. (2009). Simulation, situated conceptualization, and prediction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1281–1289. https://doi.org/10.1098/rstb.2008.0319

Barsalou, L. W., Breazeal, C., & Smith, L. B. (2007). Cognition as coordinated non-cognition. *Cognitive Processing*, *8*(2), 79–91. https://doi.org/10.1007/s10339-

007-0163-1

Barsalou, L. W., & Sewell, D. R. (1984). Constructing representations of categories from different points of view. *Emory Cognition Project Technical Report# 2*.

Barsalou, L. W., & Wiemer-Hastings, K. (2005). Situating Abstract Concepts. In D. Pecher & R. A. Zwaan (Eds.), *Grounding Cognition: The Role of Perception and Action in Memory, Language, and Thinking* (pp. 129–163). Cambridge University Press.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bates, D., & Maechler, M. (2017). *Matrix: Sparse and Dense Matrix Classes and Methods*. Retrieved from https://CRAN.R-project.org/package=Matrix

Bechara, A., & Damasio, A. R. (2005). The somatic marker hypothesis: A neural theory of economic decision. *Games and Economic Behavior*, *52*(2), 336–372. https://doi.org/10.1016/j.geb.2004.06.010

Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (2005). The Iowa Gambling Task and the somatic marker hypothesis: some questions and answers. *Trends in Cognitive Sciences*, *9*(4), 159–162; discussion 162-164. https://doi.org/10.1016/j.tics.2005.02.002

Bellin, Z. (2012). The quest to capture personal meaning in psychology. *International Journal of Existential Psychology and Psychotherapy*, *4*(1), 27.

Berry, D., & Dienes, Z. P. (1993). *Implicit Learning: Theoretical and Empirical Issues*. Psychology Press.

Bialek, M., & Terbeck, S. (2016). Can cognitive psychological research on reasoning enhance the discussion around moral judgments? *Cognitive Processing*, *17*(3), 329–335. https://doi.org/10.1007/s10339-016-0760-y

Bjorklund, D. F., & Weiss, S. C. (1985). Influence of socioeconomic status on children’s classification and free recall. *Journal of Educational Psychology*, *77*(2), 119–128. https://doi.org/10.1037/0022-0663.77.2.119

Björklund, F., Haidt, J., & Murphy, S. (2000). Moral dumbfounding: when intuition finds no reason. *Lund Psychological Reports*, *Vol 1 no 2*. Retrieved from http://lup.lub.lu.se/record/1024827

Blair, R. J. R. (1995). A cognitive developmental approach to morality: investigating the psychopath. *Cognition*, *57*(1), 1–29. https://doi.org/10.1016/0010-0277(95)00676-P

Blair, R. J. R., Jones, L., Clark, F., & Smith, M. (1997). The psychopathic individual: A lack of responsiveness to distress cues? *Psychophysiology*, *34*(2), 192–198. https://doi.org/10.1111/j.1469-8986.1997.tb02131.x

Blair, R. J. R., Peschardt, K. s., Budhani, S., Mitchell, D. g. v., & Pine, D. s. (2006). The development of psychopathy. *Journal of Child Psychology and Psychiatry*, *47*(3–4), 262–276. https://doi.org/10.1111/j.1469-7610.2006.01596.x

Bonner, C., & Newell, B. R. (2010). In conflict with ourselves? An investigation of heuristic and analytic processes in decision making. *Memory & Cognition*, *38*(2), 186–196. https://doi.org/10.3758/MC.38.2.186

Borg, J. S., Hynes, C., van Horn, J., Grafton, S., & Sinnott-Armstrong, W. (2006). Consequences, Action, and Intention as Factors in Moral Judgments: An fMRI Investigation. *Journal of Cognitive Neuroscience*, *18*(5), 803–817.

Borg, J. S., Lieberman, D., & Kiehl, K. A. (2008). Infection, incest, and iniquity: Investigating the neural correlates of disgust and morality. *Journal of Cognitive Neuroscience*, *20*(9), 1529–1546.

Boroditsky, L., Schmidt, L., & Phillips, W. (2003). Sex, Syntax, and Semantics. In D. Gentner & S. Goldin-Meadow (Eds.), *Advances in the Study of Language and Thought* (pp. 61–80). Cambridge: MIT press.

Boyd, R. (1989). What Realism Implies and What it Does Not. *Dialectica*, *43*(1–2), 5–29. https://doi.org/10.1111/j.1746-8361.1989.tb00928.x

Boyd, R. (1991). Realism, anti-foundationalism and the enthusiasm for natural kinds. *Philosophical Studies*, *61*(1–2), 127–148. https://doi.org/10.1007/BF00385837

Brand, C. (2016). *Dual-Process Theories in Moral Psychology: Interdisciplinary Approaches to Theoretical, Empirical and Practical Considerations*. Springer.

Bucciarelli, M. (2009). What is Special about Children's Deontic Reasoning? In N.A. Taatgen & H. Van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 544–549). Austin, TX: Cognitive Science Society.

Bucciarelli, M., & Daniele, M. (2015). Reasoning in moral conflicts. *Thinking & Reasoning*, *21*(3), 265–294. https://doi.org/10.1080/13546783.2014.970230

Bucciarelli, M., & Johnson-Laird, P. N. (2005). Naïve deontics: A theory of meaning, representation, and reasoning. *Cognitive Psychology*, *50*(2), 159–193. https://doi.org/10.1016/j.cogpsych.2004.08.001

Bucciarelli, M., Khemlani, S., & Johnson-Laird, P. N. (2008). The psychology of moral reasoning. *Judgment and Decision Making*, *3*, 121–139.

Byrne, R. M. J. (2015). Mental Models. In *Emerging Trends in the Social and Behavioral Sciences*. John Wiley & Sons, Inc. https://doi.org/10.1002/9781118900772.etrds0217

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*(1), 116–131. https://doi.org/10.1037/0022-3514.42.1.116

Cameron, C. D., Payne, B. K., & Doris, J. M. (2013). Morality in high definition: Emotion differentiation calibrates the influence of incidental disgust on moral judgments. *Journal of Experimental Social Psychology*, *49*(4), 719–725. https://doi.org/10.1016/j.jesp.2013.02.014

Cannon, P. R., Schnall, S., & White, M. (2011). Transgressions and Expressions

Affective Facial Muscle Activity Predicts Moral Judgments. *Social Psychological and Personality Science*, *2*(3), 325–331. https://doi.org/10.1177/1948550610390525

Carter, S., & Smith Pasqualini, M. (2004). Stronger autonomic response accompanies better learning: A test of Damasio's somatic marker hypothesis. *Cognition & Emotion*, *18*(7), 901–911. https://doi.org/10.1080/02699930341000338

Case, D. O., Andrews, J. E., Johnson, J. D., & Allard, S. L. (2005). Avoiding Versus Seeking: The Relationship of Information Seeking to Avoidance, Blunting, Coping, Dissonance, and Related Concepts. *Journal of the Medical Library Association : JMLA*, *93*(3), 353–362.

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology*, *39*(5), 752–766. https://doi.org/10.1037/0022-3514.39.5.752

Chaiken, S., & Trope, Y. (1999). *Dual-process Theories in Social Psychology*. Guilford Press.

Champely, S. (2018). *pwr: Basic Functions for Power Analysis*. Retrieved from https://CRAN.R-project.org/package=pwr

Chapman, H. A. (2018). A Component Process model of Disgust, Anger, and Moral Judgment. In K. J. Gray & J. Graham (Eds.), *Atlas of Moral Psychology* (pp. 70–80). New York, NY: The Guilford Press.

Chomsky, N. A. (1965). *Aspects of the Theory of Syntax* (Vol. 11). MIT press.

Chomsky, N. A. (1976). *Reflections on language*.

Chomsky, N. A. (2000). *New Horizons in the Study of Language and Mind*. Cambridge University Press.

Christensen, J. F., Flexas, A., Calabrese, M., Gut, N. K., & Gomila, A. (2014). Moral judgment reloaded: a moral dilemma validation study. *Emotion Science*, *5*,

607. https://doi.org/10.3389/fpsyg.2014.00607

Christensen, J. F., & Gomila, A. (2012). Moral dilemmas in cognitive neuroscience of moral decision-making: A principled review. *Neuroscience & Biobehavioral Reviews*, *36*(4), 1249–1264. https://doi.org/10.1016/j.neubiorev.2012.02.008

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, *31*(5), 489–509. https://doi.org/10.1017/S0140525X08004998

Chung, J., & Monroe, G. S. (2003). Exploring Social Desirability Bias. *Journal of Business Ethics*, *44*(4), 291–302. https://doi.org/10.1023/A:1023648703356

Cialdini, R. B., Trost, M. R., & Newsom, J. T. (1995). Preference for consistency: The development of a valid measure and the discovery of surprising behavioral implications. *Journal of Personality and Social Psychology*, *69*(2), 318–328. https://doi.org/10.1037/0022-3514.69.2.318

Cooper, J. (2007). *Cognitive dissonance: Fifty years of a classic theory* (Vol. xi). Thousand Oaks, CA: Sage Publications Ltd.

Costa, A., Foucart, A., Hayakawa, S., Aparici, M., Apesteguia, J., Heafner, J., & Keysar, B. (2014). Your Morals Depend on Language. *PLOS ONE*, *9*(4), e94842. https://doi.org/10.1371/journal.pone.0094842

Cowley, M. B., & Byrne, R. M. J. (2005). *When Falsification is the Only Path to Truth* (SSRN Scholarly Paper No. ID 2339585). Rochester, NY: Social Science Research Network. Retrieved from https://papers.ssrn.com/abstract=2339585

Crockett, M. J. (2013). Models of morality. *Trends in Cognitive Sciences*, *17*(8), 363–366. https://doi.org/10.1016/j.tics.2013.06.005

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, *24*(4), 349–354. https://doi.org/10.1037/h0047358

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLOS ONE*, *8*(3), e57410. https://doi.org/10.1371/journal.pone.0057410

Cushman, F. A. (2013). Action, Outcome, and Value A Dual-System Framework for Morality. *Personality and Social Psychology Review*, *17*(3), 273–292. https://doi.org/10.1177/1088868313495594

Cushman, F. A., Young, L., & Greene, J. D. (2010). Multi-system Moral Psychology. In J. M. Doris (Ed.), *The Moral Psychology Handbook* (pp. 47–71). Oxford; New York: Oxford University Press.

Cushman, F. A., Young, L., & Hauser, M. D. (2006). The Role of Conscious Reasoning and Intuition in Moral Judgment Testing Three Principles of Harm. *Psychological Science*, *17*(12), 1082–1089. https://doi.org/10.1111/j.1467-9280.2006.01834.x

Damasio, A. R. (1994). *Descartes' error: emotion, reason, and the human brain*. New York: Putnam.

Daniels, N. (1989). *Reading Rawls: Critical Studies on Rawls' A Theory of Justice*. Stanford University Press.

David, B., & Olatunji, B. O. (2011). The effect of disgust conditioning and disgust sensitivity on appraisals of moral transgressions. *Personality and Individual Differences*, *50*(7), 1142–1146. https://doi.org/10.1016/j.paid.2011.02.004

Davidson, P., Turiel, E., & Black, A. (1983). The effect of stimulus familiarity on the use of criteria and justifications in children's social reasoning. *British Journal of Developmental Psychology*, *1*(1), 49–65. https://doi.org/10.1111/j.2044-835X.1983.tb00543.x

De Neys, W. (2006). Dual Processing in Reasoning: Two Systems but One Reasoner. *Psychological Science*, *17*(5), 428–433. https://doi.org/10.1111/j.1467-9280.2006.01723.x

De Neys, W. (2012). Bias and Conflict: A Case for Logical Intuitions. *Perspectives* *on Psychological Science*, *7*(1), 28–38. https://doi.org/10.1177/1745691611429354

De Neys, W. (2014). Conflict detection, dual processes, and logical intuitions: Some clarifications. *Thinking & Reasoning*, *20*(2), 169–187. https://doi.org/10.1080/13546783.2013.854725

De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition*, *106*(3), 1248–1299. https://doi.org/10.1016/j.cognition.2007.06.002

De Neys, W., & Schaeken, W. (2007). When people are more logical under cognitive load: Dual task impact on scalar implicature. *Experimental Psychology*, *54*(2), 128–133. https://doi.org/10.1027/1618-3169.54.2.128

DeLancey, C. (2001). *Passionate Engines : What Emotions Reveal about the Mind and Artificial Intelligence: What Emotions Reveal about the Mind and Artificial Intelligence*. Oxford: Oxford University Press.

Dewey, M. (2017). *metap: meta-analysis of significance values*.

Doris, J. M. (Ed.). (2010). *The Moral Psychology Handbook*. Oxford; New York: Oxford University Press.

Doris, J. M., & Plakias, A. (2008). How to Argue about Disagreement: Evaluative Diversity and Moral Realism. In W. Sinnott-Armstrong (Ed.), *Moral psychology Volume 2, The cognitive science of morality: intuition and diversity* (pp. 47–76). London: MIT.

Dreyfus, H. L., & Dreyfus, S. E. (1990). What is moral maturity? A phenomenological account of the development of ethical expertise. *Universalism vs. Communitarianism*, 237–264.

Dunn, B. D., Dalgleish, T., & Lawrence, A. D. (2006). The somatic marker hypothesis: A critical evaluation. *Neuroscience & Biobehavioral Reviews*, *30*(2), 239–271. https://doi.org/10.1016/j.neubiorev.2005.07.001

Dupoux, E., & Jacob, P. (2007). Universal moral grammar: a critical appraisal.

Trends in Cognitive Sciences, *11*(9), 373–378. https://doi.org/10.1016/j.tics.2007.07.001

Dupoux, E., & Jacob, P. (2008). Response to Dwyer and Hauser: Sounding the retreat? *Trends in Cognitive Sciences*, *12*(1), 2–3. https://doi.org/10.1016/j.tics.2007.10.009

Dwyer, S. (2009). Moral Dumbfounding and the Linguistic Analogy: Methodological Implications for the Study of Moral Judgment. *Mind & Language*, *24*(3), 274–296. https://doi.org/10.1111/j.1468-0017.2009.01363.x

Dwyer, S., & Hauser, M. D. (2008). Dupoux and Jacob's moral instincts: throwing out the baby, the bathwater and the bathtub. *Trends in Cognitive Sciences*, *12*(1), 1–2. https://doi.org/10.1016/j.tics.2007.10.006

Eden, A., & Tamborini, R. (2016). Moral Intuitions: Morality Subcultures in Disposition Formation. *Journal of Media Psychology: Theories, Methods, and Applications*. https://doi.org/10.1027/1864-1105/a000173

Elfenbein, H. A., & Ambady, N. (2002). Is there an in-group advantage in emotion recognition? *Psychological Bulletin*, *128*(2), 243–249. https://doi.org/10.1037/0033-2909.128.2.243

Epstein, S. (1994). Integration of the cognitive and the psychodynamic unconscious. *American Psychologist*, *49*(8), 709–724. https://doi.org/10.1037/0003-066X.49.8.709

Epstein, S., & Pacini, R. (1999). Some Basic Issues Regarding Dual-Process Theories from the Perspective of Cognitive-Experimental Self Theory. In S. Chaiken & Y. Trope (Eds.), *Dual-process Theories in Social Psychology* (pp. 462–482). Guilford Press.

Eskine, K. J., Kacinik, N. A., & Prinz, J. J. (2011). A Bad Taste in the Mouth: Gustatory Disgust Influences Moral Judgment. *Psychological Science*, *22*(3), 295–299. https://doi.org/10.1177/0956797611398497

Evans, J. S. B. T. (1989). *Bias in human reasoning: Causes and consequences* (Vol.

ix). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Evans, J. S. B. T. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin & Review*, *13*(3), 378–395. https://doi.org/10.3758/BF03193858

Evans, J. S. B. T. (2007). On the resolution of conflict in dual process theories of reasoning. *Thinking & Reasoning*, *13*(4), 321–339. https://doi.org/10.1080/13546780601008825

Evans, J. S. B. T. (2008). Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition. *Annual Review of Psychology*, *59*(1), 255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629

Evans, J. S. B. T. (2010). *Thinking twice: two minds in one brain*. New York: Oxford University Press.

Evans, J. S. B. T. (2011). Dual-process theories of reasoning: Contemporary issues and developmental applications. *Developmental Review*, *31*(2–3), 86–102. https://doi.org/10.1016/j.dr.2011.07.007

Evans, J. S. B. T., & Curtis-Holmes, J. (2005). Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning. *Thinking & Reasoning*, *11*(4), 382–389. https://doi.org/10.1080/13546780542000005

Evans, J. S. B. T., & Over, D. E. (2013). *Rationality and Reasoning*. Psychology Press.

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford CA: Stanford University Press.

Fine, C. (2006). Is the emotional dog wagging its rational tail, or chasing it? *Philosophical Explorations*, *9*(1), 83–98. https://doi.org/10.1080/13869790500492680

Fischer, D. G., & Fick, C. (1993). Measuring Social Desirability: Short Forms of the Marlowe-Crowne Social Desirability Scale. *Educational and Psychological Measurement*, *53*(2), 417–424. https://doi.org/10.1177/0013164493053002011

Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review*, (5), 5–15.

Forsterlee, R., & Ho, R. (1999). An Examination of the Short form of the Need for Cognition Scale Applied in an Australian Sample. *Educational and Psychological Measurement*, *59*(3), 471–480. https://doi.org/10.1177/00131649921969983

Fox, J., & Weisberg, S. (2011). *An R Companion to Applied Regression* (Second). Thousand Oaks CA: Sage. Retrieved from http://socserv.socsci.mcmaster.ca/jfox/Books/Companion

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, *19*(4), 25–42.

Freiman, C., & Nichols, S. (2011). Is Desert in the Details? *Philosophy and Phenomenological Research*, *82*(1), 121–133. https://doi.org/10.1111/j.1933-1592.2010.00387.x

Friard, O., & Gamba, M. (2015). BORIS - Behavioral Observation Research Interactive Software (Version 2.72). Italy. Retrieved from http://www.boris.unito.it

Fujita, K., Trope, Y., Liberman, N., & Levin-Sagi, M. (2006). Construal Levels and Self-Control. *Journal of Personality and Social Psychology*, *90*(3), 351–367. https://doi.org/10.1037/0022-3514.90.3.351

Geipel, J., Hadjichristidis, C., & Surian, L. (2016). Foreign language affects the contribution of intentions and outcomes to moral judgment. *Cognition*, *154*, 34–39. https://doi.org/10.1016/j.cognition.2016.05.010

Gigerenzer, G. (2008). Moral Intuition = Fast and Frugal Heuristics. In W. Sinnott-

Armstrong (Ed.), *Moral psychology, Volume 2, The cognitive science of morality: intuition and diversity*. London: MIT.

Giner-Sorolla, R. (2018). A Functional Conflict Theory of Moral Emotions. In K. J. Gray & J. Graham (Eds.), *Atlas of Moral Psychology* (pp. 81–87). New York, NY: The Guilford Press.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples. *Journal of Behavioral Decision Making*, *26*(3), 213–224. https://doi.org/10.1002/bdm.1753

Gray, K. J., Schein, C., & Ward, A. F. (2014). The myth of harmless wrongs in moral cognition: Automatic dyadic completion from sin to suffering. *Journal of Experimental Psychology: General*, *143*(4), 1600–1615. https://doi.org/10.1037/a0036149

Greene, J. D. (2007). Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains. *Trends in Cognitive Sciences*, *11*(8), 322–323.

Greene, J. D. (2008). The Secret Joke of Kant's Soul. In W. Sinnott-Armstrong, *Moral Psychology Volume 3: The neurosciences of morality: emotion, brain disorders, and development* (pp. 35–79). Cambridge (Mass.): the MIT press.

Greene, J. D. (2013). *Moral tribes: emotion, reason, and the gap between us and them*.

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition*, *107*(3), 1144–1154. https://doi.org/10.1016/j.cognition.2007.11.004

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science (New York, N.Y.)*, *293*(5537), 2105–2108.

https://doi.org/10.1126/science.1062872

Greenspan, S. L. (1986). Semantic flexibility and referential specificity of concrete nouns. *Journal of Memory and Language*, *25*(5), 539–557. https://doi.org/10.1016/0749-596X(86)90010-0

Griffiths, P. E. (1997). *What Emotions Really Are: The Problem of Psychological Categories*. Chicago: University of Chicago Press.

Gubbins, E., & Byrne, R. M. J. (2014). Dual processes of emotion and reason in judgments about moral dilemmas. *Thinking & Reasoning*, *20*(2), 245–268. https://doi.org/10.1080/13546783.2013.877400

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–834. https://doi.org/10.1037/0033-295X.108.4.814

Haidt, J. (2007). The New Synthesis in Moral Psychology. *Science*, *316*(5827), 998–1002. https://doi.org/10.1126/science.1137651

Haidt, J., & Björklund, F. (2008). Social Intuitionists Answer Six Questions about Moral Psychology. In W. Sinnott-Armstrong (Ed.), *Moral psychology Volume 2, The cognitive science of morality: intuition and diversity* (pp. 181–217). London: MIT.

Haidt, J., Björklund, F., & Murphy, S. (2000). Moral dumbfounding: When intuition finds no reason. *Unpublished Manuscript, University of Virginia*.

Haidt, J., & Hersh, M. A. (2001). Sexual Morality: The Cultures and Emotions of Conservatives and Liberals. *Journal of Applied Social Psychology*, *31*(1), 191–221. https://doi.org/10.1111/j.1559-1816.2001.tb02489.x

Haidt, J., Koller, S. H., & Dias, M. G. (1993). Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology*, *65*(4), 613–628. https://doi.org/10.1037/0022-3514.65.4.613

Hammond, K. R. (1996). *Human judgement and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice* (Vol. xi). New York, NY, US: Oxford University Press.

Harman, G. (2000). *Explaining Value and other Essays in Moral Philosophy*. Clarendon Press.

Harman, G., Mason, K., & Sinnott-Armstrong, W. (2010). Moral Reasoning. In J. M. Doris (Ed.), *The Moral Psychology Handbook* (pp. 206–245). Oxford; New York: Oxford University Press.

Harmon-Jones, E., & Harmon-Jones, C. (2007). Cognitive Dissonance Theory After 50 Years of Development. *Zeitschrift Für Sozialpsychologie*, *38*(1), 7–16. https://doi.org/10.1024/0044-3514.38.1.7

Harris, C. L., Ayçiçeĝi, A., & Gleason, J. B. (2003). Taboo words and reprimands elicit greater autonomic reactivity in a first language than in a second language. *Applied Psycholinguistics*, *24*(4), 561–579. https://doi.org/10.1017/S0142716403000286

Hauser, M. D. (2006a). *Moral minds: How nature designed our universal sense of right and wrong.* New York: Harper Collins.

Hauser, M. D. (2006b). The liver and the moral organ. *Social Cognitive and Affective Neuroscience*, *1*(3), 214–220. https://doi.org/10.1093/scan/nsl026

Hauser, M. D., Cushman, F. A., Young, L., Kang-Xing Jin, R., & Mikhail, J. (2007). A Dissociation Between Moral Judgments and Justifications. *Mind & Language*, *22*(1), 1–21. https://doi.org/10.1111/j.1468-0017.2006.00297.x

Hauser, M. D., Young, L., & Cushman, F. A. (2008). Reviving Rawls's Linguistic Analogy: Operative Principles and the Causal Structure of Moral Actions. In W. Sinnott-Armstrong (Ed.), *Moral psychology Volume 2, The cognitive science of morality: intuition and diversity* (pp. 107–155). London: MIT.

Hayakawa, S., Tannenbaum, D., Costa, A., Corey, J. D., & Keysar, B. (2017). Thinking More or Feeling Less? Explaining the Foreign-Language Effect on Moral Judgment. *Psychological Science*, 0956797617720944. https://doi.org/10.1177/0956797617720944

Heine, S. J., Proulx, T., & Vohs, K. D. (2006). The Meaning Maintenance Model: On the Coherence of Social Motivations. *Personality and Social Psychology Review*, *10*(2), 88–110. https://doi.org/10.1207/s15327957pspr1002_1

Higgins, E. T., Bargh, J. A., & Lombardi, W. (1985). Nature of Priming Effects on Categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(1), 59–69.

Hinzen, W. (2012). The philosophical significance of Universal Grammar. *Language Sciences*, *34*(5), 635–649. https://doi.org/10.1016/j.langsci.2012.03.005

Huber, S., & Huber, O. W. (2012). The Centrality of Religiosity Scale (CRS). *Religions*, *3*(3), 710–724. https://doi.org/10.3390/rel3030710

Huebner, B., Dwyer, S., & Hauser, M. D. (2009). The role of emotion in moral psychology. *Trends in Cognitive Sciences*, *13*(1), 1–6.

Hulsey, T. L., & Hampson, P. J. (2014). Moral expertise. *New Ideas in Psychology*, *34*, 1–11. https://doi.org/10.1016/j.newideapsych.2014.02.001

Hume, D. (2000). An enquiry concerning the principles of morals. *Hume Studies*, *26*(2), 344–346.

IBM Corp. (2015). SPSS (Version 24.0) [WIndows]. Armonk, NY: IBM Corp.

Isen, A. M., & Daubman, K. A. (1984). The influence of affect on categorization. *Journal of Personality and Social Psychology*, *47*(6), 1206–1217. https://doi.org/10.1037/0022-3514.47.6.1206

Jacobson, D. (2008). Does Social Intuitionism Flatter Morality or Challenge It? In W. Sinnott-Armstrong (Ed.), *Moral psychology Volume 2, The cognitive science of morality: intuition and diversity* (pp. 219–232). London: MIT.

Jacobson, D. (2012). Moral Dumbfounding and Moral Stupefaction. In *Oxford studies in normative ethics* (Vol. 2, p. 289).

Johnson-Laird, P. N. (2006). *How we reason*. Oxford ; New York: Oxford University Press.

Jones, A., & Fitness, J. (2008). Moral hypervigilance: The influence of disgust sensitivity in the moral domain. *Emotion*, *8*(5), 613–627. https://doi.org/10.1037/a0013435

Juhos, C., Quelhas, A. C., & Byrne, R. M. J. (2015). Reasoning about intentions: Counterexamples to reasons for actions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(1), 55–76. https://doi.org/10.1037/a0037274

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *99*(1), 122–149. https://doi.org/10.1037/0033-295X.99.1.122

Kahneman, D. (1973). *Attention and Effort*. Prentice-Hall Inc.

Kahneman, D. (2011). *Thinking, fast and slow*. London: Allen Lane.

Kant, I. (1959). *Foundation of the metaphysics of morals (L.W. Beck, Trans.)*. Indianapolis: Bobbs-Merrill.

Keil, F. C. (1989). *Concepts, kinds, and cognitive development* (Vol. xv). Cambridge, MA, US: The MIT Press.

Kennett, J., & Fine, C. (2009). Will the Real Moral Judgment Please Stand Up? *Ethical Theory and Moral Practice*, *12*(1), 77–96. https://doi.org/10.1007/s10677-008-9136-4

Kitcher, P. (2011). *The Ethical Project*. Harvard University Press.

Kohlberg, L. (1969). *Stages in the development of moral thought and action*. New York: Holt, Rinehart & Winston.

Kohlberg, L. (1971). *From is to Ought: How to Commit the Naturalistic Fallacy and Get Away with it in the Study of Moral Development*.

Kohlberg, L. (1985). Kohlberg's stages of moral development. *Theories of Development. Upper Saddle River, NJ: Prentice-Hall*, 118–136.

Kross, E., & Ayduk, O. (2008). Facilitating Adaptive Emotional Analysis: Distinguishing Distanced-Analysis of Depressive Experiences From Immersed-Analysis and Distraction. *Personality and Social Psychology*

Bulletin, *34*(7), 924–938. https://doi.org/10.1177/0146167208315938

Kruglanski, A. W. (2013). *The Psychology of Closed Mindedness*. Psychology Press.

Kruglanski, A. W., Atash, M. N., De Grada, E., Mannetti, L., & Pierro, A. (2013). *Need for Closure Scale (NFC). Measurement Instrument Database for the Social Science*.

Kruglanski, A. W., & Webster, D. M. (1996). Motivated closing of the mind: 'Seizing' and 'freezing.' *Psychological Review*, *103*(2), 263–283. https://doi.org/10.1037/0033-295X.103.2.263

Kruglanski, A. W., Webster, D. M., & Klem, A. (1993). Motivated resistance and openness to persuasion in the presence or absence of prior information. *Journal of Personality and Social Psychology*, *65*(5), 861–876. https://doi.org/10.1037/0022-3514.65.5.861

Landy, J. F., & Goodwin, G. P. (2015). Does Incidental Disgust Amplify Moral Judgment? A Meta-Analytic Review of Experimental Evidence. *Perspectives on Psychological Science*, *10*(4), 518–536. https://doi.org/10.1177/1745691615583128

Lanteri, A., Chelini, C., & Rizzello, S. (2008). An Experimental Investigation of Emotions and Reasoning in the Trolley Problem. *Journal of Business Ethics*, *83*(4), 789–804. https://doi.org/10.1007/s10551-008-9665-8

Latif, D. A. (2000). The Link Between Moral Reasoning Scores, Social Desirability, and Patient Care Performance Scores: Empirical Evidence from the Retail Pharmacy Setting. *Journal of Business Ethics*, *25*(3), 255–269. https://doi.org/10.1023/A:1006049605298

Lenth, R. V. (2016a). *estimability: Tools for Assessing Estimability of Linear Predictions*. Retrieved from https://CRAN.R-project.org/package=estimability

Lenth, R. V. (2016b). Least-Squares Means: The R Package lsmeans. *Journal of Statistical Software*, *69*(1), 1–33. https://doi.org/10.18637/jss.v069.i01

Lerner, M. J., & Goldberg, J. H. (1999). When Do Decent People Blame Victims? The Differing Effects of the Explicit/Rational and Implicit/Experiential Cognitive Systems. In S. Chaiken & Y. Trope (Eds.), *Dual-process Theories in Social Psychology* (pp. 627–640). Guilford Press.

Levitin, D. J. (1994). Absolute memory for musical pitch: Evidence from the production of learned melodies. *Perception & Psychophysics*, *56*(4), 414–423. https://doi.org/10.3758/BF03206733

Liao, S. M. (2011). Bias and Reasoning: Haidt's Theory of Moral Judgment. In *New Waves in Ethics* (pp. 108–127). Palgrave Macmillan, London. https://doi.org/10.1057/9780230305885_7

Liao, S. M., Wiegmann, A., Alexander, J., & Vong, G. (2012). Putting the trolley in order: Experimental philosophy and the loop case. *Philosophical Psychology*, *25*(5), 661–671. https://doi.org/10.1080/09515089.2011.627536

Liberman, N., Sagristano, M. D., & Trope, Y. (2002). The effect of temporal distance on level of mental construal. *Journal of Experimental Social Psychology*, *38*(6), 523–534. https://doi.org/10.1016/S0022-1031(02)00535-8

Liberman, N., & Trope, Y. (1998). The Role of Feasibility and Desirability Considerations in Near and Distant Future Decisions: A Test of Temporal Construal Theory. *Journal of Personality and Social Psychology*, *75*(1), 5–18.

Liberman, N., & Trope, Y. (2008). The Psychology of Transcending the Here and Now. *Science*, *322*(5905), 1201–1205. https://doi.org/10.1126/science.1161958

Liberman, N., Trope, Y., & Stephan, E. (2007). Psychological distance. In A. W. Kruglanski & E. T. Higgins (Eds.), *Social Psychology: Handbook of Basic Principles*.

Liljenquist, K., Zhong, C.-B., & Galinsky, A. D. (2010). The Smell of Virtue Clean Scents Promote Reciprocity and Charity. *Psychological Science*, *21*(3), 381–

383. https://doi.org/10.1177/0956797610361426

Lombrozo, T. (2009). The Role of Moral Commitments in Moral Judgment. *Cognitive Science*, *33*(2), 273–286. https://doi.org/10.1111/j.1551-6709.2009.01013.x

Lüdecke, D. (2018). *sjstats: Statistical Functions for Regression Models*. Retrieved from https://CRAN.R-project.org/package=sjstats

Machery, E. (2010). Explaining why experimental behavior varies across cultures: A missing step in “The weirdest people in the world?” *Behavioral and Brain Sciences*, *33*(2–3), 101–102. https://doi.org/10.1017/S0140525X10000178

Machery, E. (2012). Delineating the Moral Domain. *Baltic International Yearbook of Cognition, Logic and Communication*, *7*(1). https://doi.org/10.4148/biyclc.v7i0.1777

Machery, E., & Mallon, R. (2010). Evolution of Morality. In J. M. Doris (Ed.), *The Moral Psychology Handbook* (pp. 3–46). Oxford; New York: Oxford University Press.

MacNab, S. (2016, January 23). MSPs to consider ‘abhorrent’ call to legalise incest. *The Scotsman*. Retrieved from http://www.scotsman.com/news/politics/msps-to-consider-abhorrent-call-to-legalise-incest-1-4009185

Maguire, E. A., Woollett, K., & Spiers, H. J. (2006). London taxi drivers and bus drivers: A structural MRI and neuropsychological analysis. *Hippocampus*, *16*(12), 1091–1101. https://doi.org/10.1002/hipo.20233

Maia, T. V., & McClelland, J. L. (2004). A reexamination of the evidence for the somatic marker hypothesis: What participants really know in the Iowa gambling task. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(45), 16075–16080. https://doi.org/10.1073/pnas.0406666101

Maia, T. V., & McClelland, J. L. (2005). The somatic marker hypothesis: still many questions but no answers: Response to Bechara et al. *Trends in Cognitive*

Sciences, *9*(4), 162–164. https://doi.org/10.1016/j.tics.2005.02.006

Mallon, R. (2008). Reviving Rawls's Linguistic Analogy Inside and Out. In W. Sinnott-Armstrong (Ed.), *Moral psychology Volume 2, The cognitive science of morality: intuition and diversity* (pp. 146–155). London: MIT. Retrieved from http://capitadiscovery.co.uk/mic/items/163688

Mallon, R., & Nichols, S. (2011). Dual Processes and Moral Rules. *Emotion Review*, *3*(3), 284–285. https://doi.org/10.1177/1754073911402376

Marwick, B. (n.d.). *wordcountaddin: Word counts and readability statistics in R markdown documents*.

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314–324. https://doi.org/10.3758/s13428-011-0168-7

May, J. (2014). Does Disgust Influence Moral Judgment? *Australasian Journal of Philosophy*, *92*(1), 125–141.

McCloskey, M. E., & Glucksberg, S. (1978). Natural categories: Well defined or fuzzy sets? *Memory & Cognition*, *6*(4), 462–472. https://doi.org/10.3758/BF03197480

McGregor, I. (2006a). Offensive Defensiveness: Toward an Integrative Neuroscience of Compensatory Zeal After Mortality Salience, Personal Uncertainty, and Other Poignant Self-Threats. *Psychological Inquiry*, *17*(4), 299–308. https://doi.org/10.1080/10478400701366977

McGregor, I. (2006b). Zeal Appeal: The Allure of Moral Extremes. *Basic and Applied Social Psychology*, *28*(4), 343–348. https://doi.org/10.1207/s15324834basp2804_7

McGregor, I., Zanna, M. P., Holmes, J. G., & Spencer, S. J. (2001). Compensatory conviction in the face of personal uncertainty: Going to extremes and being oneself. *Journal of Personality and Social Psychology*, *80*(3), 472–488. https://doi.org/10.1037/0022-3514.80.3.472

McGuire, J., Langdon, R., Coltheart, M., & Mackenzie, C. (2009). A reanalysis of the personal/impersonal distinction in moral psychology research. *Journal of Experimental Social Psychology*, *45*(3), 577–580. https://doi.org/10.1016/j.jesp.2009.01.002

McHugh, C. (2017). *desnum: Creates some useful functions*. Retrieved from https://github.com/cillianmiltown/R_desnum

McHugh, C., McGann, M., Igou, E. R., & Kinsella, E. L. (2017). Searching for Dumbfounding. *Open Science Framework*. https://doi.org/None

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, *7*(1), 3–35. https://doi.org/10.1037/1076-8971.7.1.3

Mendez, M. F., Anderson, E., & Shapira, J. S. (2005). An Investigation of Moral Judgement in Frontotemporal Dementia. *Cognitive and Behavioral Neurology*, *18*(4), 193–197.

Mervis, C. B., & Rosch, E. H. (1981). Categorization of Natural Objects. *Annual Review of Psychology*, *32*(1), 89–115. https://doi.org/10.1146/annurev.ps.32.020181.000513

Mesquita, B., Barrett, L. F., & Smith, E. R. (2010). *The Mind in Context*. New York: Guilford Press.

Metcalfe, J., & Mischel, W. (1999). A hot/cool-system analysis of delay of gratification: Dynamics of willpower. *Psychological Review*, *106*(1), 3–19. https://doi.org/10.1037/0033-295X.106.1.3

Mikhail, J. (2000). *Rawls' Linguistic Analogy: A Study of the 'Generative Grammar' Model of Moral Theory Described by John Rawls in 'A Theory of Justice.' (Phd Dissertation, Cornell University, 2000)* (SSRN Scholarly Paper No. ID 766464). Rochester, NY: Social Science Research Network. Retrieved from https://papers.ssrn.com/abstract=766464

Mikhail, J. (2007). Universal moral grammar: theory, evidence and the future.

Trends in Cognitive Sciences, *11*(4), 143–152. https://doi.org/10.1016/j.tics.2006.12.007

Milgram, S. (1974). *Obedience to Authority: An Experimental View*. New York: Harper and Row.

Morris, S. A., & McDonald, R. A. (2013). The Role of Moral Intensity in Moral Judgments: An Empirical Investigation. In A. C. Michalos & D. C. Poff (Eds.), *Citation Classics from the Journal of Business Ethics* (pp. 463–479). Springer Netherlands. https://doi.org/10.1007/978-94-007-4126-3_23

MSPs throw out incest petition. (2016, January 26). *BBC News*. Retrieved from http://www.bbc.com/news/uk-scotland-scotland-politics-35401195

Mugg, J. (2015). The dual-process turn: How recent defenses of dual-process theories of reasoning fail. *Philosophical Psychology*, *0*(0), 1–10. https://doi.org/10.1080/09515089.2015.1078458

Murray, N., Sujan, H., Hirt, E. R., & Sujan, M. (1990). The influence of mood on categorization: A cognitive flexibility interpretation. *Journal of Personality and Social Psychology*, *59*(3), 411–425. https://doi.org/10.1037/0022-3514.59.3.411

Nadelhoffer, T., & Feltz, A. (2008). The Actor–Observer Bias and Moral Intuitions: Adding Fuel to Sinnott-Armstrong's Fire. *Neuroethics*, *1*(2), 133–144. https://doi.org/10.1007/s12152-008-9015-7

Nakamura, K. (2013). A closer look at moral dilemmas: Latent dimensions of morality and the difference between trolley and footbridge dilemmas. *Thinking & Reasoning*, *19*(2), 178–204.

Narvaez, D. (2005). The neo-Kohlbergian tradition and beyond: Schemas, expertise, and character. In G. Carlo & C. Pope-Edwards (Eds.), *Nebraska symposium on motivation* (Vol. 51, p. 119).

Narvaez, D. (2008). The Social Intuitionist Model: Some Counter-Intuitions. In W. Sinnott-Armstrong (Ed.), *Moral psychology Volume 2, The cognitive science* *of morality: intuition and diversity* (pp. 233–240). London: MIT.

Narvaez, D., & Lapsley, D. K. (2005). The psychological foundations of everyday morality and moral expertise. *Character Psychology and Character Education*, 140–165.

Nelson, G. K., & Klausmeier, H. J. (1974). Classificatory behaviors of low-socioeconomic-status children. *Journal of Educational Psychology*, *66*(3), 432–438. https://doi.org/10.1037/h0036429

Nichols, S. (2002). Norms with feeling: towards a psychological account of moral judgment. *Cognition*, *84*(2), 221–236. https://doi.org/10.1016/S0010-0277(02)00048-3

Nichols, S., & Knobe, J. (2007). Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions. *Noûs*, *41*(4), 663–685.

Nichols, S., & Mallon, R. (2006). Moral dilemmas and moral rules. *Cognition*, *100*(3), 530–542. https://doi.org/10.1016/j.cognition.2005.07.005

Niedenthal, P. M., Brauer, M., Halberstadt, J. B., & Innes-Ker, Å. H. (2001). When did her smile drop? Facial mimicry and the influences of emotional state on the detection of change in emotional expression. *Cognition & Emotion*, *15*(6), 853–864. https://doi.org/10.1080/02699930143000194

Nussbaum, M., & Kahan, D. (1996). Two Conceptions of Emotion in Criminal Law. *Columbia Law Review*, 269.

Oden, G. C. (1977). Fuzziness in semantic memory: Choosing exemplars of subjective categories. *Memory & Cognition*, *5*(2), 198–204. https://doi.org/10.3758/BF03197362

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Panksepp, J. (2007). Affective Neuroscience and the Ancestral Sources of Human Feelings. In H. Cohen & B. Stemmer (Eds.), *Consciousness and Cognition:*

Fragments of Mind and Brain (pp. 173–188). Elxevier Academic Press.

Panksepp, J., & Panksepp, J. B. (2000). The seven sins of evoutionary psychology. *Evolution and Cognition*, *6*(2), 108–131.

Park, H. S., Levine, T. R., Kingsley Westerman, C. Y., Orfgen, T., & Foregger, S. (2007). The Effects of Argument Quality and Involvement Type on Attitude Formation and Attitude Change: A Test of Dual-Process and Social Judgment Predictions. *Human Communication Research*, *33*(1), 81–102. https://doi.org/10.1111/j.1468-2958.2007.00290.x

Petrinovich, L., & O'Neill, P. (1996). Influence of wording and framing effects on moral intuitions. *Ethology and Sociobiology*, *17*(3), 145–171. https://doi.org/10.1016/0162-3095(96)00041-6

Petty, R. E., Cacioppo, J. T., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, *48*(3), 306–307.

Petty, R. E., Feinstein, J. A., Blair, W., & Jarvis, G. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psych. Bull*, 197–253.

Pizarro, D. A., & Bloom, P. (2003). The intelligence of the moral intuitions: A comment on Haidt (2001). *Psychological Review*, *110*(1), 193–196. https://doi.org/10.1037/0033-295X.110.1.193

Prinz, J. J. (2005). Passionate Thoughts: The Emotional Embodiment of Moral Concepts. In D. Pecher & R. A. Zwaan (Eds.), *Grounding Cognition: The Role of Perception and Action in Memory, Language, and Thinking* (pp. 93–114). Cambridge University Press.

Prinz, J. J. (2008a). Is Morality Innate? In *Moral Psychology Volume 1: The evolution of morality adaptations and innateness*. Cambridge, Mass.; London, England: The MIT press.

Prinz, J. J. (2008b). Resisting the Linguistic Analogy: A Commentary on Hauser, Young, and Cushman. In W. Sinnott-Armstrong (Ed.), *Moral psychology*

Volume 2, The cognitive science of morality: intuition and diversity (pp. 157–170). London: MIT.

Proulx, T., & Inzlicht, M. (2012). The Five “A”s of Meaning Maintenance: Finding Meaning in the Theories of Sense-Making. *Psychological Inquiry*, *23*(4), 317–335. https://doi.org/10.1080/1047840X.2012.702372

R Core Team. (2017a). *foreign: Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, Weka, dBase, ...* Retrieved from https://CRAN.R-project.org/package=foreign

R Core Team. (2017b). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rand, D. G. (2012). The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology*, *299*(Supplement C), 172–179. https://doi.org/10.1016/j.jtbi.2011.03.004

Rawls, J. (1971). *A theory of justice*. Harvard university press.

Reber, A. S. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General*, *118*(3), 219–235. https://doi.org/10.1037/0096-3445.118.3.219

Reed, E. S. (1996). *Encountering the World: Toward an Ecological Psychology* (First Edition edition). New York: Oxford University Press.

Rest, J. R. (1979a). *Defining issues test*. University of Minnesota, Minnesota moral research project.

Rest, J. R. (1979b). *Revised manual for the Defining Issues Test: An objective test for moral judgment development*. Minnesota Moral Research Projects.

Rest, J. R. (1986). *DIT: Manual for the defining issues test*. Center for the Study of Ethical Development, University of Minnesota.

Roedder, E., & Harman, G. (2010). Linguistics and Moral Theory. In J. M. Doris (Ed.), *The Moral Psychology Handbook* (pp. 111–146). Oxford; New York:

Oxford University Press.

Roffee, J. A. (2014). No Consensus on Incest? Criminalisation and Compatibility with the European Convention on Human Rights. *Human Rights Law Review*, *14*(3), 541–572.

Rosch, E. H. (1973a). Natural categories. *Cognitive Psychology*, *4*(3), 328–350. https://doi.org/10.1016/0010-0285(73)90017-0

Rosch, E. H. (1973b). On the internal structure of perceptual and semantic categories. In *Cognitive development and the acquisition of language* (pp. xii, 308). Oxford, England: Academic Press.

Rosch, E. H., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605.

Royzman, E. B., Atanasov, P., Parks, A., & Gepty, A. (2014). CAD or MAD? Anger (not disgust) as the predominant response to pathogen-free violations of the divinity code. *Emotion*, *14*(5), 892–907. https://doi.org/10.1037/a0036829

Royzman, E. B., Kim, K., & Leeman, R. F. (2015). The curious tale of Julie and Mark: Unraveling the moral dumbfounding effect. *Judgment and Decision Making*, *10*(4), 296–313.

Royzman, E. B., Landy, J. F., & Goodwin, G. P. (2014). Are good reasoners more incest-friendly? Trait cognitive reflection predicts selective moralization in a sample of American adults. *Judgment & Decision Making*, *9*(3), 176–190.

Rozin, P., Haidt, J., & MacCauley, C. (2009). *Disgust: The body and soul emotion in the 21st century*. na.

Rozin, P., Lowery, L., Imada, S., & Haidt, J. (1999). The CAD triad hypothesis: A mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). *Journal of Personality and Social Psychology*, *76*(4), 574–586. https://doi.org/10.1037/0022-3514.76.4.574

Rozin, P., Markwith, M., & McCauley, C. (1994). Sensitivity to indirect contacts with other persons: AIDS aversion as a composite of aversion to strangers, infection, moral taint, and misfortune. *Journal of Abnormal Psychology*, *103*(3), 495–504. https://doi.org/10.1037/0021-843X.103.3.495

Russell, P. S., & Giner-Sorolla, R. (2011a). Moral Anger Is More Flexible Than Moral Disgust. *Social Psychological and Personality Science*, *2*(4), 360–364. https://doi.org/10.1177/1948550610391678

Russell, P. S., & Giner-Sorolla, R. (2011b). Social Justifications for Moral Emotions: When Reasons for Disgust Are Less Elaborated Than for Anger. *Emotion*, *11*(3), 637–646. https://doi.org/10.1037/a0022600

Russell, P. S., & Giner-Sorolla, R. (2013). Bodily Moral Disgust: What It Is, How It Is Different From Anger, and Why It Is an Unreasoned Emotion. *Psychological Bulletin*, *139*(2), 328–351. https://doi.org/10.1037/a0029319

Sabini, J. (1995). *Social psychology*. New York; London: Norton.

Sabo, J. S., & Giner-Sorolla, R. (2017). Imagining Wrong: Fictitious Contexts Mitigate Condemnation of Harm More Than Impurity. *Journal of Experimental Psychology: General*, *146*(1), 134–153. https://doi.org/10.1037/xge0000251

Saltzstein, H. D., & Kasachkoff, T. (2004). Haidt's Moral Intuitionist Theory: A Psychological and Philosophical Critique. *Review of General Psychology*, *8*(4), 273–282. https://doi.org/10.1037/1089-2680.8.4.273

Sauer, H. (2017). *Moral Judgments as Educated Intuitions*. MIT Press.

Schellenberg, E. G., & Trehub, S. E. (2003). Good Pitch Memory Is Widespread. *Psychological Science*, *14*(3), 262–266. https://doi.org/10.1111/1467-9280.03432

Schmidt, D. (2016). The Effects of Cognitive Load and Stereotyped Groups on Punitiveness. *CMC Senior Theses*.

Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H. (2008). Disgust as Embodied Moral Judgment. *Personality & Social Psychology Bulletin*, *34*(8), 1096–

1109. https://doi.org/10.1177/0146167208317771

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*, *84*(1), 1–66. https://doi.org/10.1037/0033-295X.84.1.1

Schnell, T. (2011). Individual differences in meaning-making: Considering the variety of sources of meaning, their density and diversity. *Personality and Individual Differences*, *51*(5), 667–673. https://doi.org/10.1016/j.paid.2011.06.006

Schwitzgebel, E., & Cushman, F. A. (2012). Expertise in Moral Reasoning? Order Effects on Moral Judgment in Professional Philosophers and Non-Philosophers. *Mind & Language*, *27*(2), 135–153. https://doi.org/10.1111/j.1468-0017.2012.01438.x

Shweder, R. A., Much, N. C., Mahapatra, M., & Park, L. (1997). The 'Big Three' of morality (autonomy, community, divinity) and the 'Big Three' explanations of suffering. In A. M. Brandt & P. Rozin (Eds.), *Morality and health* (pp. 119–169). Routledge.

Singmann, H., Bolker, B., & Westfall, J. (2015). *afex: Analysis of Factorial Experiments*. Retrieved from https://CRAN.R-project.org/package=afex

Sinnott-Armstrong, W. (2008a). Framing moral intuitions. In W. Sinnott-Armstrong (Ed.), *Moral psychology Volume 2, The cognitive science of morality: intuition and diversity* (pp. 47–76). London: MIT.

Sinnott-Armstrong, W. (2008b). *Moral Psychology Volume 1: The evolution of morality adaptations and innateness*. Cambridge, Mass.; London, England: The MIT press.

Sinnott-Armstrong, W. (2008c). *Moral psychology Volume 2, The cognitive science of morality: intuition and diversity*. London: MIT.

Sinnott-Armstrong, W. (2008d). *Moral Psychology Volume 3: The neurosciences of morality: emotion, brain disorders, and development*. Cambridge (Mass.): the

MIT press.

Sinnott-Armstrong, W., Young, L., & Cushman, F. A. (2010). Moral Intuitions. In J. M. Doris (Ed.), *The Moral Psychology Handbook* (pp. 206–245). Oxford; New York: Oxford University Press.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3–22. https://doi.org/10.1037/0033-2909.119.1.3

Smith, E. R., & DeCoster, J. (2000). Dual-Process Models in Social and Cognitive Psychology: Conceptual Integration and Links to Underlying Memory Systems. *Personality & Social Psychology Review (Lawrence Erlbaum Associates)*, *4*(2), 108–131.

Sneddon, A. (2007). A Social Model of Moral Dumbfounding: Implications for Studying Moral Reasoning and Moral Judgment. *Philosophical Psychology*, *20*(6), 731–748. https://doi.org/10.1080/09515080701694110

Srull, T. K., & Wyer, R. S. (1979). The role of category accessibility in the interpretation of information about persons: Some determinants and implications. *Journal of Perconality and Social Psycholo Gy*, 1660–1672.

Stanley, M. L., Dougherty, A. M., Yang, B. W., Henne, P., & De Brigard, F. (2017). Reasons Probably Won't Change Your Mind: The Role of Reasons in Revising Moral Decisions. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0000368

Staub, E. (2013). *Positive Social Behavior and Morality: Social and Personal Influences*. Elsevier.

Steger, M. F., Kashdan, T. B., Sullivan, B. A., & Lorentz, D. (2008). Understanding the Search for Meaning in Life: Personality, Cognitive Style, and the Dynamic Between Seeking and Experiencing Meaning. *Journal of Personality*, *76*(2), 199–228. https://doi.org/10.1111/j.1467-6494.2007.00484.x

Stich, S. (1993). Moral philosophy and mental representation. *The Origin of Values*, 215–228.

Strahan, R., & Gerbasi, K. C. (1972). Short, homogeneous versions of the Marlow-Crowne Social Desirability Scale. *Journal of Clinical Psychology*, *28*(2), 191–193. https://doi.org/10.1002/1097-4679(197204)28:2<191::AID-JCLP2270280220>3.0.CO;2-G

Sun, R., Slusarz, P., & Terry, C. (2005). The Interaction of the Explicit and the Implicit in Skill Learning: A Dual-Process Approach. *Psychological Review*, *112*(1), 159–192. https://doi.org/10.1037/0033-295X.112.1.159

Tabossi, P. (1988). Effects of context on the immediate interpretation of unambiguous nouns. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 153–162. https://doi.org/10.1037/0278-7393.14.1.153

Tabossi, P., & Johnson-Laird, P. N. (1980). Linguistic Context and the Priming of Semantic Information. *Quarterly Journal of Experimental Psychology*, *32*(4), 595–603. https://doi.org/10.1080/14640748008401848

Takeuchi, A. H., & Hulse, S. H. (1993). Absolute pitch. *Psychological Bulletin*, *113*(2), 345–361. https://doi.org/10.1037/0033-2909.113.2.345

Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist*, 204–217.

Thomson, J. J. (1986). *Rights, Restitution, and Risk: Essays in Moral Theory*. Harvard University Press.

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making*, *11*(1), 99–113.

Tomasello, M. (2003). *Constructing a Language*. Harvard University Press.

Tomasello, M. (2014). *A Natural History of Human Thinking*. Harvard University Press.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, *39*(7), 1275. https://doi.org/10.3758/s13421-011-0104-1

Topolski, R., Weaver, J. N., Martin, Z., & McCoy, J. (2013). Choosing between the emotional dog and the rational pal: A moral dilemma with a tail. *Anthrozoös*, *26*(2), 253–263. https://doi.org/10.2752/175303713X13636846944321

Triskiel, J. (2016). Psychology Instead of Ethics? Why Psychological Research Is Important but Cannot Replace Ethics. In C. Brand (Ed.), *Dual-Process Theories in Moral Psychology: Interdisciplinary Approaches to Theoretical, Empirical and Practical Considerations* (pp. 77–98). Springer.

Trope, Y., & Liberman, N. (2003). Temporal construal. *Psychological Review*, *110*(3), 403–421. https://doi.org/10.1037/0033-295X.110.3.403

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293–315. https://doi.org/10.1037/0033-295X.90.4.293

Unipark, Q. (2013). *QuestBack Unipark.(2013).*

Valdesolo, P., & DeSteno, D. (2006). Manipulations of Emotional Context Shape Moral Judgment. *Psychological Science*, *17*(6), 476–477. https://doi.org/10.1111/j.1467-9280.2006.01731.x

van den Bos, K. (2018). On the Possibility of Intuitive and Deliberative Processes Working in Parallel in Moral Judgement. In K. J. Gray & J. Graham (Eds.), *Atlas of Moral Psychology* (pp. 31–39). New York, NY: The Guilford Press.

Wallbott, H. G. (1991). Recognition of emotion from facial expression via imitation? Some indirect evidence for an old theory. *British Journal of Social Psychology*, *30*(3), 207–219. https://doi.org/10.1111/j.2044-8309.1991.tb00939.x

Wheatley, T., & Haidt, J. (2005). Hypnotic Disgust Makes Moral Judgments More Severe. *Psychological Science*, *16*(10), 780–784. https://doi.org/10.1111/j.1467-9280.2005.01614.x

Wickham, H. (2007). Reshaping Data with the reshape Package. *Journal of Statistical Software*, *21*(12), 1–20.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. Retrieved from http://ggplot2.org

Wickham, H. (2011). The Split-Apply-Combine Strategy for Data Analysis. *Journal of Statistical Software*, *40*(1), 1–29.

Wickham, H. (2016). *scales: Scale Functions for Visualization*. Retrieved from https://CRAN.R-project.org/package=scales

Wickham, H., & Chang, W. (2017). *devtools: Tools to Make Developing R Packages Easier*. Retrieved from https://CRAN.R-project.org/package=devtools

Wickham, H., Francois, R., Henry, L., & Müller, K. (2017). *dplyr: A Grammar of Data Manipulation*. Retrieved from https://CRAN.R-project.org/package=dplyr

Wiegmann, A., Okan, Y., & Nagel, J. (2012). Order effects in moral judgment. *Philosophical Psychology*, *25*(6), 813–836. https://doi.org/10.1080/09515089.2011.631995

Wielenberg, E. J. (2014). *Robust Ethics: The Metaphysics and Epistemology of Godless Normative Realism*. OUP Oxford.

Winston Chang. (2014). *extrafont: Tools for using fonts*. Retrieved from https://CRAN.R-project.org/package=extrafont

Wong, Y. K., & Wong, A. C.-N. (2014). Absolute pitch memory: Its prevalence among musicians and dependence on the testing context. *Psychonomic Bulletin & Review*, *21*(2), 534–542.

Yankouskaya, A., Humphreys, G. W., & Rotshtein, P. (2014). Differential interactions between identity and emotional expression in own and other-race faces: Effects of familiarity revealed through redundancy gains. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(4), 1025–1038. https://doi.org/10.1037/a0036259

Yeh, W., & Barsalou, L. W. (2006). The Situated Nature of Concepts. *The American Journal of Psychology*, *119*(3), 349. https://doi.org/10.2307/20445349

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist*, *35*(2), 151–175. https://doi.org/10.1037/0003-066X.35.2.151

Zaykin, D. V. (2011). Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. *Journal of Evolutionary Biology*, *24*(8), 1836–1841. https://doi.org/10.1111/j.1420-9101.2011.02297.x

Zhong, C.-B., & Liljenquist, K. (2006). Washing Away Your Sins: Threatened Morality and Physical Cleansing. *Science*, *313*(5792), 1451–1452. https://doi.org/10.1126/science.1130726

Zhong, C.-B., Strejcek, B., & Sivanathan, N. (2010). A clean self can render harsh moral judgment. *Journal of Experimental Social Psychology*, *46*(5), 859–862.

List of Tables

List of Figures

List of Abbreviations

BORIS	Behavioural Observation Research Interactive Software
CRSi7	Centrality of Religiosity Scale
CRT	Cognitive Reflection Test
DDE	Doctrine of Double Effect
DIT	Defining Issues Test
LA	Linguistic Analogy
MIC	Mary Immaculate College
MLQ	Meaning in Life Questionnaire
MSP	Member of Scottish Parliament
NCS	Need for Cognition Scale
NFC	Need for Closure
SES	Socio-economic status
SIM	Social Intuitionist Model
SPSS	Statistical Package for the Social Sciences
UL	University of Limerick
UMG	Universal Moral Grammar

Appendices

Appendix A: moral scenarios

Heinz

In Europe, a woman was near death from a very bad disease, a special kind of cancer. There was one drug that the doctors thought might save her. It was a form of radium for which a druggist was charging ten times what the drug cost him to make. The sick woman's husband, Heinz, went to everyone he knew to borrow the money, but he could only get together about half of what it cost. He told the druggist that his wife was dying, and asked him to sell it cheaper or let him pay later. But the druggist said, "No, I discovered the drug and I'm going to make money from it." So, Heinz got desperate and broke into the man's store to steal the drug for his wife. The druggist had Heinz arrested and charged (Haidt et al., 2000).

Trolley

A Trolley is hurtling down a track towards five people. It will kill them all on impact. Paul is on a bridge under which it will pass. He can stop it by putting something very heavy in front of it. As it happens, there is a very fat man next to him – Paul's only way to stop the trolley is to push him over the bridge and onto the track, killing him to save five. Paul decides to push the man (adapted from Greene et al., 2001).

Cannibal (original)

Jennifer works in a medical school pathology lab as a research assistant. The lab prepares human cadavers that are used to teach medical students about anatomy. The cadavers come from people who had donated their body to science for research. One night Jennifer is leaving the lab when she sees a body that is going to be discarded the next day. Jennifer was a vegetarian, for moral reasons. She thought it was wrong to kill animals for food. But then, when she saw a body about to be cremated, she thought it was irrational to waste perfectly edible meat. So she cut off a piece of flesh, and took it home and cooked it. The person had died recently of a heart attack, and she cooked the meat thoroughly, so there was no risk of disease (Haidt et al., 2000).

Cannibal (revised)

Jennifer works in a medical school pathology lab as a research assistant. The lab prepares human cadavers that are used to teach medical students about anatomy. The cadavers come from people who had donated their body for the general use of the researchers in the lab. The bodies are normally cremated, however, severed cuts may be disposed of at the discretion of lab researchers, One night Jennifer is leaving the lab when she sees a body that is going to be discarded the next day. Jennifer was a vegetarian, for moral reasons. She thought it was wrong to kill animals for food. But then, when she saw a body about to be cremated, she thought it was irrational to waste perfectly edible meat. So she cut off a piece of flesh, and took it home and cooked it. The person had died recently of a heart attack, and she cooked the meat thoroughly, so there was no risk of disease

Incest

Julie and Mark, who are brother and sister, are travelling together in France. They are both on summer vacation from college. One night they are staying alone in a cabin near the beach. They decide that it would be interesting and fun if they tried making love. At very least it would be a new experience for each of them. Julie was already taking birth control pills, but Mark uses a condom too, just to be safe. They both enjoy it, but they decide not to do it again. They keep that night as a special secret between them, which makes them feel even closer to each other (Haidt et al., 2000).

Appendix B: Sample statements to challenge judgements

Trolley

- Do you accept that five people would have died if Paul didn't push the man?
- And this man is the only way available to stop the trolley? (Paul does not weigh enough)
- Do you agree that in stopping the trolley Paul saved the lives of five people?

Cannibal

- The body had been donated for research, it was to be discarded the next day. You must agree then that it had obviously fulfilled its purpose?
- Do you accept that the body was already dead?
- And do you accept that there was no risk of disease?

Heinz/Druggist

- Do you agree that the druggist has to make a living?
- And do you accept that Heinz broke into the druggist's store?
- And do you accept that he stole from him?

  -
- Do think that Heinz should try to save his wife's life?
- And do you agree that he tried to get the money together
- And do you accept that Heinz tried to negotiate with the druggist

Incest

- Do you not agree that any concerns regarding reproductive complications are eased by their using of two forms of contraception?
- And do you accept that they are both consenting adults, and that they both consented and enjoyed it?
- And do you concede that nobody else was affected by their actions?

Appendix C: post discussion questionnaire

How sure were you about your judgement						
1	2	3	4	5	6	7
Not at all						Extremely sure

How much did you change your mind?						
1	2	3	4	5	6	7
Not at all						Extremely

How confused were you?						
1	2	3	4	5	6	7
Not at all						Extremely confused

How irritated were you?						
1	2	3	4	5	6	7
Not at all						Extremely irritated

How much was your judgement based on reason?						
1	2	3	4	5	6	7
Not at all						Extremely

How much was your judgement based on “gut” feeling?						
1	2	3	4	5	6	7
Not at all						Extremely