DOCTORAL THESIS

Morality as Navigation

Building a Moral Map and Compass from Constructive Sentimentalism

Raymond-Barker , Brett Alexander

Award date:
2022

Awarding institution:
University of Roehampton

Morality as Navigation: Building a Moral Map and Compass from Constructive Sentimentalism

by

Brett Alexander Raymond-Barker

A thesis submitted in partial fulfilment of the requirements for the degree of PhD

Department of Psychology

University of Roehampton

2022

Abstract

Moral psychologists tend to regard Humean philosophy favourably, although appear to have overlooked the updated version of Hume's treatise proffered by Prinz (2004a, 2004b, 2009). Three studies, targeted towards areas of contention between Moral Foundations Theory (Graham et al., 2013) and the Theory of Dyadic Morality (Schein & Gray, 2018), show emotions are elicited by (im)moral events, can contribute to moralization, and may act to amplify or suppress judgements of severity - findings which appear supportive of Prinz's claim that morals are constructed from emotions. Study 1 provides a conceptual replication of Gray and Keeney's (2015) research, with results challenging their claim that violations of purity are just a weird type of harm. Associations found in Study 1, between harm-anger and impurity-disgust, were also apparent in Study 2 - which provides an open-ended test of, and finds support for, the emotion-content relationships hypothesized under Constructive Sentimentalism. Study 3 provides an extended conceptual replication of Seidel and Prinz (2013a), using a 'content-free' emotion induction paradigm in combination with the investigatory framework outlined by Cameron, Lindquist and Gray (2015), finding an influence of interoceptive awareness on moral judgements. Arguments are advanced to establish purity and harm as being at least equally important, and to contend that the vast majority of moral violations contain mixed moral content - explaining the frequent co-occurrence of anger and disgust in response to moral transgressions. Following Constructive Sentimentalism (Prinz, 2009), moral judgements are postulated to require two points of reference, whereby 'Autonomy', 'Harm', and 'Other' may be aligned to one axis, and 'Continuity', 'Purity', and 'Self' aligned to another. This approach is shown to accommodate different theories of morality into a common theoretical framework and provide a means of orientating research findings and themes within moral psychology via reference to the tools, methods and practices of navigation.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank Dr. Gina Pauli, Dr. Amanda Holmes, and Professor Marcia Worrell for their support, help, assistance, kindness, and encouragement over the course of my education and the completion of this thesis - I am very grateful for all you have done.

I would also like to thank Professor Roger Giner-Sorolla and Dr. John Rae for their time, input, and enthusiasm during the examination of the thesis. The final text has undoubtedly been improved following our discussions and your feedback.

I would further like to thank everyone who has contributed, in any of a variety of ways, over the course of my studies. This includes Professor Mick Cooper, Dr. Jon Silas, Dr. Margot Crossman, Dr. Edith Steffen, Dr. Eva Eppler, Dr Lewis Goodings, and the many students I have had the privilege to teach - with additional thanks due to Dr John Rae, Professor Marcia Worrell, and my supervisors Dr. Gina Pauli and Dr. Amanda Holmes.

Lastly, I would like to thank those who have indirectly contributed to this thesis. This includes all authors whose work is cited within the text - with particular thanks due to Professor Jesse Prinz, Professor Jonathan Haidt, and Professor Kurt Gray - as well as thanks to all the participants who took part in the studies conducted for this thesis.

I dedicate this thesis, with love, to my mother, Eve, and to Luisa, Isabella and Emrys.

"Whenever you meet someone, ask yourself first this immediate question:

'What beliefs does this person hold about the good and bad in life?'

Because if he believes this or that about pleasure and pain and their constituents, about fame and obscurity, death and life, then I shall not find it surprising or strange if he acts in this way or that way, and I shall remember he has no choice but to act as he does."

Marcus Aurelius - Meditations 8:14

Chapter 1 - Introduction

The following thesis revolves around the theory of Constructive Sentimentalism (Prinz, 2009), which offers an empirically informed account of morality drawing parallels with Hume's 'Enquiry Concerning the Principles of Morals' (1751). The core claim of the sentimentalist approach advocated for by these philosophers is that morals have an emotional foundation - they are built on 'the passions', to which reason is subservient. Constructive Sentimentalism is of particular interest as moral psychology has trended away from rationalist and/or prescriptive approaches, such as Kohlberg (1994) who classified developmental stages in relation to Kantian ethics, in favour of more descriptive approaches which emphasize emotions or intuitions in accounting for moral phenomena. However, although many of the authors advancing these latter approaches regard Hume favourably, the few that cite Prinz (2009) do so only in support of claims that emotions have a role in morality - they do not seem to give Constructive Sentimentalism as much attention as it would appear to warrant.

Prinz provides a comprehensively updated Humean account of concepts (2004a), emotions (2004b), and morals (2009), which seems better developed than

existing theories on offer within moral psychology. Furthermore, this material may, with development, provide a means of connecting seemingly distinct areas of psychology through its emphasis on emotion. If emotions are a form of perception, as Prinz (2004b) contends, this may provide an avenue to connect morality with ecological approaches to perception (and action) via emotion. Similarly, if emotion is a constituent of moral judgement, as Prinz (2009) contends, this may provide an avenue for establishing links with various therapeutic approaches – particularly those with a focus on emotions (e.g., van Deurzen, 2014, see also Panksepp & Biven, 2012). Given that Prinz's approach, in addition to providing explanation for various moral phenomena, may be able to connect psychology from 'bottom' (e.g., perceptual approaches to emotion) to 'top' (e.g., existential approaches to therapy), it would seem imprudent not to examine Constructive Sentimentalism further.

Of particular interest is that Prinz (2009) follows the recommendation of attending to the 'imperceptible change in the copulations of propositions' which Hume (1739) suggests would 'subvert all the vulgar systems of morality'. The passage in the Treatise (1739, p.335), concerning the deduction of 'ought' from 'is', has been the subject of various interpretations. In the simplest case, it states that for something to appear in the conclusion of an argument it must have appeared in the premises of that argument. This is not controversial. Indeed, the very issue is this 'altogether inconceivable' change whereby a prescriptive conclusion (concerning 'ought' and 'ought not') seems to follow from descriptive premises (concerning what 'is', and 'is not') – when the former is an 'entirely different relation' to the latter. However, some have interpreted the statement as claiming that such a deduction is impossible - there is an impermeable divide between matters of fact (i.e., science) and matters of value (i.e., morality). For example, there are no facts about the act of murder that entail one ought to value refraining from it. In contrast, others have interpreted it as merely noting the common absence of certain premises in moral systems - they provide 'answers' without showing how they work out all the steps necessary to arrive at those answers.

Details of this controversy aside (for discussion, see Greene, 2003; Harris, 2012), Prinz argues that 'oughts' are entailed to wrongs, such that 'if something is wrong, then one ought not to do it' – and that the concept ‘WRONG’ is constituted by an emotional disposition. From the third-person perspective this allows for the discovery of others' moral obligations - and your own if applying the third-person perspective to yourself (cf. therapy); but also, when the argument is stated from a first-person perspective - an ought can be derived from an is. According to your own standards - you ought to do what it would be wrong not to do. If Prinz is correct, and Constructive Sentimentalism provides a viable account of morality, then this would represent quite a development for moral psychology as it provides an empirically tractable definition of what counts for 'morally wrong' at the individual level, and links this to concepts of ought and obligation which are seldom addressed directly in descriptive accounts of morality.

The main thrust of the thesis is thus whether Constructive Sentimentalism (Prinz, 2009) provides a better account of morality than alternative theories on offer from moral psychologists. Constructive Sentimentalism shares ground with both Moral Foundations Theory (Graham et al., 2013) - long considered the dominant theory in moral psychology, and its most persistent critic - the Theory of Dyadic Morality (Schein & Gray, 2018). Constructive Sentimentalism, like Moral Foundations Theory, argues something may be considered morally wrong without necessary recourse to 'harm' - defined as an intentional agent causing damage to a vulnerable patient (Schein & Gray, 2018). It advocates for 'moral pluralism' (Graham et al., 2013) - there is *more than one* way in which something can be morally wrong, rather than 'harm pluralism' (Schein & Gray, 2018) - there are a wide variety of ways in which a range of agents might cause damage to all sorts of vulnerable patients. However, like the Theory of Dyadic Morality, Constructive Sentimentalism favours constructionist accounts of psychology over the (massively) modular approaches drawn on by Moral Foundations Theory. It argues that morals are constructed with reference to emotion, rather than with reference to several

'mental systems' - each attuned to a specific type of moral content. Constructive Sentimentalism offers a position between the Theory of Dyadic Morality and Moral Foundations Theory by providing a constructionist account which advocates moral pluralism, and this may be able to better explain the research findings taken to support each of these approaches. However, in claiming moral judgements are constituted by emotional dispositions, Constructive Sentimentalism further offers an account of the genesis of moral intuitions, which neither of the other theories provide - despite arguing that such affect-laden intuitions are responsible for moral judgements.

To allow for a three-way comparison, all research questions are targeted towards an area of common ground on which each theory stakes differing claims - the role of emotion(s) in morality. A preliminary question addresses an area of contention related to these claims - whether there are different types (foundations; domains) of moral content (Graham et al., 2013; Prinz, 2009) such as 'fairness', 'loyalty', 'authority', and 'sanctity', or whether different types of moral content are just taxonomic categories which label varieties of perceived harm (Schein & Gray, 2018). The focus here is on the extent to which violations of 'sanctity' may be construed as simply 'weird harms'. Subsequent questions focus on three hypothesized relationships between emotions and moral judgements. Firstly, that moral judgements *elicit* emotions. Secondly, that emotions can *amplify* moral judgements. Thirdly, that emotions can act to *moralize* non-moral content.

The key theme common to all studies conducted as part of this thesis is the question of whether specific emotions have some form of (exclusive) correspondence relationship with different types of moral content. This hypothesis, advanced by both Moral Foundations Theory and Constructive Sentimentalism, but rejected by the Theory of Dyadic Morality, claims that anger is (characteristically) associated with the violation of norms about persons (or foundations of 'harm' and 'cheating'), and that disgust is associated with violations of norms about the (perceived) natural order of nature (the

'degradation' foundation). This provides an additional layer to the other hypotheses, such that emotions of anger and disgust are hypothesized to operate with *specific* and/or *exclusive* regard to their associated content type. For example, that disgust amplifies judgements of degrading acts and perceptions of degradation, but does not amplify perceptions of harm or judgements of harmful acts, whereas anger amplifies perceptions of harm and judgements of harmful acts, but not judgements of degrading acts or perceptions of degradation.

Each of the research questions, in addition to addressing hypothesized roles of emotion in moral judgement, serves as a means of evaluating the validity of Constructive Sentimentalism's approach to morality and illustrating its explanatory power. The choice of focusing on an existing research theme which has already received considerable attention - that of emotion specificity in moral judgements - rather than exploring the many ways in which Prinz's approach might be developed, follows from concerns regarding the extent to which the results of psychological research can be reproduced (Pashler & Wagenmakers, 2012; Nosek, Spies & Motyl, 2012; Earp & Trafimow, 2015). Indeed, that one area cited as contributing to problems with replication relates to issues involving theory specification (e.g., Klein, 2014; although see Trafimow & Earp, 2016) provides additional justification for drawing on Prinz's empirically informed approach to moral philosophy. Constructive Sentimentalism is better specified and more rigorously argued for than either of the other two theories, and thus more empirically tractable. In addition to providing a broader account of the philosophical underpinnings of morality, it identifies the necessary and sufficient conditions by which a judgment may qualify as moral, expands upon this in considering the nature of moral obligations, and details how morality may be constructed with reference to independently motivated (and non-moral) accounts of concepts and emotions (Prinz, 2004a,b). In contrast, Moral Foundations Theory has taken a pragmatic (i.e., non-systematic, ad-hoc) approach to aspects of theory construction (e.g., foundation criteria), and the Theory of Dyadic Morality endorses a majority of these same claims - with the notable objection of moral pluralism.

The approach taken in addressing the research questions also follows from concerns about replication. Each of the studies conducted can be considered as providing an extended conceptual replication of existing research that relates to a key premise of the theories under discussion. Research designs and materials are tailored towards paradigms and stimuli drawn from the respective research under discussion in each case, with preference being given to 'author approved' versions where practical. This was done to help ensure both that each theoretical approach receives a fair trial, and to minimize the degrees of freedom available for authors in defending their theories. Related to this, the basic aims, rationale, hypotheses, design, materials, and proposed analyses for each of the three studies were pre-registered (see van't Veer & Giner-Sorolla, 2016) on the Open Science Framework (osf.io). This was done with a view to allowing work to be checked, facilitating any subsequent replication attempts, and more generally – to provide transparency in line with the principles of (open) science.

The justification for this approach is already implied, but worth restating in full. Prinz provides comprehensive and detailed groundwork for a theory of morality – Constructive Sentimentalism (Prinz, 2009) – which has received comparatively little attention in the literature generated by moral psychologists. The theory shares common ground with both the most cited theory of moral psychology – Moral Foundations Theory (Graham et al., 2013) – and its most fervent challenger – the Theory of Dyadic Morality (Schein & Gray, 2018); however, it also covers more ground within moral psychology, connects more easily with other fields (e.g., therapy), and touches on lesser-charted ground in the field – 'oughts'. Thus, Constructive Sentimentalism appears to be a more powerful theory of moral psychology than those previously on offer from within the field - although this has yet to be illustrated due to the comparative lack of attention it has received.

The fit with existing theories suggests that Constructive Sentimentalism may be able to subsume substantial parts of these under its approach, and its explanatory power can be illustrated through examining key hypotheses regarding the roles of emotion in moral judgement - hypotheses which are of some importance to Prinz's argument. These hypotheses also happen to be matters of particular contention within moral psychology; recent reviews of related research have challenged the evidence base for specific/exclusive links between emotions and moral content (Cameron, Lindquist & Gray, 2015), and show that the effects of disgust induction on moral judgement appear smaller than might be expected given the hypothesized moral relevance of this emotion (Landy & Goodwin, 2015a). As such, for all the potential merit of Constructive Sentimentalism, it would seem unwise to explore this potential before examining the merit of its basic claims - especially considering both recent reviews and more general concerns regarding the ability to replicate psychological research findings. This thesis proffers extended conceptual replications of studies considered important to the theories under discussion, and illustrates how Constructive Sentimentalism accounts for evidence in comparison with other theories, so as to provide some account of the viability of different theories within moral psychology and elucidate areas of common ground between them.

1.1. Thesis Structure

The thesis begins by examining ground common to, and cited favourably by, all three theories - the work of Shweder, Much, Mahapatra and Park (1997) - which provides definition of three different moral codes, and structures morality as applied to health and explanations of suffering. This work also lays the groundwork for Rozin, Lowry, Imada, and Haidt's (1999) CAD model, which proposes each of these moral domains (Community-Autonomy-Divinity) is associated with a corresponding emotion (Contempt-Anger-Disgust respectively). This leads to discussion (and critique) of Moral

Foundations Theory (Graham et al., 2013) which, following the forerunning CAD model, also advocates for associations between emotions and moral domains. The Theory of Dyadic Morality (Schein & Gray, 2018), which offers the most persistent challenge to Moral Foundations Theory, is then outlined and critiqued before addressing a key premise of correspondence accounts - that there are different moral domains; as if violations of sanctity are merely weird varieties of dyadic harm (cf. Gray & Keeney, 2015a), then there would be no moral foundation/domain with which disgust could correspond.

Attention then turns to Constructive Sentimentalism (Prinz, 2009), providing an outline of the theory and pitting its formulation against that of the Theory of Dyadic Morality (Schein & Gray, 2018). Substantial areas of overlap between theories suggest further investigation of moral-emotion correspondences may better inform debate, leading to discussion of two reviews relevant to this area. Outlines of both Cameron, Lindquist and Gray's (2015) and Landy and Goodwin's (2015a) findings are provided and critiqued, followed by an overview of research relating to moral-emotion specificity conducted after these reviews. These dovetail to tests of the elicitation, amplification, and moralization hypotheses advanced with regard to the role of emotion in morality. The first of these investigates claims of emotion specificity with regard to the elicitation hypothesis, for both moral and immoral actions, as proposed by Prinz (2009). The second conceptually replicates and extends on the work of Seidel and Prinz (2013a), using the framework provided by Cameron et al. (2015), to investigate claims of emotion specificity with regard to the amplification and moralization hypotheses. The thesis concludes with a view to establishing common ground, illustrated by drawing parallels with navigational equipment and practices, which may reconcile the theories under discussion with respect to the current evidence base. It also explores avenues for future research arising from the product of this reconciliation.

The resultant view from common ground draws analogies between morality and navigation, suggesting that emotions constitute the mechanism by which a moral compass may operate. On this approach, moral foundations can be analogized with landmarks which loom large on the landscape, such that notions of correspondences between emotions and moral content are concerned with the role of the compass (emotions) in orientating the map (concepts) to the ground (perception/action). Within this frame, the derivation of 'oughts' is akin to taking bearings - using the map and compass in order to determine which way to go, or how to go about getting, to your target destination. In this sense, the definition of morality employed is the broader, more traditional one relating to the constituent properties of a 'good life' (cf. Aristotle), rather than more simplified notions of good/bad and right/wrong. Of course, one might ask the question of whether one should actually do what one ought to do - a normative question seemingly beyond the scope of scientific inquiry. However, there may be a good descriptive answer to be given to this question. If you do not do that which, according to your own standards, it would be wrong not to do, then there will be consequences for yourself (yourself can also be read as - your ‘self’ - following Prinz & Nichols, 2016). Indeed, as illustrated by Shweder et al. (1997), individuals tend to account for suffering in moral terms.

1.2. The ‘Big Three’ of Suffering

Shweder, Much, Mahapatra, and Park (1997) examine discourses relating to morality and suffering, and discuss cultural differences which may influence the expression of such narratives. Their curiosity stems from an apparently common human tendency to try to make sense of suffering, and to take meaning from it, with reference to a relatively small set of 'causal ontologies' - ideas about what may be held responsible as the cause(s) of suffering. Shweder et al. (1997) claim that, worldwide, the most used ('big three') explanations for suffering can be categorised as relating to interpersonal,

biomedical, and moral causes - whereby responsibility may be (respectively) externalized, negated, or owned. They also note two points of particular interest. Firstly, that biomedical remedies seem to be sought more often that biomedical explanations are offered. Secondly, that both biomedical and interpersonal explanations for suffering are often imbued with moral implications and/or derivable from moral concerns - such that moral explanations for suffering may be more prevalent than their data suggests. The apparent mismatch between explanations given for suffering, which often have implicit (or explicit) moral relevance, and remedies sought for suffering, which in Western, Educated, Industrialised, Rich and Democratic societies (see Henrich, Heine & Norenzayan, 2010) tends toward forms of biomedical relief, provides a background to the current thesis - encapsulated by the notion that the practice and development of moral virtue, so as to conduct oneself in the 'right' way (i.e., to do what one ought), may be considered a form of preventative medicine.

1.3. The ‘Big Three’ of Morality

Shweder et al.'s (1997) 'big three' of morality - thematic clusters garnered from participants discursive responses to scenarios depicting culturally specific transgressions - provide the base for much subsequent work in moral psychology. One current area of contention is whether these themes just describe different types of moral domain - they are just taxonomic varieties of perceived (dyadic) harm (Schein & Gray, 2018), or whether these themes point to a range of distinguishable moral foundations - mental systems each of which have their own evolutionary roots, and are pre-attuned to detect certain stimuli deemed relevant for the instantiation of these systems following development within a particular culture (Graham et al., 2013). The debate concerns the extent to which these categories of moral discourse are reflective of reality, or in simple operational terms, whether different types of morally relevant events are distinguishable at a psychological level.

Shweder et al.'s (1997) ethics of autonomy relates to concerns about harm, justice, and rights - with a focus on the protection of individuals' freedom. This is typically the dominant ethic for societies which value 'individualism', and the concept of 'harm' in such societies tends to be deeper and more expansive than in 'collectivist' cultures. The self is regarded as an 'individual preference structure', which is sacred in its own right, and must be free to discover and follow its own obligations in addition to (or in spite of) any which may be imposed on it from elsewhere. This ethic is summarised by Constructive Sentimentalism (Prinz, 2009) as relating to norms concerning 'persons', and considered analogous to the moral foundations of 'care/harm' and 'fairness/cheating' (Graham et al., 2013).

The ethics of community relies on notions of hierarchy, duty, interdependence, and 'selves' - with a focus on protecting the integrity of positions constituting a society, and the narrative account of itself - such that for cultures where this is the dominant ethic notions of personal identity may be more closely associated with communal connections than distinctive individuality. The self is regarded as an office holder in a 'feudal' hierarchy, and operates with a view to 'taking care of one's own', with obligations derived from communal participation. This ethic is summarised by Constructive Sentimentalism (Prinz, 2009) as relating to norms concerning 'the natural order of persons', and considered analogous to the moral foundations of 'loyalty/betrayal' and 'authority/subversion' (Graham et al., 2013).

Shweder et al.'s (1997) ethics of divinity draws on concepts such as the natural and/or sacred order, tradition, sin, and sanctity - with a focus on protecting the numinous aspects of nature and humanity from pollution or degradation - and may be the dominant ethic for communities which are particularly spiritually inclined or religiously orientated. The self is regarded as connected to the (sacred) natural order, which encompasses all

things, such that everything forms part of an overarching moral order - the world is sacred, and individuals are obligated to uphold (divinely inspired) 'ways of life' and give due reverence to 'the forms of the world'. This ethic is summarised by Constructive Sentimentalism (Prinz, 2009) as relating to norms concerning 'the natural order', and considered analogous to the moral foundation of 'purity' or 'sanctity/degradation' (Graham et al., 2013).

1.4. The 'CAD Triad' Hypothesis

Moral Foundations Theory (Graham et al., 2013) follows from earlier work showing each of Shweder et al.'s (1997) moral themes appears to be associated with a specific emotional response. Rozin, Lowrey, Imada, and Haidt (1999) hypothesized links between 'Contempt' and violations of 'Community' ethics, 'Anger' and violations of 'Autonomy', and 'Disgust' and violations of 'Divinity'. In response to a variety of moral vignettes, Rozin et al. (1999) found a reasonable degree of support for their 'CAD Triad' hypothesis in samples of both American and Japanese participants. Choices of emotions depicted either by facial expressions, or words, tended to be selected in keeping with the hypothesized response pattern for violations of each moral domain respectively. These findings provide a common reference point for the theories under evaluation, and a means of contrasting their theoretical positions.

The Theory of Dyadic Morality (Schein & Gray, 2018) argues against the 'CAD Triad' hypothesis. Rozin et al.'s (1999) results, and other similar findings, can be accounted for with reference to shared components of emotion (e.g., core affect, conceptual knowledge), concerns regarding methodology (e.g., forced choice responding), and a dyadic definition of 'harm' (Cameron et al., 2015). There are no specific correspondences between emotions and moral domains because there is only one domain of morality - that of perceived dyadic harm. As juxtaposition, the 'CAD Triad'

both informs and is fully endorsed by Constructive Sentimentalism. Constructive Sentimentalism argues that there are two primary moral domains, concerning 'persons' (autonomy) and 'the natural order' (divinity), constructed with reference to emotions of anger and disgust respectively. It also argues for a derived domain, 'the natural order of persons' (community), which is associated with contempt – an emotion which Prinz claims is a blend of anger and disgust. Rozin et al.'s (1999) results are as would be expected if morals are emotionally constructed in the manner described by Prinz, as following Constructive Sentimentalism – domains are constructed with reference to emotion. In contrast, Moral Foundations Theory has built on the 'CAD Triad' in combination with intuitionist models of moral judgement (Haidt, 2001), expanding to four (Haidt & Joseph, 2004), five (Haidt & Joseph, 2007), and potentially six or more different types of morally relevant 'foundations', with (merely) 'characteristic' emotional responses for each (Graham et al., 2013).

Chapter 2 - On Moral Foundations Theory

Moral Foundations Theory (MFT, Graham et al., 2013) has previously been considered as the leading theory of moral psychology, and has been extensively defended by Haidt in *The Righteous Mind* (2013). The primary focus of MFT is on moral intuitions relevant for explaining individual and cultural moral differences - proposing the existence of a multitude of (modular) 'innate mental systems', which have arisen during human evolutionary history, are shaped by cultural experience, and from which morality is derived and constrained. These 'mental systems', which MFT calls 'foundations', are attuned in advance of experience towards detecting information relating to different types of moral content. This approach has garnered considerable attention, and provided a reference point for many investigations in the area, which has resulted in substantial empirical material. The wealth of supportive literature allows the authors to claim the theory has *pragmatic validity* - that it is scientifically useful for inquiry, allowing researchers to answer questions and pose new ones, despite being an incomplete account of moral phenomena. Graham et al. argue that it is the pluralistic nature of MFT's approach – the recognition of different types of moral content - that allows for this kind of validity; and which has demonstrable benefits over monistic approaches which consider morality can be reduced to 'one system' - such as 'justice' (Kohlberg, 1994), 'harm' (Schein & Gray, 2018), or 'well-being' (Harris, 2012). In particular, MFT is concerned with *Mapping the Moral Domain* (Graham et al., 2011; 2013), defined as comprising everything relevant to individual consideration in matters of right and wrong.

2.1. Four Key Claims

In addition to arguing for pluralism, MFT makes three further claims. There is a 'first draft' of the moral mind, such "that the human mind is organized in advance of experience so that it is prepared to learn values, norms, and behaviors related to a diverse set of recurrent adaptive social problems" (Graham et al., 2013, p.63). This universal draft is subsequently edited in variable ways through socio-cultural developmental experiences, such that the individual learns the 'web of shared meanings and evaluations' of their culture in order to "successfully navigate the moral “matrix” [the web] he or she actually experiences" (p.64). Finally, 'intuitions' - rapid, automatic, effortlessly affective evaluations - typically precede (strategic) reasoning, such that moral evaluations are "more a product of the gut than the head, bearing a closer resemblance to aesthetic judgment than principle-based reasoning" (p.66). These intuitions, which vary between cultures, can be grouped into familial categories, each of which is argued to have effectively solved a recurrent social challenge faced during the course of human evolution - and as there were a plurality of such challenges, there are a plurality of moral foundations.

2.2. Moral Judgements, Intuitions, and Reasoning

MFT endorses the following relevant definitions drawn from Haidt (2001). “*Moral judgments* are […] defined as evaluations (good vs. bad) of the actions or character of a person that are made with respect to a set of virtues held to be obligatory by a culture or subculture” (p. 1028); “*moral reasoning* can […] be defined as conscious mental activity that consists of transforming given information about people in order to reach a moral judgment”; and “*moral intuition* can be defined as the sudden appearance in consciousness of a moral judgment, including an affective valence (good–bad, like–dislike), without any conscious awareness of having gone through steps of searching,

weighing evidence, or inferring a conclusion" (p.1029). Accordingly, the key claims of Moral Foundations Theory can be summarised such that moral intuitions are elicited via events that set off culturally shaped mental systems (foundations). These systems are preferentially pre-attuned to detecting *different types of moral content* – with any moral reasoning that follows from the intuitive judgement generally seeking to support and justify the intuitions in question. This can be illustrated with reference to studies of 'moral dumbfounding' (Haidt, Bjorklund & Murphy, 2000), wherein participants are seemingly unable to provide reasons as to why they judge some ostensibly harmless actions as immoral (e.g., consensual incest, carefully framed cannibalism), yet typically maintain this judgement despite their inability to articulate any rational justification for their position.

2.3. Foundation Criteria

For a construct (e.g., fairness) to be considered 'foundational' it must be 'common in third-party normative judgements', such that members of the community typically condemn (or praise) transgressing (or promoting) actions classified as relating to that concept, even if the actions have no direct consequences for the person making the judgement. It must also typically elicit 'automatic affective evaluations' among individuals witnessing actions categorised as instances of such (e.g., cheating). Foundations should be 'culturally widespread', such that most human cultures should exhibit the concept in some form or other. There should also be 'evidence for innate preparedness', such that precursors to the foundation are apparent in infants and non-human primates; and the evolutionary model should be able to demonstrate an adaptive advantage to individuals (or groups) who managed to attune themselves to the foundation - for example, they discovered 'reciprocal altruism' (Trivers, 1971) was a 'good trick' (Dennett, 2014).

2.4. Foundation Candidates

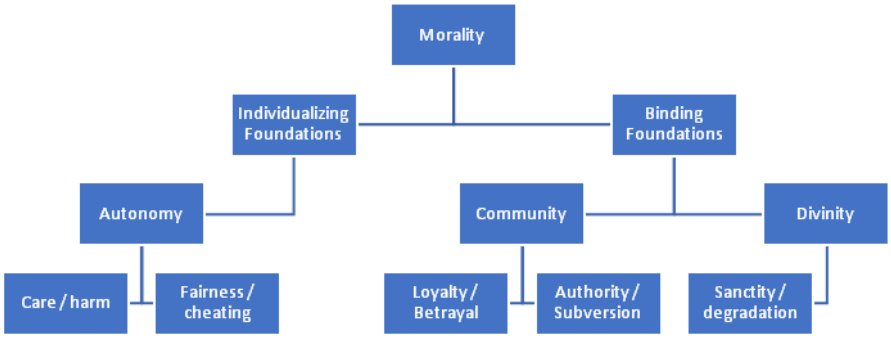
Graham et al. (2013) claim there are at least five different 'foundations' which meet these criteria. The ‘Care/harm’ foundation, originating from the challenge of caring for offspring, is triggered by perceiving instances of suffering, distress, cruelty, care, nurturance, and kindness. The ‘Fairness/cheating’ foundation provided a means to solve issues arising in the context of relationships and non-zero-sum exchanges, initially being attuned to detecting instances of cooperation and cheating among social interactions, and underpinning virtue concepts such as trustworthy, and just. The ‘Loyalty/betrayal’ foundation addressed the challenge of forming and maintaining cohesive alliances, is concerned with detecting threats or challenges to the group, and is related to values of patriotism and self-sacrifice. The ‘Authority/subversion’ foundation dealt with the problem of navigating social hierarchies, being attuned to detecting signs of rank or status, and is related to traits such as obedience and deference. And the ‘Sanctity/degradation’ foundation is argued to originate from the challenge of avoiding disease, pathogens, and contaminants – relating to a ‘behavioural immune system’, and the emotion of disgust, it is concerned with virtues of piety, chastity, temperance, cleanliness, and purity.

The eliciting stimuli which set off each of these mental systems currently extend beyond those which each foundation originally had success at detecting. Cute cartoon characters may trigger the 'Care' system, the 'Fairness' system may be triggered through interacting with a broken vending machine, sports teams may activate the 'Loyalty' foundation, 'Authority' foundation systems may be at work when interacting with police officers - and may be particularly sensitized amongst military service personnel, and the 'Sanctity' module may underlie concerns regarding immigration and genetically modifying food - being triggered by concerns about 'purity'. Graham et al. (2013) argue certain emotions are *characteristically associated* with each of the foundations; most notably linking anger with perceiving violations relating to 'harm' and 'cheating', and

disgust with actions perceived as 'degrading or impure', but also linking respect and fear with 'authority/subversion', and group pride with the foundation of 'loyalty'.

These five foundations readily translate into Shweder et al.'s (1997) 'big three' moral domains, with Care and Fairness relating to the ethic of autonomy, Loyalty and Authority to the ethic of community, and Sanctity to the ethic of divinity. These foundations may also be collapsed, such that the former two can be considered 'individualizing' foundations relating to autonomy based concerns of justice, rights, and the prevention of suffering – where the individual is the 'locus of moral value'; whereas the latter three may be considered 'binding' foundations, relating to concerns around social cohesion, duty, tradition, order, and other such concepts which consider the group as the 'locus of moral value' (Graham, Haidt, & Nosek, 2009). This range of classification levels is shown in Figure 2.1.

Figure 2.1. Moral Classification within the MFT framework.



Graham et al. (2013) state their list of five foundations is unlikely to be a complete list, with investigation of several other potential foundations described as ongoing. Their best candidate for a sixth foundation is that of 'Liberty/oppression', associated with individualizing foundations, and this foundation has demonstrated further pragmatic validity – with Iyer et al. (2012) showing that those of a 'Libertarian'

political persuasion identify concerns regarding liberty as more relevant than concerns relating to any of the other five foundations when considering whether something is right or wrong. Graham et al. (2013) also identify other likely candidates, with suggestions of foundations relating to 'Efficiency/waste', 'Ownership/theft' (property rights), and 'Honesty/deception'; and note further suggestions for new foundations (e.g. 'Industry' and 'Modesty'), plus potentially splitting existing ones (e.g. food vs. sexual purity). Furthermore, Graham et al. (2013) acknowledge critiques that MFT does not sufficiently attend to considerations of basic motives (e.g. approach/avoid - Sheikh & Janoff-Bulman, 2010) or relational contexts between involved individuals or groups (Rai & Fiske, 2011), providing even more space for theory expansion. However, regardless of how many new foundations may be added, or how the five they argue for might be reduced, Graham et al. (2013) present a strong argument that morality relates to more than one 'mental system', 'cognitive template', or 'foundation' – and detail the ways in which researching morality using a Moral Foundations Theory approach has demonstrated the pragmatic validity of pluralism.

2.5. Research drawing on Moral Foundations Theory

It is beyond the scope of the current discussion to review the whole range of research findings drawing on Moral Foundations Theory (for a list of primary research see https://moralfoundations.org/publications), but it is worth highlighting some of the key findings of relevance. Of MFT's four key claims, the majority of critiques are focused toward the claim of pluralism, with comparatively few critiques targeted towards 'intuitionism' or 'innateness' *per se*, and no apparent challenges to MFT's account of 'cultural learning' (Graham et al., 2013). This is likely because the claim of pluralism is the only one which is primarily related to morality and values, as research on topics of innateness, cultural learning, and intuitive thinking, has been extensively pursued in all manner of psychological studies. Although MFT draws on this literature to explain how

'moral foundations' may have been formed and shaped, and lay out how moral intuition, reasoning, and judgement might operate, this literature may also be used to to premise a monistic, one-system approach to morality. Indeed, there seems to be much agreement between theories with regard to the relevance of 'cultured intuitions', with disagreement focused more on the mechanics of these (i.e. whether culturally inculcated moral intuitions are 'harm-centric', or 'foundationally orientated'). Accordingly, the focus in what follows relates primarily to claims regarding (moral) pluralism.

MFT reports that the 'moral domain' is comparatively more restricted in cultures which are Westernized, Educated, Industrialized, Rich, and Democratic (WEIRD - Henrich, Heine, & Norenzayan, 2010), than in non-WEIRD cultures. This is demonstrable with reference to Shweder et al. (1997) where some of the vignettes they used to investigate the moral discourse in Orissa, India, would been seen by westernized individuals as depicting distinctly conventional transgressions which hold only as a matter of local custom, rather than high level violations of community and divinity moral codes, such as a widow eating fish two or three times a week. In contrast, other vignettes such as 'beating an insubordinate wife', which is not always considered a transgression in this culture (Shweder et al., 1997, p. 135), may be considered as relatively serious violations of 'western' moral codes, and reasoned about in different terms. For 'WEIRD' individuals, the husband's actions are likely considered wrong because they violate the wife's autonomy, whereas for non-WEIRD individuals, any wrongness is likely conceived of in terms of violating the natural/sacred order, rather than in terms of 'harm' or 'rights'. Furthermore, some of their vignettes illustrate substantial cultural differences regarding morality. A woman cooking food for her family and sleeping in the same bed as her husband during her menstrual period is considered a high-level violation of both autonomy and divinity moral codes in Orissa, whereas individuals from WEIRD cultures may consider anyone holding or expressing the attitude that such behaviour constitutes any kind of violation as a moral transgression in itself.

Socio-cultural differences in the breadth of the moral domain are also reported by Haidt, Koller, and Dias (1993), who investigated responses to vignettes depicting ostensibly harmless 'offenses' across high and low socio-economic status individuals in three cities. They found high-SES groups took a more permissive stance towards the actions in the stories, whereas low-SES groups took a more moralizing stance towards them - noting that for moralizing groups, moral judgements were better predicted by questions regarding whether the individual would be bothered by seeing the action, rather than whether someone was seen as being harmed. Additionally, research by Kim, Kang, and Yun (2012) found Korean participants tended to be more concerned with moral attitudes relating to 'purity' than participants from the United States, and that this concern was less affected by political ideology amongst Koreans. They also report a consistent trend between morals and politics, whereby participants identifying as politically liberal tended to discount concerns regarding loyalty, authority, and purity, compared to those identifying as conservatives, who tended to emphasize a similar level of concern across all foundations.

2.6. Moral Foundations and Political Orientation

A substantial portion of research drawing on MFT reports similar differences between liberals and conservatives across moral foundations, finding a generally consistent trend which forms the basis of MFT's most widely known claim. Liberals appear mostly concerned with matters relating to the 'individualizing' foundations, with little apparent regard for the 'binding' foundations, whereas conservatives tend to show similar levels of concern across all moral foundations – such that, graphically speaking, concern ratings of harm and fairness slope slightly downward, and concern ratings of loyalty, authority, and impurity, slope upwards, when moving along the liberal-conservative dimension from left to right.

Graham, Haidt, and Nosek (2009) conducted a series of studies which showed these political differences seem robust across different types of measurement. Responses to the Moral Foundations Questionnaire – a measure of the moral relevance of each of the foundations - followed the expected pattern; the greatest difference in scores between individualizing and binding foundations were reported by those self-identifying as strongly liberal – whereas strongly conservative participants reported the fewest differences between these scores. This pattern was also present over moral judgements relating to different foundations; and over responses to moral trade-offs, with scores on the Moral Foundations Sacredness Scale – a measure of how much a person would need to be paid to complete an action typically regarded as violating a foundational value - showing liberals were more willing to violate 'binding' foundations than were conservatives. Furthermore, the pattern of emphasis was also present in an analysis of naturally occurring text, with Graham et al. (2009) reporting sermons from 'liberal' churches placed more emphasis on 'Harm/care' and 'Fairness/cheating' than sermons from 'conservative' churches, which talked more about matters relating to 'Authority/subversion' and 'Sanctity/degradation'. They also found suggestive evidence of differential category use, where liberals used more 'Loyalty/betrayal' relevant words, but conservatives used them in ways consistent with endorsing the foundation, whereas liberals typically used these words in rejecting the foundation.

Differences in naturally occurring text have also been found across a range of studies. Among the results, Day et al. (2014) found that framing an issue with reference to a particular moral foundation resulted in greater reference to the framed foundation when asked to provide a few written points in support of their judgements about the issue. Stolerman and Lagnado (2020) analysed newspaper articles regarding human rights, reporting that these articles emphasized individualizing foundations, but conservative newspapers made more use of terms relating to the binding foundations than liberal newspapers. McAdams et al. (2008) report liberals and conservatives use markedly different language when asked for details regarding the nature of their religious

and moral beliefs, and how these developed, with conservatives using more 'binding' language and liberals more 'individualizing' language – a trend also found by Rempala, Okdie, and Garvey (2016), with participants drawing on different language when justifying their choice of political party. Harper and Hogue (2019) report that official materials from the 'Leave' campaign in the United Kingdom's 2016 referendum used more words related to 'Authority' and 'Liberty' than those of the 'Remain' campaign.

Furthermore, the pluralistic approach used by MFT can illustrate ideological differences 'above and below left-right' dimensions. Haidt, Graham and Joseph (2009) factor analysed over 20,000 responses from American participants to various personality measures and the Moral Foundations Questionnaire. Their results supported a four-factor solution, where groupings can be summarised according to whether they place high or low emphasis on individualizing foundations and whether they place high or low emphasis on binding foundations – for example, high emphasis on all foundations may relate to a 'Religious Left' ideology, in contrast to a low emphasis on all foundations, which appeared related to 'Libertarianism'. Haidt and Graham support these findings with reference to pre-existing ideological 'master narratives', illustrating how each of these draws (to varying extents) on different foundations to advance a particular vision or worldview.

2.7. Challenges to Moral Foundations Theory

However, Davis et al. (2016) suggests the apparent relationships between moral foundations and political orientation may be differentially affected by measurement issues, such as sample demographics. They propose that because MFT has relied on predominantly White samples, and because religiosity and conservatism tend to be positively related within such samples - but not within Black samples - that the extent to which liberal-conservative differences can be generalized may be overstated. Davis et

al. raise concerns regarding how the 'binding' foundations are measured, and how certain foundational virtues may be understood and upheld by different individuals/groups, which call into question both the strength, and validity, of MFT's politics-based findings. Similarly, Kugler, Jost and Noorbaloochi (2014) demonstrate MFT's liberal-conservative differences may be attributable to higher levels of Authoritarianism amongst conservatives and lower levels of Social Dominance Orientation amongst liberals. They found 'binding' concerns positively associated with support for discrimination and intergroup hostility, whereas 'individualizing' concerns were negatively associated with these constructs - although this too may follow from how MFT measures the relevance of the binding foundations.

Janoff-Bulman and Carnes (2013, 2016) suggest that the reported relationships between moral foundations and political orientation may stem from how MFT has operationalised the binding foundations. Their 'Model of Moral Motives' combines two basic motives (approach/avoidance) directed across different areas of moral focus (self, other, group), producing a six-celled taxonomy allowing it to identify moral concerns yet to be incorporated by MFT - which includes the self-focused concerns of 'industriousness' and 'modesty' previously identified in another critique by Suhler and Churchland (2011). Importantly, Janoff-Bulman and Carnes propose that the 'individualizing' foundations cover prescriptive and proscriptive other-focused actions, but that MFT's binding foundations cover only proscriptive group-focused actions. Accordingly, liberal-conservative differences may be attributable to differences in basic motives, such that conservatives may place greater emphasis on matters related to 'proscribing and protecting', whereas liberals may consider 'prescribing and providing' as more morally relevant. As MFT leaves such prescriptive group-focused actions relatively unexamined, this would readily explain why liberals seem to discount concerns related to foundations which bind groups together - the things that bind liberals together may differ from those that bind conservatives together, and MFT only measures the latter.

Yet whilst these critiques may challenge MFT's politics-based findings, and suggest MFT style pluralism may benefit from refinements, they remain favourable towards 'many-system' moral theories, whereas the most persistent challenge to MFT comes via a 'one-system' theory - the Theory of Dyadic Morality (Gray, Young, & Waytz, 2012; Gray, Schein, & Ward, 2014; Gray & Keeney, 2015a; Schein, Ritter, & Gray, 2016; Schein & Gray, 2015, 2016, 2018). This theory largely agrees with MFT's four key claims, but disagrees with its pluralistic structure of content, maintaining that morality is all about perceived harm. For Dyadic Morality, perception of harm requires perceiving a causal link of suffering between two perceived minds, such that 'harm pluralism' arises from different cultural intuitions regarding who can do harm (agents), how harm can be done (causality), and what can be harmed (victims). Violations of moral foundations, such as loyalty or purity, may be considered wrong to the extent they are perceived as being related to the three elements of harm, rather than the extent to which they may be perceived as being disloyal or impure *per se* - moral foundations are just taxonomic categories detailing certain tri-partite, multi-factor mixtures of intuitively perceived harm.

Dyadic Morality has raised concerns regarding the conceptual clarity of the foundations, arguing there seems to be a high degree of overlap between foundations and other constructs, and between the foundations themselves. For example, 'Sanctity' has been shown to be strongly correlated with Right Wing Authoritarianism (0.7, Table 7. in Graham et al., 2011), and with the 'Authority' (0.8) and 'Loyalty' (0.72) foundations - which also correlate with each other (0.88), as do 'Fairness' and 'Care' foundations (0.72, Figure 2 in Graham et al., 2011). However, whilst these results may be explainable with reference to individualizing vs. binding factors, a reported correlation of >.86 between the harm and purity foundations (Gray & Keeney, 2015a,b), which are often regarded as exemplars of these factors, would seem to present a serious challenge to the moral foundations account. Given the extent to which Dyadic Morality perseveres in its critique of MFT, it would seem prudent to examine it in further detail.

Chapter 3 - On the Theory of Dyadic Morality

A comprehensive defence of the Theory of Dyadic Morality (TDM) has been offered by Schein and Gray (2018), having previously been advanced by Gray, Young and Waytz (2012), as well as elsewhere in literature (Gray, Schein & Ward, 2014; Schein & Gray, 2014; 2015; 2016). Dyadic Morality proposes that moral judgement operates via a 'harm-based' cognitive template, such that ‘dyadic harm’ involves a causal link between two perceived minds -- an intentional agent, and a vulnerable/suffering patient -- with the former causing 'damage' to the latter. TDM claims that all moral transgressions can be understood as interpersonal harms (Gray, Young & Waytz, 2012) because all such transgressions take the form of "an intentional agent causing damage to a vulnerable patient" (Schein & Gray, 2018, p.1) which violates norms and generates negative affect. The dynamic causal structure of TDM proposes reinforcing, bi-directional links between the ‘harm’ template and moral judgement. As such, perceiving any of the three elements (agency, causality, patiency) with greater salience causes acts to be judged as more immoral, but equally judgements of immorality can also cause acts to be seen as more 'harmful' in the dyadic sense. Acts judged as immoral may lead to attributing agents a greater capacity for thinking and doing (agency), attributing patients a greater capacity for vulnerability and feeling (experiencing suffering), and drive the search for causal links between agent and patient.

Making a moral judgement, or making the move from perception of dyadic harm to judgement of immorality, relies on two further components: norms, and negative affect. Schein and Gray (2018) liken these components to fire, in that moral judgements arise once all three components are sufficiently present. Moral violations always contravene norms, and norm violations typically produce negative affect in observers, but the key component is the perceived presence of dyadic harm - instances of which

are typically norm violations that elicit negative affect, such that dyadic harm may provide all three components of immorality by itself. TDM claims the perception of dyadic harm accounts for the distinction between moral violations and violations of (social) convention, such that acts which are harmful are typically both norm violations and provoke negative affect; whereas neither norm violations or negative affect, alone or in conjunction, seem sufficient to categorise an act as immoral. By way of example, Schein and Gray (2018) explain that whilst a person spitting on their own food before consuming it is disgusting, counter-normative, and might be labelled as wrong, the act seems to lack the necessary qualities to be labelled immoral -- it lacks dyadic harm.

3.1. Defining Dyadic Harm

In defining dyadic harm, Schein and Gray (2018) argue that ‘harm’ is non-binary, with acts classified subjectively based on intuitive perceptions of the three dyadic elements, rather than classifying harmful acts along more objective and rational lines that involve only physical or emotional suffering. There are a broad range of potential agents (e.g., people, corporations, governments, divine beings), potential patients (e.g., children, animals, the environment, future selves), and varieties of causing damage - which include 'spiritual defilement' and 'mental suffering' in addition to obvious physical damage. On this approach, dyadic harm and immorality lie along corresponding 'intuitively perceived continua' whereby maximally dyadic actions - those most clearly perceived as an intentional agent causing damage to a vulnerable patient, seem the most immoral; and minimally dyadic actions – those least clearly perceived as instances of dyadic harm, seem least immoral. Schein and Gray’s (2018) dyadic definition of harm readily explains why a majority of participants do not change their judgements about the immorality of the 'harmless wrongs' used in moral dumbfounding studies - as although the vignettes are written so that they do not ‘objectively’ contain harm, participants are intuitively perceiving some variety of dyadic harm. Furthermore, Schein and Gray (2018)

propose that all moral values can be understood as 'intermediaries of harm'. In this manner, valued concepts or norms, such as those described by Moral Foundations Theory (i.e., fairness, authority, loyalty, purity), can be seen as vulnerable entities. This allows for a second link to dyadic harm via the suffering or destruction of a value which, in turn, is a cause of harm. For example, failure to observe the 'correct' post-death customs can be perceived as a source of direct harm, in that suffering is caused to the spirit of the deceased, but this failure can also be perceived as an indirect source of harm - such missteps lead the failing agent to ‘lose their sanctity’ which, in turn, can be seen as a direct harm as it would result in their palpable suffering.

3.2. Defining Dyadic Mechanisms

In terms of the mechanisms underlying Dyadic Morality, moral judgement is proposed to work by means of dyadic comparison, by which the more an action seems like it involves an intentional agent causally damaging a vulnerable patient - the more immoral the action is considered to be. Moral judgements are also held to affect perception through the reciprocal process of dyadic completion, whereby the more immoral the action is considered to be - the more it seems like it involves an intentional agent causally damaging a vulnerable patient. This completion process may be ‘agentic’, ‘patientic’, or ‘causal’, depending on which element of the dyad is ambiguous, unclear, or seemingly absent. Completion is compelled by the perception of any two elements of dyadic harm, such that once perceived, observers may seek to find a causal link between wrongdoing agents and suffering patients, perceive intentional agents to account for the caused suffering of vulnerable patients, or seek to find a victim that suffers as the result of perceiving an intentional agent violating norms. Furthermore, Schein and Gray (2018) claim that comparison and completion form a dyadic loop through their mutual reinforcement, and thus account for the process of moralization.

3.3. The Main Hypotheses of Dyadic Morality

The Theory of Dyadic Morality makes a number of testable predictions. Firstly, increased perceptions of any element of the moral dyad should also increase judgements of wrongness - actions where perceptions of agency, damage, or vulnerability are more salient should be judged as more wrong. Secondly, increasing the perceived responsibility of the patient should lead to lower attributions of suffering, and increasing the perceived suffering of the agent should lead to lower attributions of responsibility - more agentic patients suffer less, and more patientic agents are less responsible. Thirdly, the primacy of dyadic harm in moral judgement, and the intuitive nature of dyadic harm, should be apparent in the association between dyadic harm and moral judgement in studies that use implicit measures or cognitive load tasks. Fourthly, moral condemnation (and by extension, the breadth of individuals moral perception and their propensity to moralize) should align with individual differences regarding threat sensitivity. Finally, exceptions to TDM should be "rare, unstable, and maintained only with effortful reasoning" (Schein & Gray, 2018, p. 12). TDM states that it would be falsified if there is no causal link between perceptions of dyadic harm and immorality, provided it is assessed intuitively and with controls in place to account for affect and norms.

3.4. Dyadic Morality and Moral Foundations Theory

Schein and Gray (2018) claim Dyadic Morality agrees with Moral Foundations Theory (MFT) on several key points, such that perceiving harm is sufficient for causing a moral judgement, that harm is the most typical moral consideration, and that morality is innate, intuitive, culturally inculcated, and pluralistic (in a way). However, TDM disagrees with MFT about claims regarding modularity, arguing instead that moral pluralism emerges through variations in norms, affect, and pluralities of dyadic harm – which

emerge through variations in types of agents, patients, and the ways in which damage can be caused. TDM also contests the existence of specific and distinct links between moral domains and particular emotions, arguing evidence for such links can be explained with reference to ‘core affect’ and conceptual knowledge. Furthermore, TDM challenges several of MFT's explanations of results, suggesting that differences between liberals and conservatives may be better explained by how they each perceive (dyadic) harm, that differences in judgements across moral domains may be explained by confounds (e.g., weirdness), and that the role of disgust in predicting moral judgement is not direct, instead being mediated by the perceived dyadic harm involved. The account of TDM put forward by Schein and Gray (2018) leads them to conclude that MFT provides a taxonomy of specific types of content (or varieties of harms, norms, and affect) which are highly likely to be moralized, but that MFT does not explain moral judgement, arguing instead that moral judgement is driven by the intuitive perception of dyadic harm. For TDM, morality operates on domain-general mechanisms, rather than domain-specific ‘foundations’.

However, whilst previous work on Dyadic Morality (e.g., Schein & Gray, 2015) has prompted clarifications regarding aspects of MFT, such as how MFT uses 'modular' definitions, and the relative emphases placed on foundations by liberals and conservatives, the most recent defence of TDM (Schein & Gray, 2018) does not appear to account for the concerns raised by Haidt, Graham, & Ditto (2015). Although Haidt et al. (2015) praise the notion of a dyadic template, highlighting work on dyadic completion as particularly merit-worthy, they identify 'strong' and 'weak' versions of TDM, both in definitions and hypotheses. The strong version requires, by definition, the perception of two interacting minds with a causal link of suffering between them, which Haidt et al. (2015) consider analogous to the Care/Harm foundation; and the strong hypothesis states this template is necessary and sufficient, such that all morality is understood in this way (Gray, Young, & Waytz, 2012). In contrast, the weak version hypothesizes that dyadic harm is the most important template, with no definitional claim to necessity -

which is uncontroversial, especially if damage or suffering may be construed as anything objectionable, and anything can take the role of the patient. Haidt et al. (2015) suggest that Schein and Gray (2015) only have sufficient evidence to back up the weak version; and although the theory is more clearly stated by Schein and Gray (2018), evidence in support of the strong version still appears to be lacking.

The main thrust of Haidt et al.'s (2015) critique targets Schein and Gray's (2015; 2018) claims of 'harm pluralism', as the means by which damage can be done under TDM (2018, p.3) are limited to what MFT would regard as 'harm' (physical destruction or mental suffering) or 'sanctity' (spiritual defilement). Yet whilst it may be feasible to argue defilement might be considered as a variety of damage under TDM, the range of damaging acts - broad enough to include anything potentially construed as objectionable - diminishes their claim to 'moral judgements are about dyads'. Furthermore, whilst being viable to argue that the type of patient (or agent) of a given act may affect moral judgement, the variety of potential patients, which can include 'foundational values', further diminishes the claim to 'moral judgements are about social relationships'.

Yet if Schein and Gray are correct that actions are judged as immoral to the extent they are perceived as conforming to a template of dyadic harm, and both dyadic harm and immorality occupy non-binary continua, then the range of potential acts may not necessarily be an issue. That values may directly or indirectly take on the role of moral patients is more problematic for TDM, as this seems to contradict the strong definition which requires two interacting minds. If the patient is a group of people, or society more broadly, then it would seem 'two (or more)' would be more accurate – or questionable whether the group can be reduced to ‘one mind’. If the patient is the environment, or a 'foundational value', then these would seem to lack the vulnerable feeling capacities required to be classified as a moral patient. Furthermore, if the patient is the self (or future self), and the agent is also the self (i.e., self-directed wrongs), then only one mind would seem to be present. These concerns regarding patiency highlight a

potential weakness of this element in Dyadic Morality, and suggest the ‘type' or 'role' of patient may need limiting. This might be achieved by adding a 'reversibility requirement' such that the patient could, in principle, take the role of the agent -- although this would likely remove 'foundational values' and 'the environment' from taking a patient role, and potentially require TDM to abandon claims that 'foundational values' are intermediaries or transformations of dyadic harm. Alternatively, it may be possible to remove the patient from the theoretical formulation, yet this would seem equally problematic for a strong version of Dyadic Morality. Although the strong version of TDM may cover what MFT terms the harm/care foundation, it seems to over-reach if claiming that all morality can be accounted for with this template; whereas the weak version of TDM, where (dyadic) harm is merely the most important template, is non-controversial.

3.5. Further Critiquing Dyadic Morality

Throughout Schein and Gray's (2018) defence of Dyadic Morality there are several points which are underemphasized, but noteworthy enough to highlight in full, as they all seem to be concessions of various kinds. Firstly, "...the mere perception of suffering and vulnerability is not enough to give rise to a robust moral judgment. Instead, one must care about the vulnerable mind via empathy (Baron-Cohen, 2011; Eisenberg & Miller, 1987)." (p.7). This suggests there is further scope to quibble with the necessity of a patient, as although empathy and morality are often associated, it is not clear that empathy is a necessary component of moral judgement, moral development, or moral conduct (Prinz, 2011). Secondly, when describing the moral-conventional distinction they pick up on the importance of blame, "[i]mmoral acts are also seen as intrinsically deserving of blame...and as intrinsically tied to outrage and other negative emotions." (Schein & Gray, 2018, p.4), and the link to emotion, both of which are emphasized by Prinz (2009). To describe an act as blameworthy seems similar to describing that act as 'agentically caused'. Thirdly, although the moral dyad is supposedly doing a lot of the

'work' in moral judgement, it seems to neither initialize nor finalise this judgement - "[h]arm-based processes of moral judgment are likely initialized only once a norm violation is noticed, and the strength of the final moral judgment hinges on its associated negative affect, which is typically integral or related to an act" (Schein & Gray, 2018, p. 17) - despite norms and affect being concepts which both feature prominently in sentimentalist theories of morality (Nichols, 2002; Prinz, 2009), and maybe sufficient for moral judgement on Nichols (2004) approach. Schein and Gray (2018) propose it is perceived (dyadic) harm that makes negative norm violations immoral. Yet if norms initialize, and emotions finalise, then it is unclear whether use of a 'harm template' is necessary.

Finally, in discussing how TDM might resolve moral disagreements, Schein and Gray (2018) suggest that in "...helping people understand that "the other side" *respects the sanctity of harm*, dyadic morality may help reduce the vindictiveness of moral conversations." (p. 4, *emphasis mine*). The inclusion of sanctity *in defence of harm* exposes further weaknesses in the formulation of Dyadic Morality. On one reading, this could be read as 'respects the primary importance, and inviolability, of intentional agents causing suffering to vulnerable patients', which seems a relatively trivial statement when considering TDMs scope for 'harm pluralism' (weak dyadic morality). Another reading seems to allow that 'sanctity' might precede 'harm'. Schein and Gray's (2018) argument redefines harm pluralistically; however, it may be just as plausible to argue for 'impure pluralism' by redefining sanctity. A further reading suggests dyadic harm is a value, or sacred norm, such that it becomes 'respects the immorality of *an intentional agent causing damage to a vulnerable patient*'; as such, 'strong' Dyadic Morality would seem to beg the question of why *this* is immoral. Whichever reading is taken, it would appear that actions or behaviours which would be classed as ‘impure’ by Moral Foundation Theory pose potentially serious issues for the Theory of Dyadic Morality as outlined by Schein and Gray (2018).

Furthermore, the importance of 'impurity/sanctity' concerns are emphasized by another theory which shares substantial common ground with both Dyadic Morality (Schein & Gray, 2018) and Moral Foundation Theory (Graham et al., 2013). Constructive Sentimentalism (Prinz, 2009) considers 'impurity' to be of equal standing to 'harm', contra TDM; although it allows that 'harm' may be the most common or salient concern, and that all 'foundational values' -- except impurity/sanctity -- might be explained as transformations or intermediaries of super-ordinate values, contra Moral Foundations Theory. Thus, whilst TDM may maintain that ‘harm’ is the most important, and perhaps most populous category, of super-ordinate values - maintaining that immorality is all about dyadic harm requires that actions typically classed as ‘impure’ can be both accounted for, and better explained, by reference to the elements of dyadic harm than by alternative approaches which consider ‘impurity’ as ‘foundational’. Accordingly, the Theory of Dyadic Morality has targeted key aspects of Moral Foundations Theory’s methodology in a way that seeks to undermine domain specificity, and secure ‘sanctity’ within the template of dyadic harm.

Chapter 4 - Impure, not ‘Just Weird’ – 'Sanctity' cannot be explained away by confounded stimuli

Morality is an innate, culturally learned, intuitive, and pluralistic phenomena – so say both Moral Foundations Theory (MFT, Graham et al., 2013) and the Theory of Dyadic Morality (TDM, Schein & Gray, 2018). However, these theories disagree with regard to what counts in terms of moral content, and the mechanisms involved in morality. MFT claims there are at least five ‘moral foundations’ – domain specific mental systems which are prepared in advance of experience to detect a range of conduct (originally) relevant to solving recurrent social problems, and which are attuned through cultural development such that individuals learn to intuitively navigate the shared meanings and evaluations of their society. In contrast, TDM claims morality is the product of domain general processes, arguing that what really matters is the intuitive perception of dyadic harm – defined as an intentional agent causing damage to a vulnerable patient – such that moral foundations are just transformations or intermediaries of dyadic harm. Thus, on MFT’s approach actions are judged as immoral with reference to intuitions generated via moral foundations, whereas under TDM actions are judged as immoral with regard to the extent they are perceived as instances of dyadic harm.

To maintain the argument that morality operates via a harm-based template, TDM needs to be able to account for the range of findings contributing to the pragmatic validity of MFT – where investigating different types of moral content has produced a wealth of literature suggesting that moral foundations function and operate differently. Yet whilst it is plausible to argue that concerns regarding fairness, loyalty, authority, and liberty, might be accounted for within the proposed template of dyadic harm – which requires perceiving two causally linked minds – it is not so easy to explain concerns

regarding sanctity/degradation, where violations may, in cases, be ostensibly harmless and lacking any verifiable victims. Furthermore, there is a substantial range of evidence suggesting the 'Sanctity/degradation' foundation is dissociable from the 'Care/harm' foundation – 'Sanctity' is better predictive of certain thoughts and behaviours, and certain cultural group differences, with judgements relevant to this foundation able to perform functions that 'Harm' cannot. 'Sanctity' also appears to have a different cognitive profile, and there are a variety of research findings that would be challenging to explain in the absence of this moral foundation.

For example, purity is predictive of moral concerns regarding suicide (Rottman, Keleman & Young, 2014) and is more concerned with transgressions relating to the self (Chakroff et al.; 2013, Chakroff & Young, 2015, Ulhmann & Zhu, 2013). The language of purity appears predictive of attitudes towards stem cell research (Clifford & Jerit, 2013), online social network distance (Dehghani et al., 2016), and the politico-religious ideology of sermons (Graham, Haidt & Nosek, 2009), as well as other ideological positions (Koleva et al., 2012; Day et al., 2014; Feinberg & Willer, 2013, 2015). Also, with regard to associated emotions, moral anger (harm-linked) appears more responsive to changes in circumstances than (purity-linked) moral disgust (Russell & Giner-Sorolla, 2011a) and, unlike disgust, also responds to considerations of intentionality (Russell & Giner-Sorolla, 2011b, Young & Saxe, 2011). Accordingly, TDM has focused on a variety of issues in order to tackle these differences, with the aim of subsuming Sanctity under the template of dyadic harm.

Schein and Gray (2018) argue one such issue is that of potential confounds acting on 'Sanctity' measures, supporting this claim with reference to research which suggests impure violations may be 'just weird'. Gray and Keeney (2015a) show that commonly used MFT impurity scenarios are both weirder, and less severe, than both commonly used MFT harm scenarios and 'naturalistic' impurity scenarios generated by participants. They also propose that weirdness and severity, rather than moral content,

may explain differences between evaluations of acts and moral character. Strikingly, Gray and Keeney (2015a) report that impure scenarios (e.g., a person having surgery to attach a tail) were rated as less impure than harm scenarios (e.g., making cruel remarks about the appearance of an overweight person), and that 'naturalistic' scenarios - participant generated examples of 'harm' and 'impurity' scenarios - were rated as more impure than MFT scenarios. Accordingly, they argue that harm scenarios appear to capture impurity better than researcher devised impurity scenarios, and that the extent to which harm and impurity seem correlated poses a problem for accounts favouring a modular approach to morality.

Extending this argument, Schein and Gray (2018) suggest findings taken in support of the existence of 'objectively harmless' moral wrongs are problematic because such cases tend to use weird scenarios during assessment, stating "MFT scenarios tapping moral judgment […] confound moral content with weirdness and severity" (p.24). As weirdness represents the extent of a norm violation, and this is a factor in judgements of immorality (along with negative affect, and dyadic harm), Schein and Gray (2018) claim that "studies arguing for a special link between purity and character (Uhlmann & Zhu, 2013; Young & Saxe, 2011) confound purity with weirdness when they use bizarre scenarios. When participant-generated examples of impurity (e.g., pornography, prostitution) are used to eliminate these confounds, any apparent differences disappear" (p.22). In short, Schein and Gray's (2018) argument can be summarized such that concerns about 'Sanctity' are really concerns about 'weird harms'.

However, Gray and Keeney's (2015) research focuses on certain *'commonly used'* MFT scenarios. Thus, whilst they may have a point that *these* scenarios might suffer from confounds, it seems this point is unlikely to generalize to all MFT scenarios in the way that Schein and Gray (2018) suggest by not restricting their statement to this effect. Furthermore, in critiquing Gray and Keeney's (2015a) research, Graham (2015, p.5) claims that "atypicality [weirdness] is in fact a primary feature of the

Purity/degradation foundation". Graham agrees that "the items chosen to represent MFT are indeed the weirdest of the weird" (p.7), but argues that this is not problematic for MFT style moral pluralism - and identifies specific issues with the methodology which may undermine Gray and Keeney's (2015a) conclusions.

In particular, Graham (2015) states that near perfect correlations between ratings of harm and impurity strongly suggest these measures were tracking wrongness, rather than content. Graham also notes that two impurity scenarios generated by participants for the second study involve transgressions across both harm and impurity domains, and thus fail to distinguish between these domains with regard to methodology. Graham further notes that the materials used to make claims about differences between act and character evaluations depict 'mixed' violations (e.g., adultery), and that the (manipulated) presence of weirdness is unrelated to the act (e.g., the same act is done by someone painted red and wearing a cape made from human hair). As such, Gray and Keeney's (2015a) third study ends up contrasting (potentially) consensual adultery with sexual assault within impurity measures, and contrasts these with harm scenarios depicting 'simple' assault (e.g., face slap, step on foot), rather than providing clear instances of harm and impurity violations. However, despite Graham's (2015) critique, Gray and Keeney (2015b) reply that they have 'disconfirmed Moral Foundations Theory on its own terms'.

To add to Graham's (2015) critique, it is not entirely clear why Gray and Keeney (2015a) choose to explain away the construct of impurity. The question asked to their participants in Study 1 is "How impure [i.e., involving sinfulness, indecency, dirtiness] is this act?", and to generate "three impure violations (''sinful, dirty, degrading, lustful, or indecent'')" in Study 2. However, although they realise that "''impure''— that is, ''sinful'' or ''indecent''— is synonymous with ''morally wrong'' (Oxford English Dictionary, n.d.)." (p.861), they do not acknowledge this actually jeopardises their claims. The further claim that "[b]ased upon naturalistic scenarios and the overlap between harm and impurity,

perhaps we can simply define participants' understanding of impurity as ‘‘(perceived) harm involving sex.’’" (p.866), seems similarly unjustified. No explanation is provided as to why overlapping responses appear to be favourably removed in order to make such a claim - murder, theft, drug abuse - all of which do not ostensibly involve sex, were generated by participants for impure violations. Furthermore, 'harm' appears in their word-cloud for impurity, but 'impurity' does not appear under harm. As such, based on Gray and Keeney's own results, it seems plausible to make an opposing claim - harmful acts are also impure.

The impurity of harmful acts may explain why these acts tend to be rated as more severe, or more morally wrong, than (solely) impure acts; there may be an additive component which is either substantially weaker, or simply unavailable, when considering impure violations which are not obviously harmful - such as receiving a blood transfusion from a child molester. Indeed, when considering acts involving 'direct physical harm', these are likely, as a matter of definition, to include desecration and/or destruction of the body (lacerations, bruising, broken bones, and so on) to some extent, and are a clear violation of the 'body as a temple' narrative associated with the purity foundation. Additionally, moral foundations not examined in Gray and Keeney's research (2015a) may also impact on ratings of wrongness. These too may be additive, such that 'unfair' harm may be seen as worse than harm alone, or may mitigate any effect, such that harming a member of an out-group may be seen as less wrong than harming a member of the in-group. However, TDM may be able to provide some account of these issues with reference to cultural conceptions of 'harm' (Schein & Gray, 2018).

Further critiquing Gray and Keeney's (2015a) research, the majority of scenarios seem to be drawn from the Moral Foundations Sacredness Scale (MFSS; Graham & Haidt, 2012) - which is a measure of how much financial recompense it would require for participants to carry out actions regarded as violating different foundations. Yet all MFSS items involve self-perpetrated actions which may not adapt well for use in alternative

contexts – such as when these ‘wrongs’ are being performed by someone other than the participant, and the question is ‘how wrong?’ rather than ‘how much?’. Indeed, the transfer of these scenarios for use as reaction-judgement scenarios seems likely to be a key factor underlying Gray and Keeney's (2015a) results, as it changes the way in which these scenarios are instrumented.

In the 'paid-to-perpetrate' format, events are predominately focused on the agents’ willingness to commit various acts of transgression, such that both 'you kicking a dog' and 'you surgically adding a tail' are primarily focused on the extent to which doing them is bad for (you) the agent. It is presumed the respondent sets some value in avoiding both such activities, and all events are instrumented on the principle of an increasing bank balance. In contrast, when these scenarios are described such that someone else is the agent, the MFSS harm scenarios cover judgements regarding the wrongness of another inflicting harm on a third-party (e.g., they kick a dog), whereas the MFSS impurity scenarios remain concerned with self-targeted transgressions (e.g., they get *themselves* a tail) despite the change in perpetrator. An individual may consider body modification wrong for them, but permissible for others, whereas kicking the dog remains wrong no matter who does it.

It seems entirely plausible that self-victimization is considered to be both weirder and less morally severe than other-victimization, particularly as the latter allows for the inference of some instrumental purpose which the former prohibits. This issue alone may explain Gray and Keeney's (2015a) results, as their 'naturalistic' participant-generated impurity scenarios include items covering actual or implied other-victimization (adultery, rape, prostitution), and items which may be seen as providing some instrumental benefit to the individual - either via gaining actual or implied earnings (making porn, stripping) or presumably gaining some form of satisfaction. In contrast, the 'commonly used' MFT scenarios mostly imply any negative consequences will only impact the offending-self, and are substantially harder to justify as providing any benefit

to the individual in question. Yet Schein and Gray (2018) might argue it is also plausible that self-victimization is less severe because it is less dyadic. However, doing so would seem to undermine Gray and Keeney's (2015a) claims that impurity conflates weirdness with severity. Commonly used MFT impurity scenarios are weirder because weirdness is a primary feature of impure violations, but they are less severe because they are *non-beneficial self-victimizing actions being performed by another person* - not because they are 'just weird'. Indeed, it has been hypothesized that moral domains may be (partly) defined by the target of the action, in addition to the type of action, such that 'impurity' is predominantly self-focused whereas 'harm' is primarily other-focused (Chakroff, Dungan, & Young, 2013; Dungan, Chakroff, & Young, 2017).

4.1. The current study (Study 1)

Assuming differences across dimensions of weirdness and severity are problematic (as per TDM), rather than just ways in which harmful and impure moral content differ (as per Graham, 2015) - the issue of principal importance is whether differences on these factors impact on 'MFT scenarios' in general (as per Schein & Gray, 2018), or whether this is limited to *'commonly used* MFT scenarios' (as per Gray & Keeney, 2015a). However, given the above critiques, the findings in relation to the commonly used scenarios may simply be accepted - the results appear readily explainable, *prima facie* predictable, and mostly replicable (see Franchin et al., 2019). Accordingly, the focus here is on the stronger claim that factors of weirdness and severity may impact on other MFT impurity scenarios, and impurity scenarios more generally, whereby impure transgressions might be considered as generally just weirder and less severe varieties of harm.

The current study improves on the previous approach in various ways. First, the majority of scenarios used have previously been validated specifically with regard to MFT taxonomy (Clifford et al., 2015), and are written as ‘observations’ of morally relevant conduct, rather than having been adapted from ‘paid to perpetrate’ events. Second, a greater number and range of scenarios are used to better ensure the results generalize across a variety of violations. Third, it addresses concerns regarding the ratings of harm and impurity, removing the word 'sinful' from items gauging impurity (which Gray & Keeney, 2015, acknowledge as being equivalent to 'morally wrong'), and adapting these two uni-polar ratings onto one bi-polar scale to more clearly assess participants interpretation of the balance of content involved (which may also help to avoid potential ceiling effects). Participants may still endorse equivalent levels of harm and impurity by selecting the mid-point of the scale, but this prevents them from treating this rating as a proxy for wrongness - as may be the case in Gray and Keeney's (2015a) results where measures of 'harm', 'impurity' and 'wrongness' are nearly perfectly correlated. Finally, as there are a limited number of validated MFT impurity scenarios, and these also seem quite weird, a range of scenarios were researcher generated with the aim of capturing less weird, but more severe, violations of the 'Sanctity' foundation to investigate whether the issue is specific to MFT scenarios or might apply across scenarios depicting impurity more broadly.

The current study also investigates the relationship between moral foundations and perception of moral character. In critiquing Gray and Keeney (2015a), Graham (2015) cites Chakroff and Young (2015) who report links between impure actions and character attributions - such that people committing impure actions are perceived as having worse moral character. Furthermore, the study examines the relationships between moral foundations and emotions, whereby violations of autonomy (i.e., harm) are thought to be associated with anger, and violations of divinity (i.e., impurity) with disgust (Rozin et al., 1999; Prinz, 2009; Graham et al., 2013) – associations which TDM argues against (Cameron et al., 2015).

Following Gray and Keeney (2015a), it is hypothesized that harm scenarios are expected to be rated as more morally wrong (i.e., more severe) than impurity scenarios (H1); following Chakroff and Young (2015), that agents are expected to be evaluated as having poorer moral character for impurity violations than harm violations (H2); and following Graham (2015), impurity scenarios are expected to be rated as more atypical (i.e., weirder) than harm scenarios (H3). Following Moral Foundations Theory (Graham et al., 2013) and Constructive Sentimentalism (Prinz, 2009), it is further hypothesized that harm scenarios are expected to be rated as more angering than impurity scenarios (H4) and impurity scenarios are expected to be rated as more disgusting than harm scenarios (H5). Finally, and as a means of assessing categorisation of content, it is hypothesized that (pre-classified) harm scenarios are expected to be rated as more harmful than impurity scenarios, and (pre-classified) impurity scenarios as more impure than harm scenarios (H6).

4.2. Method

4.2.1. Design

This study investigates participant perception of moral content (harm vs. impurity) along dimensions of wrongness, character, abnormality, anger, disgust, and moral domain 'balance' - with inclusion of a between-subject factor whereby two groups respond to conceptually matched stimuli sets.

4.2.2. Participants

G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) indicated a minimum sample size of 170 would be required to have 99% power for detecting a medium effect size (0.2) at an alpha level of .01. An opportunity sample was recruited via adverts placed on the University research participation management system (SONA), with participants offered research credit for taking part in an online survey taking around half an hour to complete. Participants self-identified as speaking fluent English, being 18 or over, and being willing to read about acts/behaviours involving violent or sexual content. Recruitment was conducted with the aim of gathering over 170 responses over the course of the academic year. The resultant sample (N = 228) is comprised of first- and second-year Psychology students studying in the United Kingdom, and mostly female (n = 187), with a mean age of 20 (SD = 4.36). The research for this project was submitted for ethics consideration under the reference PSYC 16/ 243 in the Department of Psychology and was approved under the procedures of the University of Roehampton’s Ethics Committee on 05.10.16.

4.2.3. Materials

The study uses a total of 64 short vignette-scenarios depicting a variety of transgressions typically considered either morally harmful or morally impure. The majority of scenarios (47) are drawn from existing research, with most (37) having been created and validated specifically for use with Moral Foundations Theory approaches (Clifford et al., 2015), and the other 10 scenarios being 'naturally generated' by participants in Gray and Keeney's (2015a) research. The remaining 17 scenarios were created for this study, being designed with the aim of eliciting 'impurity' concerns as described under MFT and Constructive Sentimentalism (Prinz, 2009). These are divided between two scenario sets, such that each set is comprised of 16 scenarios depicting harm-based transgressions (emotional / animal focused / physical) and 16 scenarios depicting impurity-based transgressions. The split of scenario sources between sets is shown in Table 4.1.

Table 4.1. Source of scenario by moral domain and scenario set.

Domain / Set	Set 1 (32)	Set 2 (32)
Harm (16)	14 MFV + 2 G&K	13 MFV + 3 G&K
Impurity (16)	10 MFV + 5 G&K + 1 RG	16 Researcher Generated

MFV = Moral Foundations Vignettes (Clifford et al., 2015)

G&K = 'Naturalistic' scenarios drawn from Study 2 in Gray and Keeney (2015a)

Participants were asked 6 questions about each of the 32 scenarios they were shown. 1. Is this act/behaviour morally wrong? 2. Does the person engaging in this act/behaviour have poor moral character? 3. How atypical [i.e., weird, strange, unusual/ bizarre, odd] is this act/behaviour? 4. Is this act/behaviour angering? & 5. Is this act/behaviour disgusting? - each asked for responses on a seven-point scale, with the end points labelled 'Not at all' and 'Extremely'. The sixth question, 6. This act/behaviour is.... - examined whether the event was, on balance, considered to be more harmful or

impure. This was rated on a bipolar scale, with the left end (1) labelled 'more harmful [i.e., involving physical and/or emotional suffering] than impure' and the right end (7) labelled 'more impure [i.e., involving degradation, indecency, dirtiness] than harmful'

In addition to these measures, participants were asked to complete items drawn from the harm and impurity sub-scales of both the Moral Foundations Questionnaire (MFQ; Graham et al., 2011) and the Moral Foundations Sacredness Scale (MFSS; Graham & Haidt, 2012). The former examines the relative emphasis placed on each Moral Foundation, and the latter asks how much money it would require for an individual to undertake actions generally considered violations of harm or purity. These scales were included to account for potential moderators, although were not subsequently included during analysis.

4.2.4. Procedure

Participants took part in the study by completing an online questionnaire delivered via Qualtrics (Provo, UT, USA). Participants were asked to provide informed consent, demographic information (age/sex), an ID code, and were randomly assigned to one of the two scenario sets. Scenarios were presented in a random order and questions about the scenarios were always presented in the same order. All randomisation was controlled via Qualtrics, and participants were blind as to which scenario set was presented. Participants were also asked to respond to the MFQ and MFSS items after rating the scenarios, with items presented in list-wise order alternating between harm and impurity items. Participants were presented with a debrief following completion of the study.

4.2.5. Pre-registration

A priori power calculations, statements of hypotheses, planned analyses, and a full list of scenarios are available via the pre-registration site - https://osf.io/kkfms/

4.2.6. Limited Reporting of Method and Results

The method and results reported here are limited to the first study detailed in the pre-registration document, others are not reported in full due to various methodological issues. In particular, the second study failed to recruit enough participants to provide acceptable power for analysis (N=82), and further undermined plans to analyse response differences within participants across both studies. It is also questionable as to whether simply re-phrasing the question to ask about approval for the second study provides a suitable test for second-order morals. Only pre-scripted analysis was run for the second study, with preliminary results following largely the same patterns reported below for first-order morals. Given the level of similarity, this could suggest either that first and second order morals operate similarly, such that 'approving of this action is...' is answered in a very similar way to just 'this action is...', or that the second study is simply redundant with the first study, such that participants interpret these questions in the same way. Further investigation, using a different methodology, would be required to better examine potential differences between first and second order moral judgements. This may require development of more detailed stimuli, whereby the approval is built into vignettes depicting actions with 'reactive witnesses'. These would contain both first order material, such that the perpetrator performs an action, and second order material, such that a witness observes this event and reacts approvingly/neutrally/disapprovingly towards it - with questions being asked about both perpetrator and witness.

4.3. Results

4.3.1. Response validity checks

One case was removed as the result of a participant completing the study twice, with the latter completion being removed. Five further cases were removed from the data set according to the pre-set exclusion criterion for improbably quick responses - under 10 minutes to answer ~225 items, which is equivalent to ~2 seconds per item once factoring in page loading. This left 222 responses, with 113 covering the first set of scenarios, and 109 covering the second set.

4.3.2. Data processing

Data was initially processed following the pre-scripted data processing syntax provided as part of pre-registration. This calculated means for each of the six dependent variables across both harm and impurity scenarios for each group, along with mean scores for both the harm and purity subscales of the Moral Foundations Questionnaire and Moral Foundations Sacredness Scale. Subsequent processing transformed data to nest scenarios within participants, allowing for analyses approximating the multi-level model approach used by Gray and Keeney (2015a).

4.3.3. Stimulus Validity

Means for the harm vs. impurity 'balance' item were inspected for each scenario to examine stimulus validity. Four scenarios, all included as instances of impurity violations, were rated as being more harmful than impure (i.e., $M < 4$). The majority of researcher generated scenarios successfully depicted instances of impurity rather than instances of harm, although two of these (#32 - $M = 3.88$, $SD = 1.72$; #57 - $M = 2.97$, $SD = 1.69$) produced scores below the scale midpoint. In contrast, the two other failing scenarios were the 'naturalistic' ones generated in Gray and Keeney's (2015a) study which are mentioned as depicting mixed moral content. Adultery (#27) scored just below the midpoint ($M = 3.74$, $SD = 2.28$), and rape (#30) was typically perceived as being more harmful than impure ($M = 2.85$, $SD = 2.3$). Balance ratings matched conceptual classifications for all other scenarios on average, although notably - every scenario received ratings across the full range of the item scale.

Averaging across balance scores for all participants and scenarios, the median score was 4, with a mean of 4.08 ($SD = 2.19$). The response pattern shows participants tended to favour the middle and end points of the scale overall (*MODE* = 7), with a slight preference for using intermediary points nearer the scale ends over those nearer the middle (Figure 4.1). Collapsing to average across participants, ratings also showed some degree of variability ($SD = .76$) regarding whether participants perceived their set of scenarios as being more harmful or more impure. Although around half the sample produced scores falling within the standard error of the mean (+/-0.5), a small selection of participants seemed to tend either towards perceiving all events as being more harmful (Mean < 3, $n = 13$, ~6%), or as being more impure (Mean > 5, $n = 25$, ~11%), despite only 10 participants having fairly restrictive ratings (*RANGE* < 5) and the majority making full use of the 'balance' scale (*MODE RANGE* = 6).

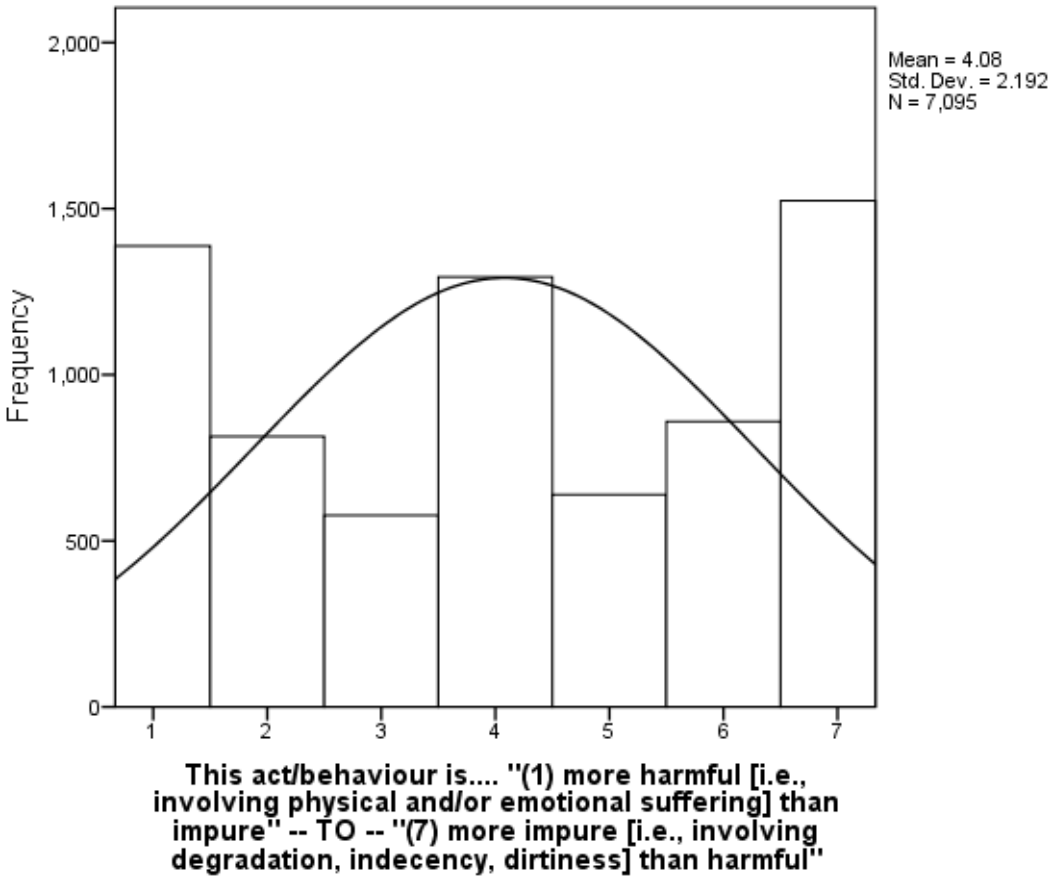


Figure 4.1. Response distribution pattern of all participants 'balance' ratings across all scenarios.

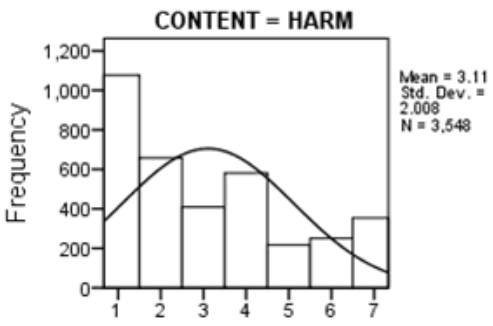


Figure 4.1a. 'Harm-rated' scenarios.

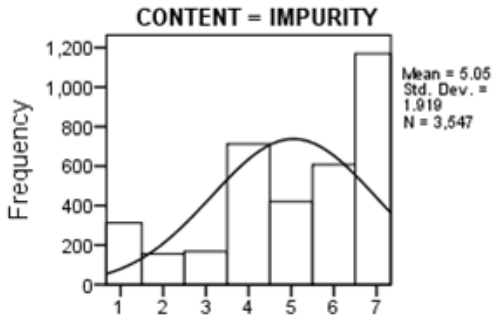


Figure 4.1b. 'Impurity-rated' scenarios.

4.3.4. Deviations from pre-registration

The primary analysis plan was executed in line with the pre-scripted analysis syntax provided at time of pre-registration. However, issues regarding stimulus validity were not fully taken into account during pre-registration and this created some degrees of freedom within the planned analyses. Proceeding as scripted leaves failing scenarios in place, such that some events which participants generally perceived as being more harmful would have been scored on the impurity factor - which is problematic for validity. In contrast, removing these scenarios unbalances the number of scenarios scored for each factor among groups - which may be problematic for ANOVA based analyses which assume equal groups. Alternatively, failing scenarios could be reclassified as instances of harm based on participant responses - which has the benefit of using all scenario data collected whilst being favourable toward participant intuitions.

There is also a question as to whether all failing scenarios should be removed, as whilst there is no attempt to claim that failing researcher generated scenarios validly depict instances of impurity, Gray and Keeney (2015a, p.862) claim scenarios depicting adultery and rape are 'naturalistically' valid violations of purity. However, the response patterns for the failing scenarios suggest selective inclusion of these would be problematic. Results showed both researcher generated scenarios scored below the factor-group mean across all dependent variables, whereas the naturalistic scenarios both scored above the factor-group means across these variables - with the exception of adultery, which was rated as less atypical than the average transgression. Selectively retaining only ‘naturalistic’ scenarios whilst comparing scenario sets would provide favour to whichever moral content factor they were associated with, although notably, treating these as harm scenarios would seem more favourable to Gray and Keeney’s (2015a) claims than treating them as impurity scenarios.

Taking stimulus validity and these potential degrees of freedom into account, analyses were initially run using ANOVA across three data variations – as per pre-registration, removing failed stimuli, and factoring stimuli based on mean participant responses. Subsequent analyses were run using multi-level models so as to approximate Gray and Keeney's (2015) approach. The multi-level approach nests scenarios within participants, such that each scenario is treated as a separate measure, and greatly increases statistical power. This also helps address any concerns regarding equal group size assumptions when using ANOVA - although multi-level analyses are also run across the three data variations. Thus, whilst results are reported for this study in line with the pre-scripted analyses, they are further reported in line with a pre-planned analysis option to approximate Gray and Keeney's approach using multi-level models, and to provide what is arguably a better analysis method for the data - multi-level models with scenarios allocated to harm and impurity factors based on average participant ratings. Furthermore, given results showed some differences across dependent variables between stimulus sets, analyses were performed to investigate these differences with regard to scenario 'source' – again based on Gray and Keeney's (2015a) approach.

4.3.5. ANOVA-based analyses

MANOVA's were run to examine the effects of scenario content and scenario set across all six of the dependent variables. Analyses were performed in triplicate to account for the degrees of freedom created by stimulus validity issues. Results here are thus reported for when scenario content is treated as initially allocated (so includes 'invalid' content), when 'invalid' scenarios are removed (so with the four scenarios allocated for impurity, but rated as harmful, removed for analysis), and when content is treated as typically rated by participants (includes all scenarios).

Main effects for content were found across all dependent measures (see Tables 4.2 and 4.3). As expected, harm scenarios were rated as more harmful and impurity scenarios were rated as more impure overall. Also as expected, impurity scenarios were rated as more atypical than harm scenarios, less angering, more disgusting, and more reflective of poor moral character (except when scenarios were treated as rated). Harm scenarios were rated as more morally wrong than impurity scenarios, but this main effect for wrongness was only found when scenarios were treated as initially allocated - harm and impurity scenarios were comparable over ratings of wrongness when scenarios were treated as rated.

Table 4.2. Main effects of content across three data treatment possibilities - ANOVA

Content	As allocated			Scenarios Removed			As rated		
Measure	*F*(1,220)	*p*	*ηp2*	*F*(1,220)	*p*	*ηp2*	*F*(1,220)	*p*	*ηp2*
'Balance'	507.79	<.001	.698	549.23	<.001	.714	563.23	<.001	.719
Wrongness	9.78	.002	.043	5.52	.02*	.024	1.57	*n.s.*	.007
Character	15.45	<.001	.066	10.19	.002	.044	4.3	.039*	.019
Atypicality	183.32	<.001	.455	246.34	<.001	.528	269.07	<.001	.55
Anger	118.24	<.001	.35	127.41	<.001	.367	117.76	<.001	.349
Disgust	6.54	.011	.029	14.86	<.001	.143	26.25	<.001	.107

*non-significant when applying Bonferroni correction for number of analyses run.

However, between-subject effects for scenario set were consistently present for balance, and anger (see Tables 4.4 and 4.5). These showed scenarios in Set 1 were rated as relating more to impurity than those in Set 2, whereas scenarios in Set 2 were rated as more angering than those in Set 1 on average.

Table 4.3. Mean ratings by Scenario Type for dependent variables across three data treatment possibilities

Content	As allocated		Scenarios Removed		As rated	
Scenarios	Harm	Purity	Harm	Purity	Harm	Purity
'Balance'	3.105	5.050	3.105	5.309	3.127	5.309
Wrongness	5.670	5.520	5.670	5.547	5.611	5.547
Character	5.433	5.238	5.433	5.260	5.369	5.260
Atypicality	4.616	5.301	4.616	5.475	4.553	5.427
Anger	5.530	4.873	5.530	4.781	5.485	4.781
Disgust	5.276	5.408	5.276	5.490	5.216	5.490

Table 4.4. Between-subject effects of Scenario Set across three data treatment possibilities

Content	As allocated			Scenarios Removed			As rated		
Measure	*F*(1,220)	*p*	*ηp2*	*F*(1,220)	*p*	*ηp2*	*F*(1,220)	*p*	*ηp2*
'Balance'	5.29	.022*	.023	13.85	<.001	.058	16.18	<.001	.068
Anger	4.24	.041*	.019	13.55	.02*	.059	11.19	<.001	.048

*non-significant when applying Bonferroni correction for number of analyses run.

Table 4.5. Mean ratings by Scenario Set for dependent variables across three data treatment possibilities

Content	As allocated		Scenarios Removed		As rated	
Scenarios	Set 1	Set 2	Set 1	Set 2	Set 1	Set 2
'Balance'	4.193	3.962	4.389	4.025	4.413	4.023
Wrongness	5.620	5.570	5.572	5.645	5.577	5.580
Character	5.363	5.308	5.306	5.388	5.311	5.318
Atypicality	4.945	4.972	5.072	5.020	5.025	4.955
Anger	5.098	5.304	4.958	5.353	4.959	5.307
Disgust	5.413	5.272	5.415	5.351	5.420	5.286

Interaction effects between content and scenario set (see Tables 4.6 and 4.7) were found for both wrongness and character, showing impure content in Set 2 was rated as both more wrong and more diagnostic of poor moral character than the harm content in the same set, whilst harm content in Set 1 was rated as both more wrong and more diagnostic of poor moral character than the impure content in the same set - an opposing pattern. An interaction for atypicality was found when scenarios were treated as allocated, with harm scenarios in Set 1 being rated as more atypical than those in Set 2, whilst impure scenarios in Set 2 were rated as more atypical than those in Set 1. However, this effect disappeared when scenarios were removed or treated as rated. A consistent interaction was also found for anger, with harm scenarios being rated as more angering than impurity scenarios in Set 1, although scenarios in Set 2 were rated as similarly angering on average. There was no interaction effect for disgust when scenarios were treated as allocated, although an effect emerged when scenarios were treated as rated. Harm scenarios in Set 1 were rated as more disgusting than those in Set 2, whilst impure scenarios in Set 2 were rated as more disgusting than those in Set 1. Finally, no interaction effects were found for 'balance'.

Table 4.6. Interaction effects (scenario type x scenario set) across three data treatment possibilities

Content	As allocated			Scenarios Removed			As rated		
Measure	*F*(1,220)	*P*	*ηp2*	*F*(1,220)	*P*	*ηp2*	*F*(1,220)	*p*	*ηp2*
'Balance'	.448	*n.s.*	.002	549.23	.044*	.018	3.22	*n.s.*	.014
Wrongness	16.56	<.001	.07	36.59	<.001	.143	56.57	<.001	.205
Character	21.38	<.001	.089	45.85	<.001	.172	70.46	<.001	.243
Atypicality	13.19	<.001	.057	3.71	*n.s.*	.017	.264	*n.s.*	.001
Anger	82.29	<.001	.272	123.11	<.001	.359	145.48	<.001	.398
Disgust	1.01	*n.s.*	.005	5.43	.021*	.024	13.8	<.001	.059

*non-significant when applying Bonferroni correction for number of analyses run.

Table 4.7. Mean ratings for dependent variables (scenario type x scenario set) across three data treatment possibilities

Content	As allocated		Scenarios Removed		As rated	
Scenarios	Set 1 Harm	Set 2 Harm	Set 1 Harm	Set 2 Harm	Set 1 Harm	Set 2 Harm
	Set 1 Impurity	Set 2 Impurity	Set 1 Impurity	Set 2 Impurity	Set 1 Impurity	Set 2 Impurity
'Balance'	3.192	3.018	3.192	3.018	3.239	3.015
	5.194	4.905	5.586	5.032	5.586	5.032
Wrongness	5.793	5.548	5.793	5.548	5.804	5.418
	5.447	5.593	5.351	5.743	5.351	5.743
Character	5.575	5.291	5.575	5.291	5.586	5.152
	5.150	5.325	5.036	5.484	5.036	5.484
Atypicality	4.695	4.538	4.695	4.538	4.601	4.504
	5.196	5.406	5.448	5.502	5.448	5.406
Anger	5.701	5.359	5.701	5.359	5.702	5.268
	4.496	5.250	4.216	5.346	4.216	5.346
Disgust	5.373	5.180	5.373	5.180	5.382	5.050
	5.453	5.363	5.457	5.523	5.457	5.523

4.3.6. Multi-level model analyses

Other research in the area has advocated using some form of mixed models over ANOVA so as to better account for variability within content across scenarios (e.g., Gray & Keeney, 2015a). These models are also better at handling both missing data and unequal group sizes (Field, 2013), making these better suited for analyses where 'failed' impurity stimuli are removed, or treated as rated by participants. Given this is the method used by Gray and Keeney (2015a), data were processed to approximate their reported analysis. Scenarios were nested within participants and a mixed-models approach was used to examine fixed main effects of scenario content and scenario set, with simple slopes analysis used to examine any interactions. For transparency, results

are reported in line with the previous analysis eventualities: with scenarios treated as allocated, with 'failed' impurity scenarios removed, and with scenarios treated as rated. Summary findings for each dependent measure are provided below, with statistics relating to main effects for content and scenario set provided in Tables 4.8 and 4.9 respectively. Statistics for simple slopes analyses are provided in Tables 4.11 and 4.12, unpacking the interactions reported in Table 4.10. Ratings for all scenarios on wrongness and atypicality are plotted in Figure 4.2.

4.3.6.1. Balance

Consistent main effects show harm scenarios were rated more towards harm, and impurity scenarios more towards impurity. Scenarios in Set 1 were rated further towards impurity than those in Set 2, and this is reflected in interaction effects showing the difference between harm and impurity ratings was greater in Set 1 than in Set 2.

4.3.6.2. Wrongness

A consistent lack of main effect for scenario set shows these were comparable over wrongness ratings. Main effects for content are present when some scenarios are conceptually misclassified, and when removing these scenarios from the analysis. However, treating scenario content as rated by participants eliminates this main effect, showing harm and impurity scenarios rated as comparably wrong. This suggests any effect of scenario content on wrongness ratings between harm and impurity is likely to be small. Also, a consistent interaction effect showing harm being rated as more wrong than impurity in Set 1, but impurity as more wrong than harm in Set 2, suggests any effects of scenario content on ratings of wrongness are likely to depend more on other aspects of the stimulus rather than simple domain classification.

4.3.6.3. Character

There was no main effect for scenario set, except for when scenarios were removed - and this would be non-significant if applying Bonferroni adjustments across analysis eventualities. The scenario sets are thus comparable in terms of overall inference about moral character. There was a consistent main effect for Content, with harm being considered more indicative of poor moral character overall. However, an interaction effect shows harm being rated as more indicative of poor moral character than impurity in Set 1, although impurity was more indicative than harm in Set 2.

4.3.6.4. Atypicality

A consistent lack of a main effect for Set shows these are comparable in terms of overall atypicality, and a consistent main effect for content shows impurity scenarios were rated as more atypical (i.e., weirder) than harm scenarios. Any interaction effects are comparatively weak, being absent when scenarios are transferred, and being non-significant if applying Bonferroni adjustments across analysis eventualities when scenarios are removed. When present, interaction effects indicate harm scenarios in Set 1 were rated as more atypical than those in Set 2, although impurity scenarios in Set 1 were rated as slightly less atypical than those in Set 2.

4.3.6.5. Anger

Consistent main effects were shown for content, with harm scenarios rated as more angering than impurity scenarios. There was also a consistent main effect for Set,

with Set 2 being more angering overall. A consistent interaction effect shows harm and impurity scenarios followed the main effect where harm was more angering in Set 1, although scenarios were rated as similarly angering in Set 2.

4.3.6.6. Disgust

Consistent main effects were shown for content, with impurity being rated as more disgusting than harm. There was an inconsistent main effect for set, with Set 1 scenarios rated as being more disgusting than Set 2 when scenarios were treated as allocated or as rated. An inconsistent interaction was present following removal or reallocation, which showed impurity scenarios were similarly disgusting between Sets 1 and 2, and harm and impurity ratings were similarly disgusting in Set 1, whereas Set 2 harm scenarios were less disgusting than Set 2 impurity scenarios.

4.3.7. Comparing Scenario Sources

The dependent variables were also examined by scenario source to further investigate claims that MFT impurity scenarios are weirder and less severe than other scenarios - and that 'naturalistic' impurity scenarios are less weird and more severe in comparison. The two researcher generated items that failed the stimulus evaluation check (Set1-#16, Set2-#9) were excluded from this analysis, whilst retaining all Gray and Keeney's impurity scenarios. This analysis therefore compares 10 MFT impurity scenarios, all 5 'naturalistic' impurity scenarios, and 15 researcher generated (RG) impurity scenarios; comparisons with this last category are thus between groups, whilst comparisons of MFT and 'naturalistic' ratings are within participants. A mixed-models approach was used to examine fixed effects of scenario source. These results are summarised below, with statistics for these comparisons reported in Table 4.13.

'Naturalistic' impurity scenarios were rated as being less impure, on balance, than both MFT and RG impurity scenarios; RG impurity scenarios were also less impure, on balance, than MFT impurity scenarios - and MFT impurity scenarios remain the most impure even if only the valid three ‘naturalistic’ scenarios are included in the analysis. 'Naturalistic' impurity scenarios were rated as less morally wrong than MFT impurity scenarios and RG impurity scenarios – an effect more pronounced if limiting to the valid three, whilst MFT and RG impurity scenarios were rated as similarly wrong. 'Naturalistic' impurity scenarios were rated as similarly indicative of poor moral character to MFT impurity scenarios, but less indicative than RG impurity scenarios – and these were rated as more indicative of poor moral character than MFT impurity scenarios. 'Naturalistic' impurity scenarios were also rated as less atypical than both MFT and RG impurity scenarios, whilst RG impurity scenarios were less atypical than MFT impurity scenarios. 'Naturalistic' impurity scenarios were rated as similarly angering to MFT impurity scenarios, but both of these were rated as less angering than RG impurity scenarios. 'Naturalistic' impurity scenarios were rated as less disgusting than both MFT and RG impurity scenarios, whilst MFT impurity scenarios were more disgusting than RG impurity scenarios.

In summary, the 'naturalistic' impurity scenarios (#27-31) generated by participants in Gray and Keeney's (2015) research were rated as less wrong, less atypical, and less disgusting, than scenarios from other sources, but were comparable with MFT impurity scenarios (#17-26) with regard to anger and judgements about moral character. However, these were also the most balanced scenarios overall, being rated as closest to the mid-point of the harm-impurity item scale. In contrast, MFT impurity scenarios were the most atypical, and most disgusting, and also most skewed towards being rated as impure; whilst researcher generated impurity scenarios (#49-64, not including #57) were the rated as being most indicative of moral character and most angering.

Table 4.8. Main effects of content across three data treatment possibilities – multi-level models

Scenario	As allocated			Scenarios Removed			As rated		
Measure	*b*	*β*	*t* *p*	*b*	*B*	*t* *p*	*b*	*B*	*t* *p*
'Balance'	-1.945	-.887	(7095) = -41.38 <.0005	-2.205	-1.101	(6649) = -47.93 <.0005	-2.182	-.996	(7095) = -47.43 <.0005
Wrong	.150	.093	(7104) = 3.95 .001	.124	.079	(6656) = 3.18 .001	.065	.04	(7104) = 1.71 *n.s.*
Character	.195	.113	(7104) = 4.76 <.0005	.173	.102	(6656) = 4.22 <.0005	.109	.063	(7104) = 2.66 .008
Atypical	-.685	-.358	(7099) = -15.22 <.0005	-.86	-.456	(6651) = -19.11 <.0005	-.923	-.482	(7099) = -20.51 <.0005
Anger	.657	.355	(7102) = 15.28 <.0005	.749	.407	(6654) = 17.41 <.0005	.704	.381	(7102) = 16.76 <.0005
Disgust	-.132	-.074	(7101) = -3.14 .002	-.215	-.124	(6653) = -5 <.0005	-.276	-.154	(7101) = -6.42 <.0005

Positive scores ‘harm > impurity’, negative scores ‘impurity > harm’.

‘Balance’ = difference (~2 scale points).

Table 4.9. Main effects of scenario set across three data treatment possibilities

Scenario	As allocated			Scenarios Removed			As rated		
Measure	*b*	*β*	*t* *p*	*b*	*B*	*t* *p*	*b*	*B*	*t* *p*
'Balance'	.232	.106	(7095) = 4.94 <.0005	.364	.166	(6649) = 7.52 <.0005	.389	.178	(7095) = 8.46 <.0005
Anger	-.206	-.111	(7102) = -4.79 <.0005	-.394	-.214	(6654) = -9.16 <.0005	-.348	-.188	(7102) = -8.29 <.0005
Disgust	.142	.08	(7101) = -3.38 .001	.065	.037	(6653) = -1.51 *n.s.*	.135	.075	(7101) = 3.14 .002

+ scores 'Set 1 > Set 2', - scores 'Set 1 < Set 2'. 'Balance' shows Set 1 is more impure.

Table 4.10. Interaction effects across three data treatment possibilities

Scenario	As allocated			Scenarios Removed			As rated		
Measure	*b*	*β*	*t* *p*	*b*	*B*	*t* *p*	*b*	*B*	*t* *p*
'Balance'	-.117	-.074	(7095) = -1.26 *n.s*	-.383	-.175	(6649) = -4.12 <.0005	-.333	-.152	(7095) = -3.65 <.0005
Wrong	.392	.242	(7104) = 5.19 <.0005	.637	.404	(6656) = 8.27 <.0005	.778	.481	(7104) = 10.12 <.0005
Character	.459	.266	(7104) = 5.62 <.0005	.773	.434	(6656) = 8.89 <.0005	.883	.511	(7104) = 10.75 <.0005
Atypical	.368	.192	(7099) = 4.11 <.0005	.212	.112	(6651) = 2.35 .019*	.152	.079	(7099) = 1.7 *n.s.*
Anger	1.1	.593	(7102) = 12.86 <.0005	1.47	.801	(6654) = 17.04 <.0005	1.56	.846	(7102) = 18.43 <.0005
Disgust	.101	.057	(7101) = 1.19 *n.s*	.255	.147	(6653) = 2.99 .003	.395	.221	(7101) = 4.64 <.0005

*non-significant when applying Bonferroni correction for number of analyses run

Table 4.11. Simple slopes comparisons of scenario sets across three data treatment possibilities

Scenario	As allocated			Scenarios Removed			As rated		
Measure	*b*	*β*	*t* *p*	*b*	*β*	*t* *p*	*b*	*β*	*t* *p*
'Balance' (Harm)	-	-	- -	.173	.079	(6649) = 2.72 *p* = .006	.223	.102	(7095) = 3.7 *p* <.001
'Balance' (Impurity)	-	-	- -	.556	.253	(6649) = 8.17 *p* <.001	.556	.253	(7095) = 8.12 *p* <.001
Wrong (Harm)	.246	.152	(7104) = 4.54 *p* <.001	.246	.156	(6656) = 4.67 *p* <.001	.386	.238	(7104) = 7.59 *p* <.001
Wrong (Impurity)	-.146	-.09	(7104) = -2.7 *p* = .007	-.392	-.249	(6656) = -6.96 *p* <.001	-.392	-.242	(7104) = -6.79 *p* <.001
Character (Harm)	.285	.165	(7104) = 4.92 *p* <.001	.285	.168	(6656) = 5.06 *p* <.001	.434	.251	(7104) =7.99 *p* <.001
Character (Impurity)	-.175	-.101	(7104) = -3.02 *p* = .003	-.448	-.265	(6656) = 7.44 *p* <.001	-.448	-.259	(7104) = -7.28 *p* <.001
Atypical (Harm)	.157	.082	(7099) = 2.49 *p* = .013*	.157	.084	(6651) = 2.56 *p* = .011	-	-	- -
Atypical (Impurity)	-.21	-.101	(7099) = -3.33 *p* = 0.01	-.054	-.029	(6651) = -.822 *n.s*	-	-	- -
Anger (Harm)	.342	.185	(7102) = 5.68 *p* <.001	.342	.186	(6654) = 5.81 *p* <.001	.435	.235	(7102) = 7.74 *p* <.001
Anger (Impurity)	-.754	-.408	(7102) = -12.51 *p* <.001	-1.13	-.615	(6654) = -17.89 *p* <.001	-1.13	-.611	(7102) = -17.74 *p* <.001
Disgust (Harm)	-	-	- -	.193	.111	(6653) = 3.31 *p* = .001	.332	.186	(7101) = 5.9 *p* <.001
Disgust (Impurity)	-	-	- -	-.063	-.036	(6653) = -1.01 *n.s.*	-.063	-.035	(7101) = -.984 *n.s.*

Positive scores 'Set 1 > Set 2', negative scores 'Set 1 < Set 2'. 'Balance' shows both harm and impurity scenarios in Set 1 were rated as more impure than their counterparts in Set 2.

*non-significant when applying Bonferroni correction for number of analyses run.

Table 4.12. Simple slopes comparisons of content across three data treatment possibilities

Scenario	As allocated			Scenarios Removed			As rated		
Measure	*b*	*β*	*t* *p*	*b*	*β*	*t* *p*	*b*	*β*	*t* *p*
'Balance' (Set 1)	-	-	- -	-2.4	1.09	(6649) = -36.06 *p* <.001	-2.35	-1.07	(7095) = -36.45 *p* <.001
'Balance' (Set 2)	-	-	- -	-2.01	-.919	(6649) = -30.95 *p* <.001	-2.02	-.92	(7095) = -31.25 *p* <.001
Wrong (Set 1)	.346	.214	(7104) = 6.46 *p* <.001	.443	.281	(6656) = 8.04 *p* <.001	.454	.281	(7104) = 8.35 *p* <.001
Wrong (Set 2)	-.045	-.028	(7104) = -.83 *n.s.*	-.195	-.124	(6656) = -3.61 *p* <.001	-.324	-.201	(7104) = -5.96 *p* <.001
Character (Set 1)	.425	.246	(7104) = 7.42 *p* <.001	.539	.319	(6656) = 9.15 *p* <.001	.55	.319	(7104) = 9.49 *p* <.001
Character (Set 2)	-.034	-.02	(7104) = -.59 n.s.	-.194	-.115	(6656) = -3.36 *p* = .001	-.332	-.192	(7104) = -5.71 *p* <.001
Atypical (Set 1)	-.502	-.262	(7099) = -8.01 *p* <.001	-.754	-.4	(6651) = -11.7 *p* <.001	-	-	- -
Atypical (Set 2)	-.869	-.454	(7099) = -13,63 *p* <.001	-.965	-.512	(6651) = -15.29 *p* <.001	-	-	- -
Anger (Set 1)	1.21	.652	(7102) = 20.18 *p* <.001	1.49	.808	(6654) = 24.07 *p* <.001	1.49	.804	(7102) = 24.77 *p* <.001
Anger (Set 2)	.109	.059	(7102) = 1.79 *n.s.*	.013	.007	(6654) = .211 *n.s.*	-.079	-.042	(7102) = -1.31 *n.s.*
Disgust (Set 1)	-	-	- -	-.087	-.05	(6653) = -1.43 *n.s.*	-.078	-.044	(7101) = -1.3 *n.s.*
Disgust (Set 2)	-	-	- -	-.343	-.197	(6653) = -5.74 *p* <.001	-.473	-.265	(7101) = -7.86 *p* <.001

+ scores 'harm > impurity', - scores 'impurity > harm'. 'Balance' = diff. harm-impurity

Table 4.13. Comparisons of scenario sources for specified impurity scenarios

Sources	Naturalistic vs MFT			Naturalistic vs RG			RG vs MFT		
Measure	*b*	*β*	*t* *p*	*b*	*β*	*t* *p*	*b*	*B*	*t* *p*
'Balance'	-1.17	-5.35	(3328) = -12.31 <.001	-.532	-.243	(3328) = -5.94 <.001	-.642	-.293	(3328) = -9.03 <.001
Wrong	-.319	-.197	(3330) = -3.71 <.001	-.435	-.269	(3330) = -5.36 .001	.116	.072	(3330) = -1.8 *n.s.*
Character	-.201	-.116	(3330) = -2.21 *n.s.*	-.399	-.231	(3330) = -4.62 <.001	.119	.115	(3330) = 2.89 .011
Atypical	-1.85	-.966	(3328) = -20.93 <.001	-1.41	-.736	(3328) = -16.84 <.001	-.44	-.23	(3328) = -6.63 <.001
Anger	.057	.031	(3328) = .56 *n.s.*	-.805	-.435	(3328) = -8.43 <.001	.861	.466	(3328) = 11.37 <.001
Disgust	-.68	-.381	(3328) = -7.64 <.001	-.399	-.223	(3328) = -4.71 <.001	-.281	-.158	(3328) = -4.19 <.001

Impurity. MFT = Scenarios #17-26, Naturalistic = Scenarios #27-32, RG = Scenarios #49-64 (excluding #57). Positive scores 'Source 1 > Source 2', Negative Scores 'Source 1 < Source 2'. Negative 'balance' scores = more central/less impure ratings.

4.3.8. Correlation analyses

Correlations between variables were examined at both response and scenario level across all scenarios - these are detailed in Tables 4.14 and 4.15 respectively. Results suggest measurements of moral wrongness and moral character may be redundant with each other, achieving both near perfect correlation between themselves at scenario level, and highly comparable correlations with other variables at both analysis levels. The results also show a substantial correlation (~.6) between wrongness

and atypicality, in contrast to Gray and Keeney (2015a) who report no such correlation between severity and weirdness. Atypicality was also correlated with disgust to a greater extent than anger, although correlations between atypicality and content or balance (i.e., impurity) were only moderate in size. However, there was no significant correlation between content or balance and disgust at scenario level, whereas there was a moderate correlation between these variables and anger. In contrast, at response level, disgust was better associated with balance (i.e., participant classified impurity) than anger, whereas anger was better associated with content (i.e., researcher classified harm) than disgust - although all these correlations are relatively weak, and the pattern of results may be distorted by differences regarding anger and disgust between scenario sets.

Table 4.14. Hierarchical Correlations (All Responses)

N = 7093-7104		***Mean***	***SD***	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1.	Scenario Set	.49	.5	.063	-.015*	-.016*	.007*	.056	-.04	-.053
2.	Scenario Content	.437	.5		-.048	-.058	.178	-.18	.037	.444
3.	Wrong	5.6	1.62			.849	.549	.719	.756	.048
4.	Character	5.34	1.73				.555	.717	.764	.061
5.	Atypicality	4.96	1.92					.426	.597	.221
6.	Anger	5.2	1.85						.711	-.068
7.	Disgust	5.34	1.79							.138
8.	Balance	4.08	2.19							

*Correlation is NOT significant at the 0.01 level (2-tailed), all other correlations are significant at the 0.01 level. Scenario Content: Harm = '0', Impurity = '1'. Scenario Set: Set 1 = '0', Set 2 = '1'.

Table 4.15. Zero-order Correlations (All Scenarios)

N = 64		*Mean (SD)*	2	3	4	5	6	7	8
1	Scenario Set	.5 (.5)	.063	-.032	-.032	.013	.11	-.077	-.099
2	Scenario Content	.5 (.5)		-.039	-.062	.462**	-.356**	.144	.916**
3	Wrong	5.6 (.8)			.982**	.674**	.878**	.93**	.012
4	Character	5.34 (.84)				.648**	.885**	.919**	.008
5	Atypicality	4.96 (1)					.433**	.814**	.462**
6	Anger	5.2 (.95)						.757**	-.313*
7	Disgust	5.34 (.94)							.206
8	Balance	4.08 (1.18)							

*Correlation is significant at the 0.05 level (2-tailed). **Correlation is significant at the 0.01 level (2-tailed). Scenario Content: Harm = '0', Impurity = '1'. Scenario Set: Set 1 = '0', Set 2 = '1'.

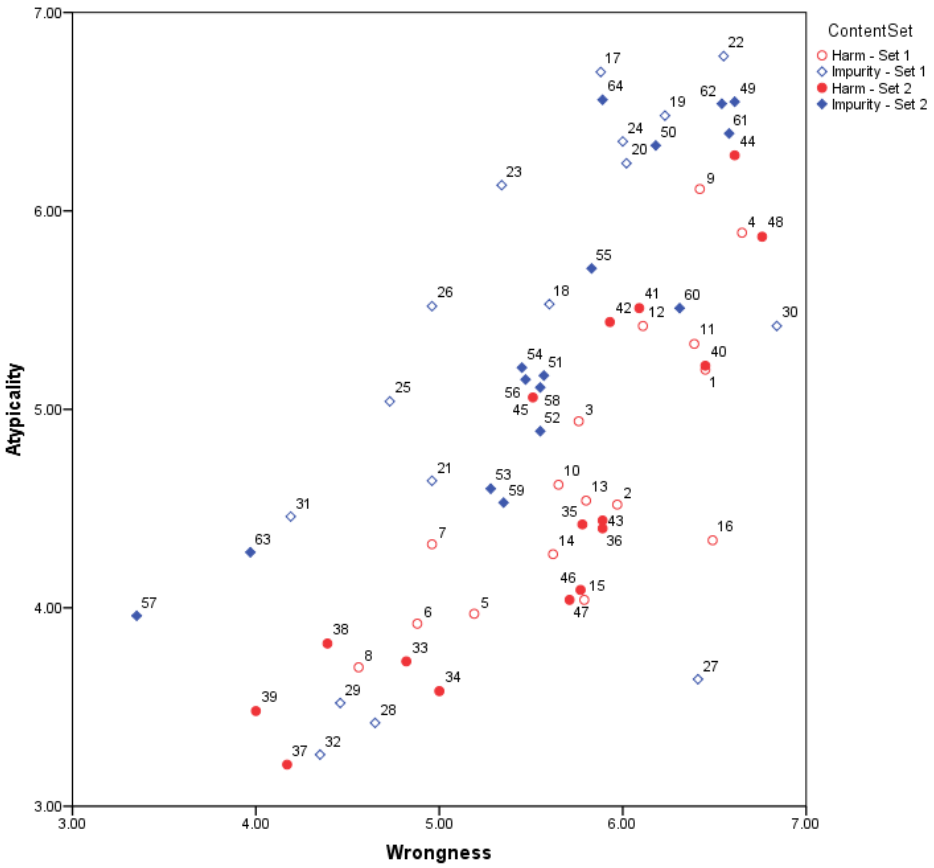


Figure 4.2. All scenarios by 'wrongness' and 'atypicality'

4.3.9. Results Summary

Harm scenarios were rated as more harmful, and impurity scenarios as more impure, with the exception of the four scenarios mentioned – the 'failures' of which can be readily explained, such that H6 can be accepted. Impurity scenarios were rated as more atypical, and generally more disgusting, than harm scenarios – H3 and H5 can be accepted. However, there is an opposing trend across ratings of wrongness and character between scenario sets, where harm scenarios were rated as more morally wrong, and more indicative of poor moral character, in the first scenario set - whereas the second set shows the opposing pattern with impurity scenarios scoring higher than harm scenarios on these variables. The results from both scenario sets are thus opposed as to whether H1 and H2 might be accepted - ratings of wrongness and poor moral character are not seemingly dependent on whether the scenarios depict violations of harm or purity. Furthermore, scenario sets varied across ratings of anger, whereby harm scenarios were more angering than impurity scenarios over the first scenario set, but scenarios depicting both types of content were rated as comparably angering in the second set – limiting support for accepting H4, as the effect was only present in one set of scenarios.

4.4. Discussion

The study aimed to investigate support for claims that both Moral Foundations Theory impurity scenarios, and (by inference) impurity scenarios more generally, are typically weirder and less severe than harm scenarios – such that these 'confounds' might be used to explain MFT's 'Sanctity' foundation in terms of 'weird harms' (Gray & Keeney, 2015a; Schein & Gray, 2018). The study also sought to examine associated claims that impurity is more associated with moral character than harm (Chakroff & Young, 2015); and more generally whether harm and impurity violations elicited comparably greater levels of anger and disgust respectively, such that harm violations characteristically elicit anger, whereas impurity violations characteristically elicit disgust (Graham et al., 2013; Prinz, 2009). The results support Graham's (2015) reply to Gray and Keeney (2015a) – that weirdness is a 'primary feature' of impurity – and generally support Moral Foundations Theory based claims regarding emotion (Graham et al., 2013) over those advanced by the Theory of Dyadic Morality (Schein & Gray, 2018).

4.4.1. The Balance of Harm and Impurity

The results show participants in this study did, on average, reliably classify scenarios in line with theoretical expectations - supporting Graham's (2015) contention that Gray and Keeney's (2015a) measures of harm and impurity are redundant with their measures of wrongness - which is to say they all relate to the same construct (immorality). Harm scenarios were not rated as being more impure than impurity scenarios in this study. Thus, whilst Gray and Keeney's (2015a) findings are replicable (Franchin et al., 2019), their findings seem to be a result of their methodology, as this study shows that participants can generally dissociate judgements of harm from those of impurity when better controls are in place.

The two apparent misclassifications of researcher generated impurity scenarios can be explained with reference to either their highly exploratory nature, in that scenario #32 (challenging religious beliefs) was chosen to make up the numbers in the first scenario set - as the likely placement of this item was the most difficult to anticipate; or in that the refusal of a lifesaving blood transfusion (#57) contains, or at least implies, an (perhaps obvious, particularly in hindsight) instance of physical harm. However, the two seeming misclassifications of the scenarios generated by participants in the second study of Gray and Keeney's (2015a) research are potentially more of an issue, given their claim that these 'naturalistic' scenarios capture the concept of impurity better than custom designed MFT counterparts.

Scenarios about adultery (#27) and rape (#30) were rated as being more about physical and/or emotional suffering than about degradation, indecency, or dirtiness – which is to say these scenarios were rated as generally being more harmful than impure. Considering both scenarios were generated by Gray and Keeney's participants for both harm and impurity concerns, and are acknowledged to contain ‘mixed’ moral content, their failure to be classified as more impure in this study is perhaps unsurprising. Furthermore, the overlap in content seems apparent from the 'balance' ratings for these two scenarios, as their ratings are the most variable (i.e., they have the highest standard deviation) out of all the impurity scenarios; and only one harm scenario, which depicts laughing at a cancer patient (#4), was rated with greater variability.

For every scenario, ratings on the harm vs. impurity ‘balance’ item ranged across all seven points of the scale. The majority of participants also made use of the full range on the 'balance' scale, although around 17% of participants tended to respond either more often or more strongly to one side of the scale than the other - with almost twice as many of these participants tending toward the impurity side. The distribution of

responses further suggests participants were generally fairly sure about whether scenarios were more harmful or impure, with the ends of the scale being selected more often than the combined total of responses favouring the respective sides for both content types. However, the middle scale option was selected more often than might be expected for both types of content, and it is notable that the (relative) opposing end of the balance item was the fifth most selected response to each content type.

Gray and Keeney (2015a) may argue that this provides further evidence that lay moral intuitions do not conform to the content boundaries argued for by modular accounts (i.e., MFT), particularly as most of the scenarios do not contain any obviously 'mixed' content. However, although 'strong modularity' is not a claim made by Moral Foundations Theory, the range of ratings provided both cast doubt on the existence of 'pure' foundation violations, and suggest considerable variability with regard to the moral perception and classification of actions. As such, Gray and Keeney's advice to ensure that stimuli only activate concerns about the foundation in question, in order to demonstrate the independent activation required for foundations to be classified as separate mechanisms, may be an unreasonable request.

For example, Clifford et al. (2015) state that "[e]ach vignette depicts a behavior violating a particular moral foundation and not others". However, their measurement is conducted with the instruction "Why is the action morally wrong? (Select the main reason.)", so the relatively forced choice nature of the question may result in overstating how 'pure' each of the scenarios is with regard to depicting a particular Moral Foundation; and even then, only 1 of the 132 scenarios they used achieved 100% agreement with regard to domain classification. When considered alongside the current results, it seems possible that most, if not all, moral violations might be seen as violating more than one foundation; and even if there is general consensus regarding which foundation has primarily or most saliently been violated for any given action - the case

that an action is immoral could also, or further, be made with reference to other foundations.

However, importantly, the results from this study show that validated MFT scenarios appear to be better associated with impurity than 'naturalistic' scenarios. On average, MFT impurity scenarios received ratings further toward the impurity end of the scale than both researcher generated and 'naturalistic' scenarios, and this holds even if the harm-rated scenarios depicting adultery and rape are removed from scores on the 'balance' item. Furthermore, at best, the three 'naturalistic' violations that were scored as being more impure might be considered marginally more associated with impurity than the three MFT scenarios scoring lowest on the 'balance' measure. This is hardly the basis from which to make a substantive claim that 'naturalistic' scenarios are better measures of impurity, especially since these also limit violation varieties to revolve around sex - whereas validated MFT measures cover a greater range of 'impure' actions, such as cousin marriage and cannibalism.

4.4.2. Atypicality / Weirdness and Wrongness / Severity

The results support Graham's (2015) claim that atypicality is a 'primary feature' of the 'Sanctity/degradation' foundation. Scenarios depicting violations of this foundation were rated as more atypical, on average, than scenarios regarding physical or emotional harm; and of all dependent measures, atypicality ratings had the highest correlation with scenario 'balance' (i.e,, impurity). However, there is some overlap on ratings of atypicality with regard to content. Harm scenarios involving the (likely) death of squirrels (#9) or cats (#44) were rated as being more unusual (6+) than many of the impurity scenarios; and the majority of MFT harm scenarios involving animals also scored highly (5+) in terms of atypicality.

Interestingly, other MFT harm scenarios scoring highly on atypicality were those depicting actions involving various forms of laughing at persons with a disability (#1,40) or cancer (#4), which are classed as instances of emotional harm; and these other scenarios were rated as close to evenly 'balanced' (~3.68), whereas highly atypical animal harm scenarios were rated more towards the harm end of the balance scale (< 3). These response patterns might be explainable with reference to (potentially) 'mixed' moral content, as these seem to contain elements of perceptions which might be argued to align with the 'Sanctity' foundation. Alternatively, these responses might be explained with reference to 'vulnerability', as the victim(s) in these scenarios differ from those in other scenarios. However, this explanation would seem to associate atypicality with elements of moral patiency/victimhood, and may be problematic for Schein and Gray's (2018) claim that atypicality is associated with the extent of the norm violation. Accordingly, more specific investigation of the moral concepts surrounding these highly atypical harms may provide a productive line of future enquiry.

Yet these results, combined with results for wrongness/severity, suggest Gray and Keeney's (2015a) contention - that weirdness and severity act as confounds with regard to scenario content - fails to generalize beyond the 'commonly used' MFT scenarios they investigated. Here, validated MFT impurity scenarios were rated as the most unusual of all the scenarios, yet both MFT harm and impurity scenarios were rated as comparably wrong. As such, although it may be the case that the MFT impurity scenarios used in Gray and Keeney's first study may be weirder, and less severe, than the MFT harm scenarios they used - this trend did not appear across the properly validated MFT scenarios developed by Clifford et al. (2015), despite the MFT impurity scenarios still typically being rated as highly weird in comparison. Furthermore, although the 'naturalistic' impurity scenarios were considered less atypical than either MFT or researcher generated impurity scenarios, they were also considered less wrong. In contrast, mean scores show 'naturalistic' harm scenarios were rated as the most morally

wrong of all harm scenarios, but were largely comparable on ratings on atypicality with MFT harm scenarios.

These findings lend support to Graham's (2015) claim that weirdness and severity are just dimensions along which violations may differ, rather than confounds; and although Graham seems to concur that impurity violations may typically be weirder and less severe, the impurity scenarios used in this study were rated as depicting similar levels of wrongness to the validated harm scenarios. Additionally, wrongness was the variable least correlated with moral content classification and ‘balance’. Thus, it would seem that whilst impurity violations are indeed weird, they are not 'just weird'.

4.4.3. Moral Character

In contrast to Gray and Keeney (2015a, Study 3), who show weirdness has a larger effect on evaluations about character than evaluations about the act, no such relationship is shown in this data. However, the current result might be explained by simple reference to the task, in that those that commit more serious transgressions are thought to have correspondingly severe deficits in their moral character - regardless of the atypicality of the action. The near perfect correlation (.85 to .98) between ratings of wrongness and moral character in this study (Tables 4.11 & 4.12) suggest redundancy within these ratings - particularly as these variables follow within-scenario-set trends as regard to content. As such, the finding that harmful events were rated as being slightly more diagnostic of moral character should be treated with caution, and further, more careful investigation undertaken to examine these potential relationships. This could follow Gray and Keeney's design, although it may be a challenge to generate stimuli which are effectively matched with regard to the action depicted, but allow for weirdness to be varied in a non-extraneous manner.

4.4.4. Anger and Disgust

Consistent with previous research regarding emotions and moral content, harm scenarios were generally rated as more angering than impurity scenarios, and impurity scenarios as generally more disgusting than scenarios depicting harm. Many of the few scenarios which do not adhere to this trend have already been mentioned, either because these conceptually impure scenarios were rated as being more about harm and therefore follow the expected trend of being more angering (#27, #32, #57), or because they were particularly unusual harm scenarios (#9, #44) - which were rated as similarly angering and disgusting. Additionally, the two scenarios rated as most wrong, which involved rape (#30) or murder (#48), were rated as similarly angering and disgusting, as was laughing at a disabled co-worker (#40). As such, the expected relationship between emotions and moral content holds for the vast majority of pre-validated scenarios, and reasonable explanations may be given for approximate emotional equivalence with regard to the scenarios discussed.

However, some of the researcher generated impurity scenarios do not share the expected relationship which, given the lack of wider content validation for these scenarios, may call into question whether these are strictly impurity scenarios. Scenarios #53 and #56 involve the destruction of historic architecture and irreplaceable items respectively, whilst #55 involves protesting a funeral, and #52 concerns ignoring a last will and testament. Such actions might be considered as being more salient with regard to the Authority, or possibly the Loyalty foundation(s), and as such may not show the expected emotion-content relationship. The researcher generated impurity scenarios were also considerably more angering than impurity scenarios from other sources, and further work would be required to ascertain whether this might be explained by overlapping moral content within these scenarios - although the strong correlations between the 'binding foundations' noted by Schein and Gray (2018) suggests this is

likely a minor issue, and thus unlikely to have any noteworthy effect when comparing harmful and impure content.

The results show similar overall correlations for both anger and disgust with regard to moral content. Increased ratings of anger correlate with scenarios being rated as more harmful, whereas increased ratings of disgust correlate with scenarios being rated as more impure. However, the results are slightly mixed in that disgust shows a weak, but consistent, correlation with impure balance ratings across both scenario sets, whereas anger shows a moderate correlation with harm ratings in the first scenario set, but only a very weak correlation in the second set. These findings may be taken to support MFT's claims of 'characteristic' associations between emotions and moral content, and suggest validated MFT scenarios may be better suited to detecting associations between emotions and moral content. However, there is also substantial overlap between anger and disgust, such that all violations typically seem to elicit both emotions to some degree. This overlap might be taken to favour more 'domain-general' accounts of morality such as TDM (Cameron et al., 2015; Schein & Gray, 2018), although it may alternatively be taken to favour suggestions that the vast majority of moral violations can be seen as violating more than one foundation. This latter interpretation seems more likely given the appearance of 'characteristic' associations, despite the frequent and substantial co-occurrence of anger and disgust, in response to both harmful and impure moral content.

4.4.5. Conclusions and Implications for the Theory of Dyadic Morality

These findings pose problems for Dyadic Morality as outlined by Schein and Gray (2018), especially as they cite Gray and Keeney's (2015a) findings when trying to

explain away differences across content areas with regard to moral judgement. The results show participants reliably classified scenarios as more harmful or more impure in accordance with a priori expectations, such that scenarios pre-classified as depicting harmful/impure content were generally rated as being respectively more harmful/impure. There was considerable agreement as to the type of moral content depicted in each scenario, although there seems to be some variation with regard to how individuals classify moral content. This finding seemingly undermines a key pillar in the argument advanced for Dyadic Morality - that 'foundational' values are 'transformations or intermediaries of harm' - as participants can readily discern violations of purity from those of harm. Whilst Dyadic Morality might maintain that some 'foundational' values may be subsumed under 'harm', it remains difficult for TDM to account for values relating to 'purity' in the same manner. The results support claims that factors of weirdness and severity are part of normal variation across moral content (Graham, 2015), rather than confounding factors which may undermine MFT-based claims regarding 'Sanctity' (Gray & Keeney, 2015a; Schein & Gray, 2018).

In particular, the results do not support claims that impurity violations are just a weirder and less severe type of moral violation. First, when all scenarios were assigned a content label based on participant responses, there was no significant difference between ratings of wrongness across harm and impurity. Second, although the first set of scenarios did show that impurity scenarios were rated as less wrong, the standardized difference between ratings was only a quarter of that shown in Gray and Keeney's (2015a) sample - and this result may be somewhat reliant on variability across the 'naturalistic' scenarios; whereas the second set of scenarios shows the opposite trend, with impurity being rated as comparably wrong to harm, and as more wrong when accounting for issues with content validity. Third, ratings of wrongness averaged across all MFT harm scenarios and MFT impurity scenarios are comparable. Fourth, researcher generated impurity scenarios were rated as comparably wrong to MFT impurity scenarios, despite being rated as less weird. In contrast, the 'naturalistic' impurity

scenarios were the least wrong scenario grouping, and the least atypical scenario grouping - such that even 'naturalistic' harm scenarios were weirder. Fifth, the results show a moderately strong correlation between wrongness and atypicality, which was not found by Gray and Keeney (2015a), suggesting that morally worse violations may also tend to be perceived as more atypical – and this is likely due to their abnormality rather than their infrequency, as Chakroff and Young (2015) argue abnormality may incorporate infrequency, but infrequency does not incorporate abnormality. Sixth, some violations of harm, which were rated as such, were also considered to be highly atypical, and some impure violations, also rated as such, were not seen as particularly atypical. Seventh, although it remains possible perceptions of atypicality impact on moral judgement via the extent of norm violations, as Schein and Gray (2018) propose, the comparability between harmful and impure moral content in terms of wrongness suggests associating atypicality with differences relating to moral patiency may be a more defensible proposition. Weirdness might be related to perceptions of the verifiability, and vulnerability, of the victim. If moral judgement operates with reference to a dyadic template, as Schein and Gray (2018) argue, then impure actions may be considered weird because they tend to be self-focused and/or lacking verifiable victims - they are *morally* weird. However, although this explanation may seem to provide some concession to TDM, accepting it would jeopardise a key premise of the theory - the requirement for perceiving two minds. Violations of impurity would be *morally* weird because they are non-dyadic, although this need not necessarily impact on the perceived moral severity of such actions.

Furthermore, ratings across certain researcher generated impurity scenarios call into question a key tenet of the 'harm-as-prototype account' favoured by Gray and colleagues (Gray, Young & Waytz, 2012; Schein & Gray, 2015, 2018). On this account, moral violations where harm is 'strikingly' obvious (e.g., #46) should be rated as being morally worse than those where any harm is less salient. However, several of the impurity scenarios rated by the same participants were considered generally more

immoral, despite any potential victim(s) in these scenarios seeming less salient. Of those with the highest wrongness ratings (6+), the majority of scenarios dealt with actions that can be summed up along the lines of 'disrespect for the dead'. Although it may be possible to argue that these are potentially concerns regarding the Authority foundation, where disrespect is considered a prototypical violation, arguing that these involve harm is much more of a challenge. If harm is defined with regard to physical and/or emotional suffering, as it is normally operationalized in literature, then being dead would seem to preclude the potential to experience such suffering; whereas if harm is defined with regard to the extent of perceiving a morally dyadic relationship, then it is not immediately obvious who the patient is in these scenarios.

Schein and Gray (2018) might choose to explain this with reference to 'patientic dyadic completion', where people are inclined to identify a victim in response to intentional agents violating moral norms. In detailing this concept, they cite DeScioli, Gilbert and Kurzban (2012) who show people see suffering, and identify victims, in scenarios which can be described as harmless, in that they lack an obvious victim. However, DeScioli et al.'s (2012) preferred interpretation of their own data suggests "that people readily fabricate victims when they are unavailable" (p. 147), which seems contrary to Schein and Gray's (2018) assertion that perceptions of a suffering patient (in combination with an intentional agent) are driving moral judgements. Yet even if we accept that 'unverifiable victims', such as the disrespected deceased person, or their surviving family, are perceived to suffer as a result of dyadic completion, these victims are not 'objectively' present. Any suffering would appear to be perceived inferentially rather than directly, and inferential perception is a substantially weaker requirement than direct perception. Indeed, harmful scenarios where there is an obvious victim might have been expected to come out as morally worse than when no other people are involved (e.g., #62), providing a challenge to claims that the immorality of an act depends on the perception of (dyadic) harm. Of course, one may perceive *potential* for (dyadic) harm when considering the use of child-resembling sex toys, but this too is a substantially

weaker requirement. All kinds of actions might have such potential, and the perception of possible future harms would seem to rely on at least some reasoning, rather than being an intuitively quick judgement. Furthermore, considering either deceased individuals, or the acting intentional moral agent (i.e., for self-directed violations), as the moral patient would seem to require drawing on concepts related to the Sanctity foundation for justification. As such, regardless of whether harm is defined according to MFT or TDM, the results suggest that actions perceived as more harmful are not necessarily more wrong than impure actions, and neither are they more impure. Thus, although impure actions may be the weirdest kind of immoral actions, they are not just weird.

Chapter 5 - On Constructive Sentimentalism

Constructive Sentimentalism (CS) has been advanced and defended by Prinz (2006b, 2007a, 2008, 2009, 2013). The philosophical groundwork for CS is substantially more developed than that offered by either Dyadic Morality (TDM; Schein & Gray, 2018) or Moral Foundations Theory (MFT; Graham et al. 2013); CS is underpinned by independent, non-moral, accounts of concepts (Prinz, 2004a), and emotions (Prinz, 2004b), and in this manner, has strong parallels with the account of morals advanced by Hume (1751) -- a philosopher many moral psychologists, especially those advancing intuitionist accounts, seem to regard favourably. There is substantial common ground between CS, TDM, and MFT; for example, all theories draw on work by Shweder et al. (1997) which suggests there are various linkages between morality and accounts of suffering, all accounts include an important role for culture in moral development, and all accounts argue for moral pluralism in various ways. However, there are also important differences between the theories (see Table 5.1.). CS differs in being an emotionist theory, rather than the intuitionist accounts offered by TDM and MFT; and further differs in that Prinz (2008) argues against the innateness of morality, in contrast to both TDM and MFT. Prinz (2004b, 2006a) also argues against modularity, similarly to TDM; but argues for both moral domains, and emotion specific links with regard to these domains, similarly to MFT. As sensibility theories, such as CS, can accommodate intuitionist claims (Prinz, 2009), and there is substantial agreement between the theoretical approaches discussed, it may be possible to integrate and unify both TDM and MFT within a CS framework.

Table 5.1. Different positions taken by different theoretical approaches to morality.

Claims of Theory	TDM	MFT	CS
Intuitionist	Yes	Yes	No
Innateness	Yes	Yes	No
Modularity	No	Yes	No
Moral Domains	No	Yes	Yes
Emotion Links	No	Yes	Yes

5.1. Outlining Constructive Sentimentalism

Constructive Sentimentalism (Prinz, 2009) proposes moral content is essentially related to emotions, offering the following schematic formulations for moral judgement; it comprises a metaphysical thesis -- moral properties are constituted by emotions -- such that "(S1') An action has the property of being morally wrong (right) just in case there is an observer who has a sentiment of disapprobation (approbation) toward it." (p.92), and an epistemic thesis -- moral concepts are constituted by emotions -- such that "(S2-W) The standard concept WRONG is a detector for the property of wrongness that comprises a sentiment that disposes its possessor to experience emotions in the disapprobation range." (p.94). According to this proposal, moral properties are defined as 'powers to cause emotions in us', and moral concepts -- which incorporate the elicited emotions -- handle the detection of these properties. Moral properties are thus analogous to colour properties, in that they depend on features of the world, such that certain events possess certain relational properties which reliably cause emotions to be elicited -- just as surfaces reflecting light of particular wavelengths reliably cause colours to be perceived -- provided the observer possesses the perceptual apparatus to detect

said emotions/wavelengths. As such, the perception of such properties depends on subjective states as well as on features of the world.

On this approach, sentiments are understood in terms of emotional dispositions -- there is something 'set up' in the observer which reliably causes an emotional experience when it is 'set off' by detecting instances of *Φ'ing.* For the sentiment to count as a moral sentiment it must be (dis)approbative, meaning a moral sentiment toward *Φ'ing* must dispose its possessor to experience both self- and other-directed emotions of blame (or praise). Sentiments of such type can be termed moral rules, and moral judgements are particular emotional manifestations of these sentiments. In general mechanistic terms, some act, event, or circumstance, is perceived and categorised as an instance of *Φ'ing* (e.g., stealing), this sets off the sentiment; contextual factors then determine emotion elicitation patterns, and the resulting emotion becomes bound to the concept of *Φ'ing,* producing a compound state which can be expressed as 'blameworthy-emotion' at *Φ'ing* (e.g., anger at stealing). This compound can be equated to the judgement *Φ'ing* is wrong, where *Φ'ing* is wrong because the emotion within the judgement was generated via a moral rule.

CS specifies the contextual factors relating to emotion elicitation patterns as depending on classification of both moral domains and the observers' relationship to the transgressing agent. Prinz (2009) details three moral domains, strongly aligned to those described by Shweder et al. (1997). The 'Divinity' ethic, akin to MFT's 'Sanctity/degradation' (impurity) foundation, is construed more broadly as 'transgressions against the perceived natural order' -- and considered as one of two fundamental domains. The other fundamental is that of 'Autonomy', defined as 'transgressions against persons', and akin to the 'Care/harm' foundation of MFT -- although certain conceptions of fairness may fall into this domain. The third class is understood as being derived from these fundamentals, thus covering 'transgressions against the natural order of persons', akin to Shweder et al.'s (1997) ethic of

'Community', or MFT's 'Loyalty/betrayal' and 'Authority/subversion' foundations -- although certain conceptions of fairness may also fall into this domain. Furthermore, the emotion elicited will be either reactive, as when someone else transgresses (other-blame), or reflexive, should the observer perceive themselves as having transgressed (self-blame). Additionally, there is a derived class that sits between reactive and reflective emotions, such that is likely to be elicited when someone close to the observer (e.g., a family member) transgresses in a way which is thought to reflect on the observer.

5.2. The Main Hypotheses of Constructive Sentimentalism

Accordingly, Constructive Sentimentalism provides a series of testable predictions. Primarily, it maintains the mappings between (reactive) emotions and moral domains reported by Rozin et al. (1999), such that transgressions against persons elicit anger, transgressions against the natural order elicit disgust, and transgressions against the natural order of persons elicit contempt -- an emotion which Prinz (2009) suggests is a blend of anger and disgust. However, it expands on these to include reflexive emotions -- where the observer perceives themselves to have transgressed -- such that self-performed transgressions against persons elicit guilt, transgressions against the natural order elicit shame, and transgressions against the natural order of persons elicit a blend of guilt and shame. Additionally, CS proposes there may be derivatives of these emotions elicited by transgressions where the observer has an emotional attachment to the transgressor. For example, an emotion of 'hurt', which Prinz contends is a blend of anger and guilt, may be elicited by the transgressions of someone loved by the observer if they violate an autonomy rule, whereas the violation of a community rule by that person may elicit 'disgrace' -- a blend of hurt and shame.

Sensibility theories, such as CS, provide further testable hypotheses through their focus on emotion. Firstly, in matters of moral judgements, having the requisite emotions in response to *Φ'ing* should serve as a better indicator of the judgement or moral stance towards *Φ'ing* being genuinely held (by an individual) than verbal behaviour simply attesting to the moral qualities of *Φ'ing*. Secondly, links between emotions and behaviour suggest transgressions in different domains will have differing influences on behaviour. In this way, violations of autonomy, which draw on anger, may predict greater aggression and desire for retribution than violations of the natural order, where disgust may motivate a strong withdrawal response. For example, people might be expected to prefer being neighbours with someone known to violate autonomy rules (e.g., a thief) over someone known to violate the perceived natural order (e.g., a bestiality practitioner). Thirdly, conventional norms are more likely to be moralized if the convention is learned through emotional conditioning. For example, children learning to wash their hands after going to the toilet should consider such a behaviour as more morally relevant if taught in a way that emphasizes emotions (e.g., "...it's disgusting to leave germs on your hands..." -- "...you would be annoyed if you got sick from someone else's germs...") rather than norms or conventions (e.g., "...it is normal to wash your hands after using..."). Fourthly, moral and conventional norms should be distinguishable by appeal, such that a norm is conventional if it depends on an appeal to customs, whereas moral norms do not necessarily appeal in this way -- murder is generally not considered wrong because 'it is not what we do round here'. Finally, CS predicts that the strength of moral judgements should be commensurate with the strength of the disposition towards (dis)approbation, such that moral rules which token intensely valenced emotional states should be judged as more morally extreme than that which token less intensely valenced (dis)approbative emotions.

5.3. Sentimentalism and Dyadic Morality

Other sentimentalist theories, such as that advanced by Nichols (2002, 2004; cited by both Prinz, 2009, and Schein & Gray, 2018), bear remarkable similarities to TDM. Nichols also proposes a bi-directional causal model, where norm violation prompts feelings of concern, and feelings of concern prompt norm retrieval/formation. Here, concern is mainly understood in terms of sympathetic distress, empathic concern, or more simply as 'caring about the vulnerable mind' (Schein & Gray, 2018, p.7). Both these theories agree that "norms + feelings" are necessary for moral judgement, although Schein and Gray argue the perception of dyadic harm is further necessary in this regard.

However, closer analysis suggests there is little difference between these theories except in terms of emphasis. Norm violations presumably require an agent to perpetrate (i.e., cause) the violation, a presumption that remains whether or not there is an observer additional to the transgressing agent. Similarly, negative affect presumably requires an agent to experience such feelings, either a victim of the violation, or an observer of same. As such, the template for dyadic harm -- agent causing damage to patient -- can be expressed in terms of norms and feelings; norm violations (by agents) causing negative affect in an observer towards an entity, (who may, or may not, also be) the patient of the transgression, which arises in lieu of said violation. The latter expression seems to contain the former implicitly, making the accounts difficult to differentiate, particularly as Schein and Gray (2018) agree with Nichols (2004) that some form of concern for the patient/victim is necessary. Furthermore, negative affect tends to co-occur with perceptions of physical 'damage', and seems definitional in cases involving mental or spiritual suffering, such that it appears simply qualifying negative affect as being unwanted or undesirable is enough to make these theories somewhat indistinguishable. Stating moral judgements arise in response to perceiving 'norm violations causing undesirable negative affect' seems similar to saying they arise in

response to perceiving 'agents causing damaging experience'. Norm violating agents are present on both accounts, as are causal links, as are concern(s) for patients. If norm violations require (intentional) agents, experiencing negative affect requires (vulnerable) patients -- for whom one is concerned, and causality is implicated in both, either through the agent causing the norm violation, or the feelings being caused by the norm violation, it is not clear how 'dyadic harm' is required in addition to "norms + feelings" – even if it may be contained within them.

5.3.1. Are "norms + feelings" sufficient for moral judgement?

Further interrogation of Schein and Gray's (2018) argument is instructive. For example, they argue "negative affect cannot *alone* distinguish the immoral from the unconventional as people feel negative affect and express anger when observing nonmoral norm violations (Brauer & Chekroun, 2005; Santee & Jackson, 1977)." (p.5). However, it is debatable whether the norm violations studied by Brauer and Chekroun (2005) are nonmoral, as most of these could be classed as transgressing the natural order of persons, following Prinz (2009), or as failing tests for Kantian universality. Indeed, 69% of participants reported the transgressions described by Brauer and Chekroun (2005) warranted the exercising of some form of 'social control' in redress, and 48% exercised forms of this when witnessing these transgressions -- most notably an 'angry look' in 28% of cases. As such, it seems plausible to argue that at least some of these participants may have moral attitudes towards the transgressions. Additionally, transgressors received more angry looks for violations that were more personally relevant to the witness, such as cutting in front of them in a line, or dropping litter in the entrance to their apartment block accommodation -- which could both be construed as instances of 'receiving disrespect' for which an angry or contemptuous look may be the predicted response (Prinz, 2009).

Additionally, Santee and Jackson's (1977) findings show that the strength of potential normative sanctions (via disapproval ratings), which may be considered similar to the exertion of 'social control', varies depending on how relevant the behaviours appear to be with regard to fraternalism, independence, and expressionism - values which bear reasonable similarity to 'moral foundations' of loyalty, liberty, and autonomy (care/harm). Santee and Jackson's (1977) findings are supported by Brauer and Chekroun's (2005), in that both studies suggest participants are motivated towards exerting forms of 'social control' to curtail certain behaviours, and this varies to the extent which they (may) moralize these behaviours. As before, although 'non-moral' norm violations may elicit negative affect, it is still by no means clear that the violation is construed as non-moral by the person experiencing said affect. Yet both findings can be readily explained by Constructive Sentimentalism, simply in that if some participants do have moral attitudes to the behaviours investigated in these studies, then their reported motivations are in line with what would be expected were emotions part of moral judgements.

Furthermore, Schein and Gray (2018) continue their argument with reference to Nichols (2002) example of spitting on something before you consume it, describing the act as one that is "certainly both counter-normative and disgusting, but lacks the authority-independent, objective-seeming punch of other immoral deeds." (p.5). However, here Schein and Gray seem to be exploiting an intuition pump, as Nichols (2002) reports no significant differences between spitting in - and then drinking - a glass of water, and hair pulling (a ‘harm’ violation), with regard to ratings of permissibility, seriousness, or authority-contingency; and although justifications for why each behaviour was bad were different in these instances, the majority of participants directly appealed to disgust in justifying the spitting case. This was in contrast to justifications about why it was bad to drink soup directly from the bowl, which invoked etiquette or rules -- a pattern which concurs with Prinz's account of the moral-conventional distinction -- and is further supported via a more 'conventional' ratings pattern, in that

soup drinking was considered more permissible, less serious, and more authority dependent than the water spitting case. Thus, the results reported by Nichols would seem to suggest that such spitting behaviour may be considered immoral, at least by some participants, even if it only occupies a fairly innocuous position on a continuum of (im)morality.

Of course, Schein and Gray could concur that some participants are construing immorality in these actions, arguing that such participants are, in fact, perceiving dyadic harm. Support for this approach is provided via Royzman, Leeman, and Baron (2009), who improve on Nichols (2002) methodology, showing that the 'social transcendence' (authority/rule-independence and generalizability) of transgressions only correlates with, and is predicted by, perceiving the behaviour to be harmful such that it has a 'negative affect on others'. This effect of 'harm' is shown to be independent of being 'grossed out' by the act, and is present across both the relatively minor transgressions described by Nichols (2002) and the canonical case for 'moral dumbfounding' -- consensual incest (see Haidt et al., 2000). As 'domain specific intuitions' and affect showed no predictive power, nor a relationship with 'harm', Royzman et al. conclude such dissociation poses problems for accounts of moral judgements which consider a majority of concerns about harm as arising from post hoc reasoning processes (e.g., Haidt, 2001). This conclusion would likely be welcomed by Schein and Gray, as it suggests an important role for 'harm' in moral judgement.

However, Royzman et al.'s (2009) findings may be taken to support Prinz's sentimentalist approach, even if they do not support the one advanced by Nichols (2002). With regard to consensual incest, descriptive statistics indicate just over half of Royzman et al.'s participants responded as seeing someone presently 'harmed' by the actions, similar numbers stated the wrongness of the action was 'socially transcendent', and just over two thirds responded that the action would increase the likelihood of future 'harms'. Taking the most generous of these figures still leaves just under a third of

participants claiming consensual incest is wrong, yet reporting not to perceive anyone being negatively affected by this act -- which would seem to pose a problem for Dyadic Morality. In comparison, over four fifths of participants thought the transgressors should feel guilty about their actions, and according to CS, this 'self-directed' disapprobation is what would be expected following the violation of a moral rule. The remainder not responding this way are also more readily explainable, in that shame may be the more appropriate emotional response -- and this was not examined. Furthermore, some participants may be construing 'should feel' as an expectancy, even if most are using it as 'ought to feel'. The design allows participants may be recognising the siblings apparent lack of a moral rule against incest, such that the siblings should not feel guilty because they do not consider the action as immoral, even though the participant does construe it this way and reports it as such.

In short, the three citations Schein and Gray (2018) token in advancing their argument that "norms + feelings" are insufficient for moral judgement fail to provide them with substantive support for the claim. Furthermore, the findings of each study may be easily explained via Constructive Sentimentalism. At worst, each source can be turned against Schein and Gray's argument, as each suggests at least some participants may construe negative 'nonmoral' norm violations as immoral, and may be doing so without recourse to 'harm' -- the dyadic template may be unnecessary. At best, Royzman et al. (2009) might be taken as tentative support for Dyadic Morality over Nichols sentimentalism – even if TDM cannot rule out Prinz's sentimentalism. Yet even if the best-case argument were accepted, it requires allowing a broad definition of 'harm', such as that used by Royzman et al. and Dyadic Morality. However, authors aligned to Moral Foundations Theory have highlighted issues with precisely this concept of 'harm pluralism' (Haidt et al., 2015). The definition of harm is of the utmost importance for Dyadic Morality, and as such, the question of whether "norms + feelings" are sufficient for moral judgement can be rephrased to approach from a harm-centric perspective – and contrasted with Constructive Sentimentalism.

5.3.2. Is perceiving 'harm' necessary for moral judgement or are there 'harmless' wrongs?

Schein and Gray's (2018) definition of 'harm' and its role in moral judgement may be formalised in line with that provided by Prinz, such that: (DM1) An action has the property of being dyadically harmful just in case there is an observer who perceives it as an instance of *an intentional agent (iA) causing (→) damage (d) to a vulnerable patient (vP).* The action also has the property of being morally wrong because instances of dyadic harm are norm violations which generate negative affect*,* and (im)morality is constructed by norms and affect in conjunction with the dyadic template. Notably TDM uses '(dyadic) harm' to refer to the presence of all elements, such that harm = *iA→dvP* = immoral (as dyadic harm itself is a norm violation that generates negative affect). The links between each element allow for separate questions to be posed with regard to 'harm' and immorality, closely reflecting the possible 'dyadic completions' described by Schein and Gray (2018) - although the formulation above tokens four elements - for reasons which will become apparent, whereas Schein and Gray suggest there are only three in that '*→d*' is not separated.

Pairings of these three elements, agency, causality, and patiency, are used by Schein and Gray (2018) in detailing three types of dyadic completion. First, "[t]he presence of **evil** agents and **suffering** patients (A P) compels *causal dyadic completion*, the perception of a causal link between them." (p. 19, **emphasis** mine). Second, "isolated suffering patients (→P) compel *agentic dyadic completion*, the perception of intentional agents to account for their suffering.". Third, Schein and Gray (2018, p.19) state "[p]erhaps the most important form of dyadic completion is *patientic dyadic completion* in which norm violations completed by intentional agents (A→) compel people to see moral patients harmed by those acts (DeScioli, Gilbert, & Kurzban, 2012;

Gray et al., 2012; Gray et al., 2014).". However, as emphasized, there may be an issue with the first of these forms of completion. This becomes apparent when using the expanded notation forms for these pairings.

There is no obvious issue in the third case, *iA→d,* nor in the second, *→dvP,* but the first case, *iA vP,* does not seem valid. Using 'suffering', rather than merely 'vulnerable', in the (A P) case would seem to necessitate an expansion to *iA dvP.* Yet if this is so, the qualifier 'isolated' does not seem to justify the addition of '→' in the second case, as the patient must also be 'isolated' in the (A P) case by definition -- it cannot be *iA dvP*. Accordingly, causal dyadic completion seems to rely on damage being smuggled into the formula, by imputing suffering onto the patient, as the un-noted form of *iA vP* would read "[t]he presence of evil agents and vulnerable patients (A P) compels *causal dyadic completion*, the perception of a causal link between them." -- a statement which seems substantially less plausible, especially if "evil" is replaced with "intentional".

Furthermore, the concept of causal completion does not seem well supported by Schein and Gray's (2018) cited material. Alicke's (1992) findings do indeed show that blameworthy actions are considered the prepotent causes of unfortunate events, but these findings revolve around the attributions of causality, for which there are several interrelated causal factors -- causality is not absent. Likewise, Knobe (2003) shows the extent to which an action is rated as being deserving of blame or praise is correlated with ratings of whether the side effect of that action was intended -- the cause of the action is utterly unambiguous. Both Alicke and Knobe's findings are focused on agents and causes, *iA→d,* rather than agents and patients. The causing of damage, *→d* (or *→dvP*), is at the centre of the scenarios used already, so attributing responsibility is just asking the extent to which the agent present (*iA*) is causally linked to this damage (*→d*). This remains the case even if 'causing damage' is separated, such that they are about responsibility for causation – or blameworthiness (*iA→*) – of damage (*d*). As such, if *d* can be isolated in this way then the presence of *vP* seems optional, whereas if these are

bound (such that *dvP*) then investigating blameworthiness (*iA*→) leaves nothing for ‘completion’.

Schein and Gray's (2018) last citation in supporting claims of causal completion is the only one which might be interpreted as involving a lack of 'cause'. Here, they token causal completion as explaining "why evil thoughts seem to cause the suffering of others more than good thoughts (Pronin et al., 2006)." (p.19). However, Pronin, Wegner, McCarthy and Rodriguez's (2006) findings show a 'belief in personal causation' is not limited to instances of perceived harm. The effect occurs regardless of whether the thoughts prior to the outcome are positive or negative, and importantly, further regardless of whether the outcome is desirable or undesirable. Pronin et al. report having prior thoughts relevant to an outcome seemed to increase feelings of personal responsibility for having caused that outcome -- even if said thoughts and outcomes were incongruous. As such, participants instructed to think game relevant thoughts (i.e., concerning player contribution) rated themselves as being very slightly more responsible for the outcome of that game than those instructed to think of related but irrelevant thoughts (i.e., concerning player identification); and participants thinking a lot about a sports game whilst cheering for their team felt 'very slightly' more personally responsible for the outcome than those not thinking much about the game, regardless of whether their team won or lost. Those cheering for the losing team considered themselves as having similar levels of influence on the outcome to those cheering for the winning team. In more formal terms, this again seems to emphasize agents and causes, as patients are not obviously present in Pronin et al.'s (2006) latter studies and even if we accept there are patients, such that supporters of the losing team 'suffer', the results show outcome variation has no effect on ratings of personal responsibility. What matters is having the relevant thoughts, rather than the specific content of those thoughts. Accordingly, whilst evil thoughts may seem to cause suffering, cheering for your side may also appear to cause them losing. This seems contrary to what causal completion

might predict and effectively nullifies Schein and Gray's (2018) use of this citation, leaving their argument for causal completion unsupported.

However, although the concept of causal completion may seem to want for evidence, it also seems to be the least important of the three. Dyadic Morality might simply accept the causal component, *→d,* must be linked to either the agent or patient -- there are no (A P) cases. The most important requirement, that of two minds, would seem to remain relatively unscathed; so agentic and patientic dyadic completions would seem to remain viable. Yet as seen in the critique of causal completion, the causal component seems substantially more 'bound' to the agency side of the dyad, making it further questionable whether an 'isolated suffering patient', *→dvP,* is different to a patient which is only suffering, *dvP,* or merely vulnerable, *vP*. As such, agentic completion might be expressed as *dvP*, and patientic completion as *iA→*.

Yet even if agentic completion is granted, such that people do seek to explain suffering, and typically account for this phenomenon in moral terms (cf. Shweder et al., 1997), this motivation towards explanation seems to be different to that underlying patientic completion. Even if agentic completion were interpreted in stronger terms, as a motivation to 'find the wrong doer!', it still seems importantly separable from the more 'look at the wrong that has been done!' motivation of patientic completion. Interestingly, and running contrary to the earlier assertion, the only common ingredient in both motivations is that 'wrong' has occurred. Expressed in formulaic terms, this common denominator would seem to be *→d*, which in formal terms, equates 'wrong' to 'caused suffering'.

This approach allows separate understandings of 'harmless wrongs' to emerge. Can something wrong be 'harmless' in that it does not involve some perception of caused suffering? Or more broadly, in line with Harris (2012), can something be

regarded as being morally relevant if it is not perceived as relating (in some way) to broadly construed notions regarding well-being? The answer to these questions is likely 'no'. In contrast, the question of whether something can be 'harmlessly' wrong if it does not involve some perceptions of patiency might be answered in the affirmative -- there may be 'victimless' wrongs, even if there are no 'unsuffering' wrongs.

If this is the case, then whilst it is possible that perceiving agentically caused suffering (with no clear victim), *iA→d,* might be sufficient to judge something as being wrong, such a possibility seems less plausible for cases where suffering is caused to a vulnerable patient in the apparent absence of an agent (*→dvP*), such as that occurring via a natural disaster. Even if people attribute suffering arising from a tsunami as having some causal relationship to a supreme being, they seem to regard such 'blameworthiness' differently to normal -- the supreme being is typically seen as doling out punishment for *other* immoral actions, rather than being blamed for the tsunami – the (direct) causal source of the suffering. Thus, whilst agentic dyadic completion may indeed compel the perception of agents in accounting for suffering, it is not clear that this compulsion is necessarily driven towards the apportioning of blame, or the attributing of moral responsibility, to the agent perceived as causing the suffering in question. This suggests agentic completion may operate differently in relation to matters of moral judgement than patientic completion; and if there are victimless wrongs, such that *iA→d* is sufficient for moral judgement, then patientic completion would seem to fit the description of 'post-hoc' by definition. As such, perceptions of harm, in the dyadic sense of requiring a victim, may be unnecessary for moral judgement. There may be victimless wrongs.

Investigating the key sources cited by Schein and Gray (2018) in support of patientic completion is once again instructive. DeScioli, Gilbert, and Kurzban (2012) do suggest that there is an important role for victims in moral thought, showing that victims are perceived as having been wronged, even though victims are 'objectively' absent in

certain events, such as flag burning, dog eating, and grave desecration. Yet, as previously shown in Royzman et al.'s (2009) research, some participants rated scenarios as wrong and stated there was no victim. Almost half saw no victim for burning the flag, almost a quarter saw no victim for dog eating -- although three-quarters nominated an 'unverifiable' victim (i.e., the dog), and, in line with Royzman et al.'s (2009) findings, around a third of participants rated incest as wrong despite identifying no victims. A similar results pattern is also shared by participants in Gray, Schein, and Ward's (2014) study, where questions of whether impure actions have a victim were rated, at most, as being at the midpoint between 'definitely no' and 'definitely yes'. However, this result shows participants were unsure of whether a victim was present in 'harmless' (victimless) impurity scenarios, particularly under time pressure, but this is not the same as saying participants perceived victims, which is what Gray, Schein, and Ward (2014) claim their results show.

Additionally, whilst DeScioli et al. (2012) acknowledge their findings might be taken as evidence that suffering victims are fundamental to moral judgement, they favour the opposing possibility -- victim fabrication -- as being more likely interpretation given that participants tended to nominate 'unverifiable victims'. They suggest patientic completion is indicative that victim suffering can be considered essential for moral 'models', but that it is not an input essential for moral 'computations' (e.g., judgements). Furthermore, DeScioli et al. show that for certain offenses, such as suicide or drug use, only one person is thought be wronged. That 'the self' can be both agent and patient, and that 'victimless wrongs' are readily rated as wrong even when participants explicitly state there are no victims, severely undermines claims of *dyadic* necessity in moral judgement. Even allowing that the self can be both agent (i.e., the 'current' self) and patient (i.e., the 'future' self), perceiving dyadic harm is not necessary for moral judgements because such judgements are still made by participants who themselves identify action(s) as having no victims.

In short, claims regarding causal dyadic completion do not appear valid, and instances of agentic dyadic completion may occur without necessarily considering the causal agent as the wrong doer (e.g., for natural disasters). Furthermore, causality tends to be linked with agents, and the perception of victims does not appear necessary for moral judgements – suggesting patientic dyadic completion is a post-hoc process.

5.4. Contrasting Constructive Sentimentalism and Dyadic Morality

The ability of Constructive Sentimentalism to address the concerns identified by decompiling Schein and Gray's (2018) approach may provide testament to the explanatory power of Prinz's, and being able to account for TDM in terms of CS may allow the former to be subsumed into the latter. Indeed, Prinz suggests that it is the agent-patient relationship, rather than moral domain classification, which is most important in matters of moral approbation -- the dyadic template may yet be the focus for 'right', even if it is not of primary concern in matters of 'wrong'. Furthermore, given the agreed upon importance of harm for morality (in the MFT sense), it can be argued TDM still provides an important contribution -- even if it is incorrect in asserting the necessity of dyadic harm for judgements of immorality.

Prinz argues that a focus on the victim, common to both Dyadic Morality and Nichols sentimentalism, is misplaced. The emotions involved in moral judgements are generally focused, particularly initially, on the act or toward the perpetrator; any concern for victims is usually an afterthought in such matters. This ties in with the critique of causal completion, where attributions of causality are shown to be more closely bound to transgressors than victims. Additionally, the necessity of a patient for moral judgement has been substantially diminished by the apparent existence of both victimless wrongs and self-focused wrongs (e.g., suicide) discussed in the critique of patientic completion.

Furthermore, the remaining viable elements of Dyadic Morality, that of an intentional agent and causality, seem closely related to concepts of blame and praise.

Prinz defines (dis)approbative emotions, which constitute moral judgements, as those relating to blame or praise which can be directed at oneself or another, such that for an action to be judged morally wrong, it needs to be an action which disposes the observer to experience disapprobative emotions of *both* self- and other-blame towards it. Furthermore, in defining emotions more generally, Prinz (2004b) suggests emotions are best understood as 'felt perceptions of the organism-environment relationship with regard to well-being'. If correct, this readily accounts for the apparent impossibility of 'unsuffering wrongs' discussed in the critique of dyadic completion. If moral judgements are constituted by emotional dispositions, and emotions are perceptions concerning well-being, then moral judgements are constituted by perceptions concerning well-being. This also fits neatly with the critique of causal completion; it explains how the causal element of suffering, (→)*d,* or in emotion terms, perceptions concerning well-being, seemingly must be linked to agents or patients, and why there are no obvious (A P) cases. On Prinz's approach, it is these felt perceptions concerning well-being - emotions - which are the common denominator necessary for moral judgement; the perception of patients, or even agents, are not essential elements.

That emotions are constituents of moral judgement, rather than merely causally connected components as they are on other approaches (Graham et al., 2013; Nichols, 2004; Schein & Gray, 2018), further provides a more satisfactory account of moral dumbfounding, and studies where acts are rated as wrong despite having no 'objective' or 'verifiable' victims, as well as also answering the question begged in Schein and Gray's approach. According to Prinz, the stupefaction of participants in ‘dumbfounding’ studies occurs as a result of hitting a 'grounding norm' -- a basic value where reasoning bottoms out in affect. For example, questions might be asked with respect to each of MFT's foundations: Why is imprisoning the innocent wrong? Why is incest wrong? Why

is causing chaos and disorder wrong? Why is betrayal wrong? Why is cheating wrong? Why is harming children wrong? -- the responses to such questioning are likely to be ones of astonished incredulity -- 'they just are!'. Participants may be unable to articulate reasons for their judgements precisely because these judgements are constituted by emotions -- for Prinz, the responders' dumbfoundedness means 'it is wrong because I am disposed to experience disapprobation towards such acts'. The judgement is self-justifying, in that the person perceiving the act is disposed to experience an (ecologically sensitive) emotion in response - which is constitutive of wrongness. The act is considered wrong because of the way in which the participant reacts to it, and whilst the response may be rationalized, the reaction itself is arational. To answer the question Schein and Gray (2018) beg, 'why is it immoral for an intentional agent to cause suffering to a vulnerable patient?' -- it is because dyadic harm is being perceived as violating an affectively constituted norm. This remains the case for acts where the perception of dyadic harm is incomplete, explaining why actions often rated as having no victim (e.g., flag-burning), and actions which are lacking a clear dyad due to their self-inflicted nature (e.g., suicide), are still rated as morally wrong. As such, affectively constituted norms (sentiments), which contain elements that both TDM and CS propose are necessary for moral judgement, may also be sufficient for moral judgement. In contrast, the perception of dyadic harm does not seem necessary for moral judgement, although it may be sufficient provided the judge has the necessary sentiment(s) towards any such harmful actions.

5.5. Constructive Sentimentalism on Dyadic Morality

From this position, Constructive Sentimentalism can directly address the questions posed and answered by Schein and Gray (2018) regarding moral content. (a) An act is immoral if the person judging said act has a sentiment of disapprobation towards it. (b) The synthetic definition of harm, in the dyadic sense, is plausible, but not necessary as neither patients nor agents are required for an act to be judged immoral.

Accordingly, whilst each of the three elements: agency, causality, vulnerability, may be perceived separately, even intuitively, and likely influence moral judgement, only the causal element of suffering is necessary -- the act relates, in some way, to a broadly defined concept of well-being. Furthermore, it is questionable whether acts perceived as causing spiritual defilement can be placed in the same 'causal' category as acts perceived to cause physical destruction, and even if they can, it is further questionable whether the 'continuum of immorality' they occupy is a linear one between minimal and maximal dyadicness. (c) The 'fuzziness' of the dyadic template is to be expected; but just as there may be ambiguity around perceiving agency, patiency, or suffering, and variations in the salience of these perceptions, these variations and ambiguity are similarly present in determining whether an act is categorised as violating a moral rule (sentiment). (d) There are harmless wrongs, in the victimless sense -- that some wrongs are seen as wrong without any perception of victims. Moral dumbfounding results are explainable by grounding norms in emotion. (e) There is at least one other moral value which is not a transformation or intermediary of harm, and this value encompasses 'Sanctity/degradation' or 'impurity' (Graham et al. 2013) concerns into the 'perceived natural order' – and transgressions in this area may be separable from those perceived as being against persons.

Constructive Sentimentalism can also comment on the corresponding questions concerning moral mechanisms. (a) We make moral judgements as the result of a sentiment being 'set off' by an immoral act it has been 'set up' to detect. The emotional manifestation of this sentiment constitutes the judgement the act is wrong. The strength of the judgement may be associated mostly with the strength of the disposition it resulted from, although judgements as to the extent of immorality may also depend on the extent to which the act is perceived to violate a (moral) sentimental rule (categorisation), and the conditions under which the judgement is elicited. (b) Moral judgements may affect perception, and there may be room for 'completion' processes to operate, but these centre around acts which are perceived as concerning well-being --

the key component is 'causal suffering', and this alone may compel the search for both agents and patients. However, CS is more conciliatory on the next questions, answering that (c) moral judgements may indeed be extended and entrenched via a (dyadic) loop, although this need not be the only way of doing so. It also agrees with the last two points, such that Prinz both (d) argues against modularity, and (e) favours a constructionist account. Indeed, whilst CS disagrees with TDM about moral content, it is more conciliatory with regard to TDM's account of moral mechanisms.

However, any conciliation regarding mechanisms is likely due to both CS and TDM providing constructionist accounts of morality, and both arguing against modularity in this regard. Furthermore, in terms of content, CS is closer to Moral Foundations Theory, in that it advocates in favour of more than one moral domain – such that morality is about more than just perceived (dyadic) harm. Yet the theories differ in accounting for the role played by emotions, and how emotions relate to moral content. CS argues for specific links between emotions and moral content, similarly to Moral Foundations Theory which argues for 'characteristic associations' – both theories hypothesizing linkages of anger with 'Autonomy'/'Care-harm' violations, and disgust with 'Divinity'/'Sanctity-degradation' violations. TDM argues against any such associations, claiming that the appearance of specificity arises from statistically concealed overlaps and insufficient experimental controls, suggesting such links can be explained by domain general characteristics of core affect and conceptual knowledge of emotion categories (Cameron, Lindquist, & Gray, 2015; Schein & Gray, 2018).

The disagreement between CS and TDM regarding links between emotions and moral content is especially concerning given both are constructionist accounts of morality, and this issue is further amplified in that Cameron et al. (2015) note 'constitutive appraisal models' of emotion are more similar to the constructionist account they are advancing. CS argues moral judgements are constituted by emotions (Prinz, 2009), and draws on Embodied Appraisal Theory (Prinz, 2004b) in defining emotions,

making CS seem likely to be at least somewhat aligned with 'constitutive appraisal models' - and making the opposing positions of CS and TDM regarding emotions and moral content particularly striking.

Correspondences between emotions and moral content, or an absence of such correspondences, thus provide further means of examining which of the theories, Constructive Sentimentalism (Prinz, 2009), Moral Foundations Theory (Graham et al., 2013), or the Theory of Dyadic Morality (Schein & Gray, 2018), is best able to account for the evidence from studies examining the hypothesized correspondences. In particular, the existence of specific and exclusive links between discrete emotions (e.g., disgust) and certain moral domains (e.g., 'sanctity') would be favourable for both CS and MFT, but problematic for TDM; whereas non-exclusive links, whereby typical moral emotions (e.g., anger and disgust) are elicited across different types of moral violation (e.g., harm, unfairness, betrayal, subversion, degradation, oppression), are claimed to be favourable towards TDM (Schein & Gray, 2018). However, as numerous studies have examined the proposed links between emotions and moral content, and have found somewhat mixed result, a closer review of the available evidence is warranted.

Chapter 6 - On Emotion Specificity

Specific correspondences between emotions and moral content, whereby anger is (only) linked to harm, and disgust is (only) linked to impurity, have been the subject of numerous studies within moral psychology. Moral Foundations Theory (MFT; Graham et al., 2013) argues in favour of these 'characteristic associations' between emotions and moral content, as does Constructive Sentimentalism (CS; Prinz, 2009) - both of which draw on the Community-Autonomy-Divinity model (Rozin et al., 1999). In contrast, the Theory of Dyadic Morality (TDM; Schein & Gray, 2018) argues against any such specific associations, and especially against exclusive associations - whereby a specific emotion relates *only* to its hypothesized moral domain. However, whilst an absence of such correspondences would be a relatively minor problem for Moral Foundations Theory, in that none of its key claims rely on the existence of such links, it potentially poses a major issue for Constructive Sentimentalism given this theory argues morals are constructed via emotions. In contrast, the presence of such emotion-content associations would present a serious issue for Dyadic Morality. If morality is all about harm, such that concerns regarding impurity are just transformations or intermediaries of harm concerns, then there should be no room for such correspondences as there is only one kind of moral content - dyadic harm.

Recent reviews of studies investigating correspondences between emotions and moral domains (Cameron, Lindquist, & Gray, 2015; Landy & Goodwin, 2015a) suggest evidence in favour of such specific correspondences is bordering on non-existent. Cameron et al. focus on links between emotions and moral content, claiming only one of the studies they reviewed (Seidel & Prinz, 2013a) finds evidence in favour of exclusive correspondences – although they argue this study, like many of the others they review, does not sufficiently control for potential confounds. Landy and Goodwin examine claims that moral judgements are amplified by incidental disgust, finding evidence for a small

overall effect (*d* = .11) – although they note this effect is larger (*d* = .37) for studies using disgust inductions involving direct channels into the body (i.e., via gustatory/olfactory channels). However, they also report this effect does not appear limited to 'impure' violations, and suggest the overall effect size is likely the upper limit for disgust amplification effects given the prevalence of potential confounds within the research. Furthermore, Landy and Goodwin (2015a) suggest that the effect may be eliminated entirely once accounting for publication bias. The findings of both these reviews pose problems for theories advancing emotion-content specificity hypotheses, as well as theories which propose affect plays a causal role in moral judgements.

Advocates for emotion specificity, and a causal role of affect in morality, may take a few different approaches in responding to these findings should they wish to maintain their theories. The first of these is to inspect and highlight potential issues within the reviews which may have influenced their findings, particularly as arguments advanced by Cameron et al. (2015) regarding the role of affect seem to conflict with aspects of Landy and Goodwin's (2015a) methodology. The second is to examine studies conducted after these reviews took place; as more recent investigations may provide new, and potentially more reliable, evidence which challenges the reviews' conclusions. The third way involves experimentation, testing whether support for links to emotion (still) appear once addressing or controlling for potential confounds. Each of these approaches is discussed separately in turn, and with regard to key theoretical positions - although there are important areas of overlap between the arguments presented in each section.

6.1. A Constructionist Review of Morality and Emotions

Cameron, Lindquist and Gray (2015) reviewed 25 published articles claiming links between emotions and moral content. They argue the majority of these fail to find evidence of an exclusive relationship between emotions and moral content, noting the frequent co-occurrence of both anger and disgust in response to moral violations of both harm and impurity. Furthermore, Cameron et al. argue that the few studies which do report evidence suggesting exclusive links suffer from a variety of methodological issues. For example, they argue forced-choice response methods and ANCOVA-based analyses eliminate shared variances and are thus overestimating the extent of exclusivity. Cameron et al. conclude that evidence for the 'loose correspondences' between emotions and moral content may be better explained by factors relating to core affect (e.g., valence, arousal) and overlaps with conceptual knowledge of emotions (e.g., contamination relates to aspects of both disgust and impurity) – and set out an experimental framework for testing links between emotions and moral content that takes account of the concerns they highlight.

Cameron et al. (2015) also contrast constructionist frameworks with 'whole number' accounts of morality and emotion - such as the CAD model (Rozin et al., 1999), Moral Foundations Theory (Graham et al., 2013), and approaches that favour 'basic' emotions. Cameron et al. explain that 'whole number' accounts argue mental states arise from the processes of many distinct 'encapsulated mental mechanisms', whereas constructionist accounts focus on the flexible combination of the same common elements in accounting for different mental states. For example, rather than having separate systems that process anger and disgust, both emotions may arise from shared elements of 'core affect' - both are high arousal, negatively valenced emotions; and may be differentiated by conceptual knowledge - anger relating to knowledge of offense, and disgust to concepts relating to contamination.

A constructionist approach can also be seen in the relationships between emotions and moral content argued for by the Theory of Dyadic Morality (Schein & Gray, 2018), where high and low arousal maps onto perceptions of agency and patiency while positive and negative valence relate to perceptions of help and harm respectively (Gray & Wegner, 2011; Cameron et al., 2015). Accordingly, the frequent co-occurrence of anger and disgust may be explained in that high arousal, negatively valenced emotions are elicited in response to villains because villains are agents (high arousal) which cause harm (negative valence). Similarly, a lack of evidence for exclusive links between emotions and moral content (Cameron et al., 2015) may be taken as evidence against 'whole number' accounts (MFT is given as an example), but is consistent with the Theory of Dyadic Morality (Schein & Gray, 2018) which argues that moral content (constructed from norms, affect, and harm) relates to varieties of perceived harm - where harm is itself constructed from perceptions of agency, patiency, and causality.

Cameron et al. (2015, p. 3) further state "Constructionism's emphasis on domain-general ingredients and common combinatorial processes leads to different predictions from whole number accounts about the origin of different emotions and moral content, and their relation to one another.". However, this is not necessarily the case. Constructive Sentimentalism (Prinz, 2009), a constructionist theory which draws on the CAD model (Rozin et al., 1999), advocates for links between emotions and moral content in a similar way to Moral Foundations Theory (Graham et al., 2013). Detailing the differences between these theories, with reference to the frequent co-occurrence of anger and disgust, helps illustrate potential issues with the underlying premises of Cameron et al.'s review; although to their credit, they do acknowledge alternative explanations.

"Another possibility is that disgust and anger co-occur when *a transgression violates multiple types of moral content*: For instance, acting unfairly, being disloyal, and disobeying authority could all be construed as harmful (Gray, Waytz, & Young, 2012) *or* impure (Batson, 2011). This line of reasoning could provide exclusive correspondences between morality and emotions, but *sacrifices discreteness of moral content*, and is therefore more consistent with a constructionist perspective. Alternatively, one could suggest that a specific kind of moral content elicits a primary emotion (e.g., injustice results in anger), and this primary emotion causes a secondary emotion (e.g., my anger makes me feel disgust toward the violator). While retaining discreteness of moral content, this possibility *sacrifices discreteness of emotions*." (Cameron et al., 2015, p.12, *emphases* mine).

6.1.1. Moral Content and Emotions

Cameron et al.'s (2015) review proceeds on the premise that there is 'pure' moral content - such that violations of a particular kind of moral content are construed as violations of that content type alone; and contrast these 'whole number' accounts of morality with the constructionist account they favour - the Theory of Dyadic Morality (Schein & Gray, 2018). However, describing Moral Foundations Theory (Graham et al., 2013) as a 'whole number' account suggests Cameron et al. (2015) may have either misunderstood what MFT means when it refers to foundations as 'modules', or are mischaracterising MFT as advocating for a much stronger form of modularity than it actually holds. MFT states that the theory is not reliant on a modular view of the brain (Graham et al., 2013), "nor is there any requirement that the adult mind contain five “distinct” or “discrete” modules (or even sets of modules) with no overlap." (Haidt, Graham, & Ditto, 2015).

Yet even if we allow for a strong form of modularity, such that the mental systems underlying each foundation are largely (but not fully) encapsulated, multiple-content transgressions do not necessarily sacrifice discreetness of moral content - although they do readily explain the co-occurrence of anger and disgust. Unfairness, disloyalty, and subversion could be categorized as harmful *or* impure – but they could also be construed as *both* harmful *and* impure, to the extent they violate norms governing both autonomy and the natural order (cf. Prinz, 2009). Indeed, individual differences in content classification suggest harmful actions may sometimes be considered impure, and impure actions as harmful (see Chapter 4), such that all types of moral transgression may elicit both anger and disgust. However, discreteness of moral content may be preserved. In operational terms, functionally specialized 'Harm modules' may handle the harmful content, eliciting anger, and 'Sanctity modules' may handle the impure content, eliciting disgust. That morally relevant actions may be classified as wrong with respect to more than one foundation is not an argument against the discreteness of moral content. 'Pure' moral content, in the sense of relying on discrete mechanisms, may be retained even if all moral transgressions happened to violate multiple types of moral content.

By way of example, infidelity - in the sense of 'cheating' on your partner, tends to be considered morally wrong. However, it has also been categorised as relating to 'loyalty/betrayal' (Landy & Bartels, 2018), ‘impurity’ (Gray & Keeney, 2015a), and ‘harm’ (see Chapter 4) concerns. Infidelity is indiscrete – the act itself may be seen as violating multiple types of moral content. How this moral content gets processed is a separate matter to whether multiple types of content are present, and there are multiple processing possibilities which may preserve moral foundations. It may be that any one of the foundations could generate a relevant moral intuition, such that whether infidelity is categorised as an instance of cheating, betrayal, impurity, or harm depends on contextual factors. It could also be that more than one foundation generates relevant intuitions, such that infidelity is an instance of some or all of the above (e.g., a harmful

betrayal). Also, processing may happen in serial, such that it draws on foundations in turn, or in parallel, such that different intuitions appear with seeming simultaneity - and may even compete for attention. However, Cameron et al. (2015) anticipates this line of response, stating “[i]n general, appealing to co-activation of moral content or emotions undermines whole number theory claims for independent, domain-specific mechanisms" (p. 12).

However, even if we discount Moral Foundations Theory as a ‘whole number’ account, specific correspondences between moral content and emotions have also been proposed within constructionist frameworks. Whereas the Theory of Dyadic Morality has to advocate against specificity given it proposes there is only one type of moral content (i.e., dyadic harm), Constructive Sentimentalism argues in favour of correspondences through its proposal that emotions play a constitutional role in morality. On this approach, a disposition to experience emotions of both self- and other- blame – to possess a sentiment of disapprobation - in relation to infidelity is what makes it wrong. For example, according to Constructive Sentimentalism (Prinz, 2009), the emotion of jealousy underwrites infidelity norms. "We are enraged when our trust is violated, frightened about facing competing suitors, saddened by the potential loss of a lover, and disgusted by the prospect a lover has been contaminated. Thus, these emotions invariably blend together when we have been romantically betrayed, and we use the term jealously to label that blend." (Prinz, 2009, p. 280). The elicitation of both anger and disgust in response to others' infidelity is entirely expected on this account, with anger being elicited by 'autonomy' concerns (violation of trust) and disgust by 'divinity', 'sanctity', or 'natural order' concerns (bodily contamination).

Constructive Sentimentalism further suggests that the co-occurrence of anger and disgust could also be due to second-order moral rules supporting first-order moral rules, rather than the alternative suggestion by Cameron et al. (2015, p.12) whereby primary emotions elicit secondary emotions at the cost of sacrificing discrete emotions.

First-order rules cover norms regarding how people should behave, whereas second-order rules cover norms regarding how people should feel. Thus, whilst people may indeed experience anger in response to injustice, this need not result in that anger eliciting disgust towards the violator – it could be the violators' apparent lack of shame or guilt about committing the injustice is what elicits the disgust. For example, consider the statement “I am angry at your actions, but I’m disgusted that you don’t feel guilty about doing something wrong”. The violator commits two wrongs, one regarding how they behave – which elicits anger – the other regarding how they feel (or do not feel) about that behaviour - which elicits disgust. Both emotions are elicited in response to moral transgressions, but they may be tracking transgressions at different levels - an approach which may share some common ground with arguments proposing disgust is elicited in response to judgements of moral character (e.g., Chakroff & Young 2015, Giner-Sorolla & Chapman, 2017).

6.1.2. Review Summary

In short, Cameron et al.'s (2015) reasoning and conclusions are heavily reliant on premises whereby moral actions contain an exclusive type of moral content. Actions may be harmful or impure, but not both harmful and impure. However, this 'whole number' premise does not feature in the approaches they mention as representing it, and constructionist theories can provide convincing arguments in favour of correspondences between discrete types of moral content and discrete emotions. There are several explanations available as to why both harm and impurity violations elicit both anger and disgust which do not preclude exclusivity, and discounting these possibilities substantially weakens Cameron et al.'s argument. Many of the studies they cite report either relatively proportional correspondences - whereby anger is more often associated with harm, and disgust more often with impurity, or report that anger and disgust co-occur similarly for both types of content. Yet the interpretation of this evidence rests on

underlying premises, and there are good reasons to suggest the ‘exclusive content moral actions’ premise is invalid.

Leaving aside the other potential explanations for Cameron et al.’s findings, multiple-content violations are commonplace. Many of the scenarios used by Shweder et al. (1997) are rated as concerning more than one moral domain, sometimes to varying extents; and scenarios used by Rozin et al. (1999) show considerable within culture variance in domain classification, with several showing classification of the predominant domain differs between cultures. Scenarios developed by Clifford et al. (2015) with the explicit aim of triggering one specific foundation are only rated as belonging to the relevant foundation around 75% of the time, suggesting sizable individual differences with regard to content categorisation - a finding echoed by the content classification rates in Landy and Bartels (2018) study. Participants in Gray and Keeney's (2015a) study provided examples of either 'harm' or 'impurity' violations, volunteering transgressions including murder, rape, adultery, theft, and drug use in response to both categories. Furthermore, scenarios common within the studies Cameron et al. review may depict multiple-content violations – for example, the plane crash scenario used by Schnall, Benton and Harvey (2008) includes an instance of both harm (killing) and impurity (cannibalism), but is often labelled and treated in research as being about impurity alone. Thus, whilst Cameron et al. (2015) may claim to show there is ‘no evidence for specific links between moral content and discrete emotions’, they could potentially make the claim that there is little evidence against specific links. Indeed, the typicality of multiple-content violations, and the apparent rarity of ‘pure’ single content violations, suggests their alternative explanation may be the more likely one. If anger and disgust co-occur in response to multiple-content transgressions, and the vast majority of scenarios used in research depict multiple-content transgressions, then the frequent co-occurrence of anger and disgust is readily explainable. However, contra Cameron et al., this co-occurrence need not sacrifice discreteness of moral 'foundations' or emotions - moral content may be (pre)mixed at 'source', but processed via

functionally specialized mental systems, or different emotions may be operating at different (moral) levels.

6.2. Does Incidental Disgust Amplify Moral Judgements?

Landy and Goodwin (2015a) offer a meta-analytic review of 50 studies focused on the impact of incidental disgust on moral judgements – investigating three related hypotheses. The first is the elicitation hypothesis – moral violations elicit disgust – for which they conclude there is a strong evidence base. The review also investigates the moralization hypothesis – that experienced disgust leads to condemnation of actions typically regarded as amoral – reporting that from the limited number of studies reviewed which bear on this hypothesis, there seems to be a small effect (d = .21) of disgust inductions on non-moral actions. However, the main thrust of the review evaluates the amplification hypothesis – that inducing disgust increases the severity of moral judgements – concluding that such an effect seems to be small (d = .11), but is not limited to violations of purity, and may be non-existent once accounting for publication bias. Furthermore, Landy and Goodwin suggest that, given confounds are prevalent, the effect size they report may be interpreted as an upper bound on the amplification effect.

These findings, prima facie, may be taken in support of Cameron et al.'s (2015) review as they show a non-exclusive effect for disgust. However, the methodology employed by Landy and Goodwin (2015a) seems to conflict with Cameron et al.'s assertions regarding the role of emotions in moral judgement - such that if Cameron et al. are correct about non-exclusivity, then Landy and Goodwin may have underestimated their reported effect sizes. In particular, Landy and Goodwin treat sadness inductions (and potentially other emotion inductions, such as fear) as control conditions – and compute effect sizes based on comparisons between the disgust condition and 'control' conditions. Accordingly, if there were any amplification effects for sadness (or fear)

inductions on moral judgements, then comparing these with disgust inductions would produce a relatively smaller effect size for disgust amplification than comparing solely with neutral/non-emotion induction conditions.

The conflicting positions of these reviews can be illustrated with reference to Cheng, Ottati, and Price (2013), who are cited in both. For Cameron et al. (2015), Cheng et al.'s results provide evidence favourable towards assertions against specificity between moral content and emotions, showing inductions of anger, disgust, fear, and sadness all have similar impacts on moral judgement ratings. Indeed, Cameron et al. propose fear and sadness are necessary control conditions for moral-emotion specificity studies, as these emotions share important features with anger and disgust – all are negatively valenced and, with the exception of sadness, high arousal emotions. According to Cameron et al., it is these elements of shared 'core affect', in combination with conceptual knowledge of emotion categories, which may explain the apparent lack of moral-emotion correspondences. In contrast, for Landy and Goodwin (2015), Cheng et al.'s results only contribute a minimal effect size ($d < .1$) regarding the influence of disgust on moral judgements, despite all the other emotion inductions reporting statistically similar effects in comparison to the control condition. Accordingly, if all emotion inductions have equivalent effects, then collapsing sadness and neutral conditions into a single control may have reduced the reported effect size for Cheng et al.'s study significantly, and even more so if the fear condition were included in the collapse.

Additionally, in apparent contrast to Cheng et al.'s (2013) findings, Horberg et al. (Study 2, 2009) report that inducing disgust led to greater condemnation of purity violations, and greater praise of purity virtues, in comparison to inducing sadness, whereas ratings concerning violations of harm and virtues of care did not differ significantly between the two conditions. If sadness induction increases ratings of moral severity (as per Cheng et al., 2013), then the effect size calculated for Horberg et al.'s

study also may be understated given the comparison with a non-neutral condition. Furthermore, in splitting content types during analysis, Landy and Goodwin's (2015a) results suggest Cheng et al.'s (2013) second study may actually provide evidence favouring specificity (contra Cameron et al., 2015), reporting an effect size of *d* = .38 for disgust induction on purity violations in contrast to a *d* = .06 effect size for non-purity violations. Although, as Schnall et al.'s (2008) multiple-content 'plane crash' scenario is the only instance of a purity violation used by Cheng et al., the validity of this suggestion is questionable.

6.2.1. Additional Confounds

Closer inspection of the studies included in Landy and Goodwin's meta-analysis (2015a, from Table 1.) identifies the use of multiple content violations when depicting instances of impurity, and Schnall et al.'s (2008) vignettes in particular, as potential confounds - affecting both effect size estimates, and the extent to which conclusions might be generalised. For example, Zhong, Strejcek and Sivanathan's (2010) research contributes two of the largest negative effect sizes from the literature reviewed by Landy and Goodwin (2015a), with a study that is treated as employing only impurity-based judgements. Leaving aside issues with treating induced cleanliness as a control condition, there are potential problems with at least half of the 'social issues' examined in their studies. Adultery and drug-use are rated as multiple-content violations by lay participants (Gray & Keeney, 2015a, Study 2), and the latter example presumably covers specific instances of drug use, such as smoking and alcoholism. Similarly, abortion is likely another multiple-content issue, as may be the wearing of animal fur - as participants might readily perceive associations with 'harm' in these matters. It is also unclear that use of profane language, by itself, is a violation of purity. Additionally, 'obesity' and 'masturbation', and to a lesser extent, 'homosexuality' and 'premarital sex', are trailing the floor of the ratings scale and so providing no signal for amplification.

Inclusion of these instances may affect the overall size of the effect reported via Zhong et al. (2010); multiple content violations may receive limited amplification if disgust operates selectively on impure content, and 'floored' scenarios contain little to amplify.

Furthermore, a substantial number of the reviewed studies also draw frequently on the scenarios used by Schnall et al. (2008). This means these scenarios may exert significant influence on the meta-analytic outcome, and reservations are readily raised about all three of Schnall et al.'s (2008) impurity scenarios. The plane crash, in addition to depicting multiple content, may be impacted by ceiling effects. Regardless of any mitigation within the scenario, child killing seems like a candidate for an action which is 'maximally morally wrong' (cf. Schein & Gray, 2018) - as well as potentially involving mixed moral content in itself, before compounding (and confounding) the situation by eating the child's remains. In contrast, eating dog meat is framed within the scenario as a conventional transgression (i.e., it is not universally wrong), which may mitigate any moral judgement - despite such an event also being potentially subject to disgust ceiling effects (cf. Wheatley & Haidt, 2005). Also, the third of their impurity scenarios, involving a boy rubbing his aroused appendage along a kitten, is particularly weird - to the point of needlessly pushing the bounds of ecological validity. A scenario involving a dog cleaning up an indelicate peanut butter spillage may have been better suited in this regard.

Schnall et al.'s (2008) 'non-purity' scenarios also contain potential confounds. The resume scenario contains an element of violating the natural order of persons (cf. Prinz, 2009), as well as being unfair, both of which may elicit some disgust (Prinz, 2009; Schnall, 2017) - and thus provide a signal to amplify. The wallet scenario also contains undertones of fairness, framing inequality and need as potentially mitigating reasons to keep the money. Furthermore, the trolley scenario is somewhat different to the other scenarios, which may impact on its wrongness ratings. All scenarios ask, 'how wrong was (the 'immoral' action) X?', but differ in what constitutes 'Not X'. The kitten-rubbing and dog-eating impurity scenarios are simple, in that not doing X has no obvious ill

effects; and for the wallet and resume non-purity scenarios, not doing X has minimal negative outcome, whereas for the plane crash and trolley scenarios, not doing X involves highly negative outcomes. In the plane crash, not killing and eating the child may lead to three deaths - but this outcome is described as probable, rather than certain, and there is substantial ecological wiggle room available. Ignoring notions that two adults might feasibly be able to carry a boy with a broken leg, and possible rescue missions aside, the village is several days away - yet humans may survive a few weeks without food, and even if this is scaled down substantially to factor in the extreme cold, and days already travelled, there may still be enough time left to save the boy. Although 'Not X' probably has a negative outcome, the scenario may suggest its action takes place 'too soon'. In contrast, for the trolley scenario, both doing X and not doing X have a certain negative outcome - regardless of whether hitting the switch is considered wrong. Either one person is killed (by commission) or five people are killed (by omission), and killing may well be considered a multiple content violation. Furthermore, hitting the switch may be seen as the moral action (saving the most lives), such that the question should be 'how wrong is it for you to NOT hit the switch...?', to keep it in line with the questions for other scenarios.

Accordingly, if confounds within Schnall et al.'s (2008, Study 2) impurity scenarios lead to the understatement of effect sizes, then Landy and Goodwin's (2015a) meta-analytic result would also be understated. However, if scenario confounds tend towards overstating the effect, then there may not actually be any effect. Landy and Goodwin's (2015a) review suggests published studies report a greater effect size than non-published ones, and nine of the fourteen published studies reviewed draw on Schnall et al.'s scenarios. The considerable variation in the effect sizes reported provides little clarification on matters. Focusing specifically on disgust amplifying impure judgements (Landy & Goodwin, 2015a, from Table 1.): Cheng et al.'s (2013) contribution to the meta-analysis should be treated with caution given this is based solely on the plane crash scenario. Schnall et al.'s (2008) second and third studies report minimal

effects (d < .1), whereas their fourth study shows a moderate effect (d = .38), despite all using the same stimuli. Schnall et al. (2008) also find a moderate effect (d = .34), but one which Johnson, Cheung and Donnellan (2014) did not find when replicating their study. Ugazio, Lamm and Singer (2012) also report minimal effects (d < .1), although their materials either do not fully report the impure scenarios used, or treat all scenarios used by Schnall et al. (2008) as disgust-related, making validity judgements difficult. In contrast, Seidel and Prinz (2013a) use a modified (less confounded) set of Schnall et al.'s scenarios and report a large effect size (d = .92) - although Cameron et al. (2015) would argue the size of this effect reported is likely due to experimental confounds.

The remaining published studies reviewed with regard to impure judgements also provide conflicting evidence. Zhong et al.'s (2010) findings might be excused as a result of confounding, removing the two largest negative effects from the published literature; and Schnall et al.'s (2008) pilot study may be discounted for being too disgusting. Horberg et al.'s (2009) contribution may also be understated given it uses sadness as a control condition, and that some of the scenarios used may be rated at floor. These points would each suggest the effect size may be higher than Landy and Goodwin (2015a) report. However, the study by Eskine, Kacinik and Prinz (2011), which contributes the largest effect size favourable to the amplification hypothesis (d = 1.18), must also be excused. Ghelfi et al.'s (2020) multi-lab replication of this study suggests the extent to which ingesting a bitter drink increases moral condemnation is within the bounds suggested by Landy and Goodwin (2015a); and in contrast to Eskine et al., Ghelfi et al. find that this (small) effect was also present when comparing the sweet drink and control conditions. This leaves only Schnall et al.'s (2008) first study, and the two studies by Wheatley and Haidt (2005), as the remaining three of fourteen published contributions to examine the effect without including the confounds discussed above. However, these do not account for other potential confounds identified by Landy and Goodwin (2015), and their effects may still be outweighed once taking account of the effect sizes from unpublished studies.

6.2.2. Untested moderators

Schnall et al. (2015) take issue with Landy and Goodwin's (2015a) conclusions, arguing these findings support those reported in Schnall et al. (2008). Specifically, that the un-moderated effect of incidental disgust on moral judgement is minimal, that olfactory inductions are particularly effective - such that these often produce moderate effect sizes without recourse to moderating variables, and that disgust amplification effects are not specific to violations of purity. Although this last finding may be the result of the aforementioned content-relevant confounds in the scenarios used. Schnall et al. (2015) highlight three factors as potentially responsible for the differing conclusions drawn; a key moderator is not included, olfactory induction results are under-emphasized, and experimental confounds may nullify the effect being investigated. Each factor is worth noting as all bear on particular issues under investigation, but also because claims regarding experimental confounds and key moderators advanced by Schnall et al. (2015) are argued to conflict with each other (Landy & Goodwin, 2015b).

Firstly, Schnall et al. (2015) claim the effect is limited to participants who scored highly on the Private Body Consciousness Scale (Miller et al., 1981), a measure of sensitivity to bodily sensations, which is not included as part of the meta-analysis. However, this is not examined due to the limited number of studies which use such a measure (Landy & Goodwin, 2015a, b), even if it does suggest an upwards revision of the effect size once this moderator is accounted for. Secondly, Schnall et al. (2015) claim that Landy and Goodwin (2015a) understate the efficacy of olfactory induction methods. However, Landy and Goodwin (2015b) suggest such methods may be prone to confounds, in that the experimenter may provoke offense as a result of inducing disgust in participants - and this offense may explain, or at least contribute to, any amplification effect. Thirdly, and perhaps unusually, Schnall et al. (2015) claim that

studies in this area 'require' participants to mis-attribute the source of their emotions, such that if feelings of disgust are attended to prior to considering the act, they are unlikely to be considered as having arisen in response to the act - which may nullify the effect, or even reverse its direction.

Landy and Goodwin (2015b) suggest there may be some merit in Schnall et al.'s (2015) third claim, finding studies with emotion tasks prior to moral ratings show no amplification effect, whereas those without emotion tasks show some level of effect (*d* = .18). However, they note this seems to be in opposition to their first claim. It suggests attending to emotional states in advance works to suppress amplification effects, despite a general tendency to attend to such states being crucial for such effects to emerge. Furthermore, it is not clear that misattribution is strictly required. Seidel and Prinz (2013a) had participants focus on sounds which typically induce anger or disgust for one minute prior to giving ratings, and these sounds continued during the ratings phase, which is not a surreptitious induction method (cf. fart spray, unkempt experimental space) - yet this study finds a large effect size for disgust amplification over judgements of impurity, and also that anger induction amplified condemnation of autonomy violations. Sidestepping concerns about the disgust-impurity amplification in this study (cf. Cameron et al., 2015), participants may easily be attributing the induction as the cause of their negative affect - although they may be unable to negate its influence due to its persistence. It would be interesting indeed if simply administering an induction check prior to scenario rating curtailed any resultant effect size, as it seems like participants should be quite aware that the sound is the cause of any negative affect present before they rate the scenarios - so it is not clear why asking them to confirm this should make any difference under such circumstances.

6.2.3. Review Summary

In short, Landy and Goodwin's (2015a) review of the available evidence suggests that incidental disgust amplifies moral judgements to a minimal extent at best, and that this effect does not seem specific to violations of impurity. However, closer inspection of relevant meta-analytic input factors highlights an array of potential confounds which may interfere with estimates of effect sizes, particularly with regard to content-specificity estimates. Operationally, treating sadness inductions as control conditions may underestimate effect sizes if sadness exhibits a similar amplification effect to disgust (e.g., via shared valence and/or arousal - cf. Cameron et al., 2015), as it does in Cheng et al. (2013). Also, estimates of effect size may further increase if sufficient data were available regarding participants' sensitivity to bodily sensations, argued to act as a crucial moderator on the amplification effect (Schnall et al., 2015). Yet addressing the first of these issues may only yield a marginal increase, and the second would do little to alter conclusions drawn about un-moderated effect sizes (Landy & Goodwin, 2015b). However, effect size estimates are likely influenced to a greater extent, and in opposing directions by content, through confounds present within the studies under review.

To start with, some scenarios may get rated at the floor of their scale, providing a limited signal for amplification. Also, some scenarios may get rated at ceiling, either because the act is considered highly immoral (e.g., killing), or because the act may be highly disgusting (e.g., dog-eating), but both placing constraints on detecting the effect. This could either be through leaving no room to increase ratings of wrongness - any amplification cannot be measured, or through leaving no room for the induction to operate if the scenario elicits maximum disgust - the signal cannot be amplified further. The presence of floor and/or ceiling effects would suggest an upward revision to effect size estimates. Furthermore, the frequent presence of multiple-content scenarios may influence estimates regarding content-specificity, particularly if disgust behaves in a

content specific manner. Including multiple content measures for impurity may suppress any effects, whereas using multiple content measures for harm may promote effect emergence. This could revise estimates for the effect of disgust on impure judgements upward, whilst revising effect estimates over non-purity judgements downward, with the emergent outcome potentially supporting claims favourable toward specific correspondences between emotions and moral content.

Importantly, this potential divergence of content effect sizes remains plausible when factoring in potential confounds highlighted by Landy and Goodwin (2015a). Exposure to the induction materials may produce negative affect by various means, which may lead to the effect of disgust being overestimated - but even if this were to balance out with potential confounds which may lead to underestimation, induction exposure confounds would need to operate in a highly specific manner to balance with potential content-specificity confounds. If the issue is just general negative affect, it should affect all content types similarly; and if participants experience a form of nausea, it is not immediately clear why this is an issue - any amplification would still be based around (incidental) disgust. In contrast, if participants take offense or feel they have suffered a minor harm, then any amplification may not be attributable to disgust, but rather to anger elicited by the offense/harm. Yet if content specificity claims are correct, this anger should act on non-purity scenarios, which would work to further suppress the emergence of content-specificity within the results. As such, whilst there may be a little room to doubt claims about the upper bounds of the amplification effect, there seems to be good reason to doubt claims that (incidental) disgust amplifies moral judgements without regard to content type – specific correspondences between emotions and moral content cannot be ruled out, and the existence of these may be suppressing the size of the amplification effect.

6.3. Recent Evidence

Research conducted since these reviews were published, and with knowledge of either review (according to Google Scholar citation metrics in 2019), provides further evidence relating to correspondences between emotions and moral content. This research is addressed in three parts. Part I covers research which fits with Landy and Goodwin's (2015a) conclusions, and research thought to fit with Cameron et al.'s (2015) conclusions - such as that from Kollareth and colleagues, as well as research relating to similarities between anger and disgust. Part II covers research suggesting anger and disgust serve specific functions, and research into the effects of emotion sensitivity traits (primarily disgust) on moral judgements. Part III covers research focused on correspondences between emotions and moral content, all of which may be taken to oppose Cameron et al.'s (2015) claims against specificity.

6.3.1. Part I

Johnson et al.'s (2016) study addresses the claims regarding untested moderators raised by Schnall et al. (2015). Over two replication attempts of Schnall et al.'s (2008) third study, Johnson et al. (2016) report finding no evidence for a disgust amplification effect, although they did find a main effect whereby those reporting higher Private Body Consciousness scores (PBC; Miller, Murphy & Buss., 1981) tended towards providing harsher moral judgements. However, they report no interaction between PBC and the induction conditions in either study, and the only two interactions they report in their second study are also in opposition to the hypothesized effects. Asking about mood before or after rating the scenarios showed no main effect on judgements, and even if not accepting the null hypothesis for the three-way interaction - this shows participants not responding to mood items before scenario rating only made harsher moral judgements if they were *less* attentive to their bodily states. All these

findings run contrary to arguments presented by Schnall et al., (2015) with regard to amplification effects.

Wisneski and Skitka's (2017) research might be taken as supporting Landy and Goodwin's (2015a) conclusion that *incidental* disgust only has a minimal effect (at best) on judgements of immorality. Wisneski and Skitka (2017) conducted two studies examining the effects of disgust cues on moral conviction. They found only the supraliminal cueing of target relevant disgust produced an effect on judgement ratings - showing pictures of an aborted foetus increased moral convictions regarding abortion. In contrast, inducing disgust either subliminally, or disgust which was unrelated to the target, produced no such effects when the induction images depicted 'pure disgust', animal abuse, or incidentally disgusting harms to humans. These results suggest it is rather non-incidental disgust which increases ratings of moral conviction – and increased moral conviction regarding 'X' implies increased severity ratings for violations of 'X'. This finding might be taken to support the role of conceptual knowledge in moral judgement as proposed by Cameron et al. (2015). However, Wisneski and Skitka (2017) also report that ratings of disgust, rather than ratings of anger or appraisals of harm, fully mediated the effect of target-relevant disgust. They note this finding seems opposed to approaches which contend morality centres on concerns about harm (e.g., Schein & Gray, 2018).

Kollareth and colleagues provide mixed evidence regarding correspondences, having taken aim at both the CAD model (Kollareth & Russell, 2017; Kollareth, Kikutani, Shirai, & Russell, 2019), and specific links between disgust and impurity (Kollareth & Russell, 2018; 2019). Kollareth et al. (2019) report participants responded similarly to violations of autonomy and community, selecting anger more frequently than contempt for both types of violation. Yet their stimuli actually try to keep the transgression constant by changing the target of the transgression from one individual to many. Although interesting, this does not seem to be a valid approach. Moral domains are defined by the

type(s) of norm being violated (Prinz, 2009), rather than by the number of people who may suffer as a result of the action. As such, Kollareth et al. (2019) seem to be comparing autonomy violations involving single or multiple victims, rather than with community violations. Similarly, Kollareth and Russell (2018) suggest prototypical purity violations are self-directed immoral actions, finding sadness, rather than disgust, is associated with such actions. However, the vast majority of the materials used involve instances relating more to self-harming behaviour than self-directed impure behaviour, and those remaining do not always adapt well between conditions. For example, consensual incest is likely considered quite differently to incestuous rape. Both these studies contain confounds in key aspects of the methodology, such that their conclusions can provide no evidence for or against correspondences.

In contrast, using better materials, Kollareth and Russell (2017) show autonomy and community violations both primarily elicit anger, but state that disgust seems limited to instances of impurity involving sex or pathogens. However, this latter claim is only limited because participants appeared to consider child abuse as (primarily) an instance of harm, which is perhaps unsurprising given how the scenario is written. Suppose “[o]ne day you find out that one of your acquaintances had been enslaving children for sex trade for the last 7 years” (Kollareth & Russell, 2017, from Table 6. Appendix 1.). Persons enslaving children for sexual exploitation may be regarded as sub-human, and such degrading behaviour may be considered a violation of purity, but it seems highly implausible that this scenario depicts only impurity. Enslavement itself may be considered as harmful, particularly as those inflicting such deprivation of liberty seem likely to employ harmful methods to enforce it over time. As such, the study actually does show links between 'autonomy' and anger, and 'divinity' and disgust. The lack of strong links between community and contempt is less important. The community domain may be considered as derived from the autonomy and divinity domains, and contempt derived from anger and disgust (Prinz, 2009), making the apparent similarity with autonomy violations readily explainable. These results may actually be taken in favour of

specificity, as the predominant responses are what correspondence accounts hypothesize. However, the results from this study are inconclusive with regard to the exclusivity of correspondences due to frequent co-activation of emotions.

More convincing new evidence against exclusive correspondences is provided by Kollareth and Russell (2019), who compared emotion responses to sacred and non-sacred events in the presence or absence of pathogens. This provides a strong test, as ‘sacredness’ may be seen as a relatively ‘pure’ instance of foundational moral content, and the presence of pathogens across many commonly used purity violations may account for much of any disgust-impurity correspondences. Kollareth and Russell (2019) found that sacred violations elicited responses of anger, and disgust, but not 'grossed-out' - which appeared to be a response specific to pathogen presence. Additionally, they found both anger and disgust predicted immorality ratings, whereas 'grossed-out' did not. Accordingly, they suggest 'core' disgust (i.e., grossed-out) is linked to pathogens, whereas 'moral' disgust seems to operate in a similar manner to anger – a finding which echoes those of Royzman et al. (2014). Given their scenarios co-activate both anger and disgust, as well as other 'negative' emotions (e.g., sadness), Kollareth and Russell (2019) interpret these results as favouring Cameron et al.'s (2015) position against specific and exclusive correspondences.

The apparent similarity between moral disgust and anger also finds support following a different line of research. Scott, Inbar, and Rozin (2016) suggest absolutist opposition to genetically modified foods is linked to disgust - a finding expected given such opposition is hypothesized to stem from purity concerns. However, although Royzman, Cusimano and Leeman’s (2017) research concurs that trait disgust predicts opposition to genetically modified food, and also extends this finding to other 'new' technologies (e.g., stem cell research, nuclear power), their study suggests feeling 'creeped out' (i.e., fear) better predicts opposition to genetically modified food. Royzman et al. (2017) finds that although being 'disgusted' is linked with opposition, this link

disappeared when disgust was operationalised as feeling 'orally inhibited' (e.g., nausea, gagging, loss of appetite). They state, "even in the context of ostensibly pure “pathogen-linked events”, “disgusting” is as likely to refer to fear, disapproval, and epidermal discomfort as it is to refer to oral inhibition proper" (Royzman et al., 2017, p. 472) - which is to say there may be substantial disconnect between theoretical and lay usages of “disgust”. However, the link between disgust and genetically modified food might be preserved, as the scenarios used by Royzman et al. (2017) are designed to elicit 'physical disgust' - and this form of disgust is thought to share features with fear, rather than 'moral disgust', which is thought to share features with anger (Lee & Ellsworth, 2013).

Oaten et al. (2018) find further support for the claim that moral disgust shares features with anger, and also suggest that neural overlap between ‘core’ and ‘moral’ disgust stems from overlapping content in the eliciting stimulus (i.e., core disgust), rather than moral violations per se. Oaten et al. (2018) piloted a range of scenarios, the results of which showed moral disgust scenarios were rated as more disgusting than scenarios depicting more potent core disgust elicitors - ‘high’ disgust (e.g., someone vomiting profusely). The results also showed these ‘high’ disgust scenarios were rated as more disgusting than those depicting ‘matched’ disgust – scenarios which contained the same disgust elicitor as the ‘moral’ disgust scenarios (e.g., a dead rat), but with a substantially curtailed or ‘absent’ moral element (e.g., someone finds a dead rat in their kitchen vs. someone puts a dead rat in their neighbour’s kitchen). Yet when investigating neurological responses (fMRI) to these scenarios in their main study, Oaten et al. (2018) found that the moral disgust scenarios did not produce any ‘additional disgust’, suggesting that much of any disgust response to moral scenarios may be attributable to the presence of core disgust elicitors.

Oaten et al. (2018) note the contrasts between both ‘moral’ and ‘high’ conditions with the ‘matched’ condition show different activation patterns, and that there were no

common activations when contrasting high disgust with matched disgust, or moral disgust with high disgust. These results suggest that whatever accounts for the greater disgust ratings for moral scenarios differs from whatever accounts for the greater disgust ratings given to more potent core disgust elicitors. Furthermore, Oaten et al. (2018) claim that once neural activation attributable to core disgust was accounted for, the residual pattern of activation was more similar to that produced by 'moral anger' than one of 'residual disgust'. Oaten et al.'s (2018) state they "…believe this delivers compelling evidence against the idea that moral violations can themselves be disgusting in the same way that core elicitors can be." (p. 12).

This finding further suggests that any correspondences between disgust and impurity may be due to the presence of core disgust elicitors (e.g., pathogens) within the scenarios, rather than *moral* content (i.e., impurity) specifically – a finding which echoes the position advanced by Kollareth and Russell (2019). Furthermore, the similarities between 'moral' disgust and anger might be taken in support of a more harm-centric approach to moral judgement. Ratings of harm given in the pilot study for moral disgust scenarios were similar to those provided for moral anger scenarios, whereas ratings of impurity and justice differed as expected between scenario types. Also, on inspection of Oaten et al.'s (2018) scenarios it appears the main differences between the 'moral' and 'matched' scenarios are related to the presence of dyadic elements (e.g., intention). Taken together, 'moral' disgust minus 'matched' disgust leaves behind activation patterns which appear more similar to moral anger at the neurological level, and leaves behind dyadic elements at the scenario level. This might be taken to support claims that 'moral' is linked with 'anger' and 'dyads' (cf. Royzman et al., 2014), whilst suggesting 'disgust' is arising in response to particular eliciting stimuli, rather than moral content per se. These results seem highly favourable towards both Cameron et al.'s (2015) claims, and the Theory of Dyadic Morality (Schein & Gray, 2018) more generally.

6.3.2. Part II

However, although there are similarities and overlaps between anger and disgust, these emotions are not equivalent. Both Molho et al. (2017) and Tybur et al. (2019) argue anger and disgust respectively relate to direct and indirect aggressive responses. Molho et al. (2017) find ratings of anger were greater for violations targeting the self (i.e., the rater), whereas ratings of disgust were greater for other-targeting violations. They also report self-targeting violations elicited more direct aggression, although both self- and other-targeting violations elicited similar levels of indirect aggression. However, Molho et al. (2017) further report that the emotion links seem functionally specialized, as anger was not related to indirect aggression, nor was disgust related to direct aggression. These findings are echoed by Tybur et al. (2019), who report similar results in their replication of Molho et al.'s (2017) fourth study. Self-targeting violations elicited more anger than other-targeting violations, other-targeting violations elicited more disgust than self-targeting violations, and each emotion was only related to direct or indirect aggressive responses respectively. As the authors state, these findings seem more favourable towards Moral Foundations Theory (Graham et al., 2013) than the Theory of Dyadic Morality (Schein & Gray, 2018).

Further differences along self-other dimensions are demonstrated by Dungan, Chakroff, and Young (2017). Across a range of harm and purity violations, they find perceptions of purity are better predictors of moral judgments for non-dyadic (i.e., self-directed) violations, whereas perceptions of harm consistently predicted moral judgements of dyadic violations. They also find "judgments of purity violations, compared to harm violations, are relatively more sensitive to the negative impact perpetrators have on themselves versus other victims" (p. 1). Similarly, Chakroff and Young (2015) found impure actions tended to elicit more person-based (over situation-based) explanations than harmful actions, even when specifying situational factors which led to the action. These findings imply disgust should also be linked with the 'negative impact perpetrators

have on themselves' or 'person-based explanations' via its relationship to the purity domain. Evidence in favour of this implied link is provided by Giner-Sorolla and Chapman (2017), who show disgust relates to information regarding 'moral character' regardless of content type, whereas anger relates to information about actions (e.g., wrongness, consequences). In showing that 'purity' may explain certain aspects of moral judgement better than 'harm', these studies are all favourable towards theories advancing moral-pluralism (Graham et al., 2013; Prinz, 2009) over those advancing harm-pluralism (Schein & Gray, 2018), and may also help partially explain the frequent co-occurrence of anger and disgust. These emotions may each serve specific functions, whereby each relates to different information arising from the same (stimulus-eliciting) action.

Rottman, Young and Kelemen (2017) investigated how children moralize novel actions, which may be conceptually classed as relating to (im)purity (i.e., ostensibly victimless, and focused on 'the body' or 'nature'), by asking them to make judgements about the behaviours of aliens on another planet. They found that inducing incidental disgust had no significant effect on children's moral judgements in comparison to the control condition - a finding in concurrence with Landy and Goodwin's (2015a) review. Rottman et al. (2017) also found principle-based testimony (e.g., about 'harm', 'unfairness', or 'weirdness') resulted in more severe judgements, and this effect was initially strongest for interpersonal testimony - a finding which may be taken in support of Dyadic Morality (Schein & Gray, 2018); although when re-testing the children after three months this effect had attenuated to that of non-interpersonal testimony. Similarly, providing emotion-laden testimony about the actions also amplified the severity of the children's judgements, albeit to a lesser extent than principle-based testimony. Here, Rottman et al. (2017) found both anger and disgust-based testimony produced similar increases in wrongness ratings - a finding in concurrence with Cameron et al.'s (2015) review. However, they caveat this support for 'domain-general' accounts by noting preliminary evidence favouring domain-specific trait sensitivity in their supplemental

materials. Children with a disposition towards anger were more sensitive to anger-based testimony, and children with high disgust sensitivity were similarly influenced by disgust-based (but not anger-based) testimony.

Staying within this research area, Wagemans, Brandt and Zeelenberg (2018) report a series of five studies, using standardised scenarios that cover a range of foundations (Clifford et al., 2015), which show 'Disgust Sensitivity is Primarily Associated with Purity-Based Moral Judgments'. Similarly, in a pre-registered study using harm and purity scenarios also drawn from Clifford et al. (2015), Liuzza et al. (2018) find that 'Body Odor Disgust sensitivity predicts stronger moral harshness towards moral violations of purity'. Both these studies link disgust with impurity. However, Van Leeuwen et al. (2017) finds sensitivity to different types of disgust relates to stronger endorsements of differing moral foundations. Pathogen disgust had small predictive effects for the 'binding foundations' of MFT. Sexual disgust had the largest predictive effects for endorsements of 'purity', although this was also predictive of scores for all foundations except 'fairness'. Similarly, moral disgust predicted scores across all foundations, except here the effect was largest for endorsement of the 'fairness' foundation, followed by the 'harm' foundation. It is noteworthy that sexual and pathogen disgust seems to predict 'binding foundation' endorsements, but that moral disgust seems predictive of endorsements across five foundations – being most predictive of 'individualizing foundations' endorsement. Moral disgust appeared more related to violations of autonomy - hypothesized to elicit anger, whereas non-moral disgust appeared more related to non-autonomy violations. This further suggests disgust may be elicited by certain content, but not necessarily *moral* content, such that associations between disgust and impurity may be due to the common presence of non-moral disgust elicitors in impure scenarios rather than the moral impurity of the violation in question. Thus, whilst the results from Wagemans et al. (2018) and Liuzza et al. (2018) may be taken to support specific correspondences between disgust and impurity, Van Leeuwen et al.'s (2017) results lend credence to assertions that moral disgust may share features with

anger (e.g., Lee & Ellsworth, 2013, Oaten et al. 2018) - which can be taken in support of claims against the exclusivity of correspondences (Cameron et al., 2015).

Furthermore, Landy and Piazza (2019) show that trait sensitivity to a range of affective states relates to judgement extremity of both moral and conventional transgressions, and also that disgust sensitivity relates to the extremity of evaluative judgements in general. They found sensitivity to sadness, irritation, and general negative affect, was associated with normative judgements in a similar way to disgust sensitivity - individuals more affectively sensitive to these emotions provided harsher judgements of both moral and conventional transgressions. Additionally, greater disgust sensitivity was related to more extreme judgements of moral violations, conventional transgressions, imprudent actions, competence, aesthetics, and notably - morally positive actions were also met with greater praise. These findings are consilient with Cameron et al.'s (2015) claims that affect operates on moral judgement via domain general mechanisms.

6.3.3. Part III

Franchin et al. (2019) employ Gray and Keeney's (2015a) materials to investigate links between emotions and moral content. They measured the facial expressions of participants elicited in response to audio recordings of scenarios, in addition to asking them to rate the scenarios along various dimensions. Franchin et al. state they do not find direct evidence of what they term '*strong* MFT' - the expected emotional expression is the most frequent in responses *across* content, although they do find limited support for '*weak* MFT' - the expected emotion being more frequent in responses *within* content. Interestingly, their results almost completely replicate Gray and Keeney's findings with regard to scenario ratings - including that MFT harm scenarios were rated as more impure than MFT purity scenarios. However, their results for facial expressions reinforce previously advanced critiques of Gray and Keeney's

(2015a) research, and although the study is described as supporting harm-centric accounts of moral judgement (e.g., Dyadic Morality; Schein & Gray, 2018), it cannot rule out 'moral modules' (in the MFT sense) given the appearance of differential links between emotional expressions and moral content, nor can it rule out the exclusivity of such associations.

Of the expressions investigated, Franchin et al. (2019) report anger was most frequently elicited in response to harm, but that smiling, rather than expressions of disgust, contempt, or surprise, tended to be elicited in response to violations of purity. Splitting the results by scenario source revealed MFT harm scenarios elicited more anger and contempt expressions, whereas MFT purity scenarios elicited more smiles and disgust expressions. However, over the 'naturalistic' scenarios, the only difference between types of moral content was that purity violations elicited substantially more smiling than harm violations, whereas anger was elicited at comparable levels by each type of content, as was disgust, and contempt. Yet that smiling in response to violations of purity occurred across both MFT and 'naturalistic' scenarios suggests this may be a domain-specific response which is not attributable to the relative weirdness of impurity scenarios; although follow up analyses by Franchin et al. suggest that some participants did find some purity violations amusing - as indicated by the presence of 'Duchenne smiles'. Furthermore, a follow-up study using self-report measures of emotion indicated disgust was selected most frequently for MFT impurity scenarios, whereas contempt was selected most frequently for 'naturalistic' impurity scenarios. These findings might be taken in support of the assertion that smiles elicited in this setting are indicative of 'covert disgust' - the face may show amusement, but self-report measures tended toward disgust.

Importantly, all Franchin et al.'s (2019) results relating to emotional expression at scenario level seem to be better explained via Constructive Sentimentalism (Prinz, 2009). The results' apparent support for harm-based models of moral judgement seems

overly reliant on the presence of 'naturalistic' scenarios - as results over MFT scenarios may be interpreted as largely conforming to 'weak MFT' predictions. Of the two MFT harm scenarios where anger expressions were not selectively elicited, anger is either comparable with contempt - because making cruel and offensive remarks about appearance may be construed as a violation of the perceived 'natural order of persons'; or comparable with smiling (covert disgust) - because killing ants may be seen as a violation against the perceived 'natural order'. Both these assertions find some support in the follow-up self-report study. 'Ant killing' primarily elicited a 'grin-and-bear-it' (i.e., purity aligned) response, and the 'remarks' were the second most frequent contempt elicitor after 'sticking a pin into an unknown child' - a scenario which may also seem to violate the 'natural order of persons' in addition to being harmful. The frequent elicitation of contempt in response to cruel and offensive remarks is also apparent in both facial expressions and self-report ratings for the 'naturalistic' version of the ‘remarks’ scenario. Additionally, stealing - potentially another 'natural order of persons' transgression, also frequently elicited contempt on the self-report measure. However, responses to 'physically striking' or 'intentionally killing' another person, both 'naturalistic' scenarios and canonical instances of 'harm', only selectively elicited anger in self-report measures whereas emotional expressions over these violations were more mixed. These results further suggest MFT harm measures may be more reliable than 'naturalistic' ones, and illustrate how Constructive Sentimentalism may be better suited to explaining the results.

In comparison, four out of five MFT impurity scenarios elicited expressions consistent with some form of purity violation, and the one which mostly elicited anger might be explained by informational assumptions behind receiving a 'blood transfusion from a child molester'; although in the self-report follow up this primarily elicited a 'grin-and-bear-it' response, suggesting it may contain a perceived 'natural order' violation, as with 'ant killing'. However, this 'grin-and-bear-it' response might be argued as reflective of the relative lack of severity and high weirdness of these two scenarios reported in Gray and Keeney (2015a). There was also a more mixed pattern for two of the

scenarios, in that both smiles and anger were frequently expressed in response to 'soul selling' and 'tail addition surgery' - although self-report follow up suggested participants found these scenarios elicited what could be described as 'surprised enjoyment tinged with disgust', with anger not reported as elicited by either scenario. In contrast, from the 'naturalistic' impurity scenarios, three out of five mostly elicited smiling expressions. Of the two that differed, the one depicting rape reliably elicited anger - supporting critiques that this is typically perceived as a harmful violation of autonomy, even if it can also/further be classed as a violation of purity. Similarly, responses to the adultery scenario support claims this contains 'mixed' content - eliciting expressions of anger, contempt, and smiling, with contempt as the most frequent self-report selection. This mix of reactions to adultery is what might be expected on Prinz's approach (cf. with discussion of jealousy as a blended emotion, and adultery as violating more than one domain; Prinz, 2004b, p. 98-100; Prinz, 2009, p.280). Furthermore, as angry expressions were so prevalent - to the point of exclusive selection - in response to the scenario depicting rape, it seems reclassifying this action as it is perceived (i.e., as 'harm') would be sufficient for the expected differences in anger between harm and impurity scenarios to become apparent in the 'naturalistic' set.

Taking these issues into account, Franchin et al.'s (2019) results show that anger was the most frequent response to 'harm' scenarios, and was elicited more by these scenarios than by impurity scenarios, providing support for both 'strong' and 'weak' MFT. The results for disgust were more mixed, with this expression elicited more frequently by MFT impurity scenarios than MFT harm scenarios, supporting 'weak' MFT. However, as anger was expressed more frequently than disgust for MFT impurity scenarios, and 'naturalistic' impurity scenarios elicited the fewest disgust expressions overall, 'strong' MFT was unsupported for impurity. As Franchin et al. are focused on disgust-impurity links, they conclude the results do not support 'weak' MFT, and this conclusion holds even when considering the 'mixed content' present in two of the 'naturalistic' scenarios may have skewed the results. Yet Franchin et al. do show a

reliable (domain-specific) association between smiles and purity violations which is not readily explainable by domain-general accounts such as Dyadic Morality (Schein & Gray, 2018). They state the results allow for the possibility of distinct moral foundations, with specific 'emotional footprints', even if the hypothesized disgust-purity link may need revising. Furthermore, the possibility that smiles may be indicative of covert disgust cannot be excluded - a possibility which Franchin et al. (2019) acknowledge would resultantly provide support for both 'strong' and 'weak' MFT within their results.

Landmann and Hess (2018) used a selection of moral scenarios, validated by Clifford et al. (2015), to examine emotion elicitation across five moral foundations. They investigated seven emotions, five of which are considered ‘characteristic’ of a specific moral foundation (Graham et al., 2013), and found mixed results. Violations of harm and purity showed specific emotional responses, whereas violations of fairness, authority, and loyalty showed no such specificity. Disgust was the only emotion that clearly followed the expected MFT-based trend, being elicited most strongly across purity violations. Purity violations also elicited the least anger and rage ratings, and these emotions were most strongly elicited by violations of fairness and care (harm), as expected – although violations of authority and loyalty also elicited reasonably strong anger and rage. However, the results also show high ratings for contempt and resentment for violations of care and fairness, with the lowest ratings for these emotions given to authority violations. Also, fear had the lowest ratings of all emotions, and was lowest for authority ratings. These emotion ratings for authority violations have an element of specificity, in that MFT hypothesizes respect and fear as characteristic emotions arising in the context of authority/subversion, but the results run in the opposite direction to what might be expected (i.e., high contempt/resentment - as relative opposites to respect, or high fear, in response to authority violations).

Unusually, Landmann and Hess also examined compassion – which is considered characteristic of the ‘harm/care’ foundation (Graham et al., 2013), even

though none of the scenarios used depict the ‘care’ side. Compassion was elicited more strongly for violations of harm and purity in comparison to the other foundations. Given participants were asked about ‘the situation’, rather than ‘the act’, in the scenarios, it is plausible that compassion elicited by harm scenarios is victim focused. This may be taken to support Gray and Wegner’s (2011) suggestion that sympathy is typically elicited by victims - although this would then seem opposed to Dyadic Morality, as compassion might be expected to be similarly elicited across all foundations if foundations are transformations of harm (following Schein & Gray, 2018). However, as the scenarios Landmann and Hess (2018) use to depict impurity have no obvious victims, the elicitation of compassion by impurity scenarios is more challenging to explain – even if some of this may be accounted for by individual differences in moral values as measured by the Moral Foundations Questionnaire (Graham et al., 2011), as Landmann and Hess show. It is possible that people who commit impure actions elicit compassion as a result of their own self-victimization. It is also, or further possible that compassion is elicited from inferences drawn about the agent’s moral character, in that such behaviour has arisen as a result of some perceived defect in moral character - and the acquisition of this defect may have been somewhat outside of the agent’s control. Possible explanations aside, Landmann and Hess's (2018) results provide 'weak' support for MFT's hypothesized links between both disgust and impurity, and anger and autonomy (i.e., care/harm & fairness/cheating). The results also suggest possible support for a distinction between fundamental and derived moral domains (following Prinz, 2009), as compassion was elicited more strongly for violations involving fundamental domains (i.e., harm or purity) than derivative domains.

Heerdink et al. (2019) show that ‘expressions of anger and disgust drive inferences about autonomy and purity violations’. In all three of their studies, participants were presented with a content-ambiguous scenario which depicted a social interaction from an observational standpoint – not sharing unhealthy snacks, drinking alcohol in university, and a conversation relating to an unspecified behaviour. Across these

studies, Heerdink et al. report that the expression of emotion from an observer of the transgression biases the inferences made by participants regarding why the transgression was wrong. Those that read about the observer expressing anger tended towards providing autonomy-based reasons (e.g., the behaviour is selfish), whereas those reading about a disgusted observer tended towards purity-based reasoning (e.g., the behaviour is distasteful). These results favour theories which argue for specific links between emotions and moral content. Furthermore, although Heerdink et al. find correlations between anger and disgust, and autonomy and purity, which might be tokened in support of the position against exclusive specificity taken by Cameron et al. (2015), Heerdink et al. (2019) suggest that their findings are not well explained by either 'shared core affect' or 'shared conceptual activation' – both of which Cameron et al. propose as explaining these correlational associations.

Of particular concern for those arguing against specific (and exclusive) moral-emotion links is research from Tracy, Steckler, and Heltzel (2019), investigating 'The Physiological Basis of Psychological Disgust and Moral Judgments'. Tracy et al. (2019) conducted a series of double-blind studies in which the treatment group ingested ginger – a nausea inhibiting antiemetic. Tracy et al. show that, in comparison to the control (sugar pill) group, those that ate the ginger pill reported less feelings of disgust when viewing images likely to elicit moderate 'core disgust', such as snot in a napkin - a 'purity-offending' stimulus with no moral content. They further show this effect remains present for moderately severe purity violations, such that those in the ginger condition rated moderately 'impure' moral scenarios as less severe - either via a main effect, or mediated through an interaction with Private Body Consciousness. This follows from previous research (e.g., Schnall et al., 2008; Johnson et al., 2016) showing greater awareness of bodily sensations is associated with increased severity ratings in response to moral scenarios. However, whilst this effect of bodily awareness was present across both 'moderate' and 'severe' violations, regardless of content type - the effect of ingesting an antiemetic was restricted to only moderately 'impure' moral scenarios, with

no such reduction in severity ratings apparent across scenarios depicting moderate or severe violations of four other ‘moral foundations’. These findings are strongly supportive towards accounts of (exclusive) emotional specificity with regard to moral content.

Importantly, Tracy et al.'s (2019) study design is sufficient to address or bypass many of the potential issues mentioned within the review pieces (Cameron et al.,2015; Landy & Goodwin, 2015a). All the concerns raised by Cameron et al. (2015) are sufficiently addressed. Tracy et al. (2019) do not rely on the use of forced-choice options or ANCOVA analyses, and they assess moral content of varying severity across five foundations. Furthermore, as the experimental disgust manipulation operates by suppression, rather than induction, it bypasses concerns regarding conceptual knowledge activation or stimulation of core affect. This also addresses Landy and Goodwin’s (2015a) concern of confounds arising should participants take offense to the induction methodology, as well as concerns regarding potential misattribution of emotions (cf. Schnall et al., 2015), and makes the manipulation incidental (i.e., target irrelevant, and covert) whilst ensuring the elicitation of any disgust is integral (i.e., target relevant, as per Wisneski & Skitka, 2017).

Moreover, Tracy et al.’s (2019) scenarios rely largely on pre-validated materials (i.e., Clifford et al., 2015), and although critics may quibble about the impurity scenarios being perceived as containing core disgust elicitors, any presence of these seems broadly comparable between the moderate and severe impurity scenarios. As such, any presence of core disgust elicitors seems unlikely to explain differences between ratings on scenarios of different severity. That the effect is limited to moderately severe impurity violations seems more likely due to limits of the antiemetic effect, given this effect was similarly present over ratings of moderately disgusting images, but not present over ratings of highly disgusting images.

Tracy et al. (2019) suggests their results show that the disgust elicited by (at least moderate) violations of purity is 'real', as inhibiting a component of the disgust response (i.e., nausea) resulted in reduced severity ratings for this type of moral judgement. In contrast, that this effect was not present over ratings of scenarios involving other types of moral content, regardless of severity, also suggests that any self-reported disgust in response to non-purity moral violations is likely metaphorical. Tracy et al. (2019) argues their results provide evidence for a causal connection between feelings of physiological nausea and psychological disgust, and further suggest that this physiological experience may constitute part of the participants' purity-based moral judgements. They state that "[a]pparently, when we witness a purity-based moral infraction of some ambiguity (i.e., a moderately severe violation), we feel nauseous, and this feeling tells us that what we are seeing is wrong."(p. 27). This is extremely similar to how Prinz (2009) suggests moral judgements operate.

6.4. Recent Evidence in the context of Reviews

Research conducted after the reviews were published has primarily focused on correspondences between emotions and moral content, with apparently little research investigating amplification effects following Landy and Goodwin's (2015a) review. As such, there is limited material from which to challenge Landy and Goodwin's conclusions. Johnson et al. (2016) found no evidence of an amplification effect for incidental disgust, nor did Royzman et al. (2017). Wisneski and Skitka (2017) found a medium sized amplification effect for target-relevant disgust on moral judgement, but report this effect is not present for incidental disgust. Only the study by Tracy et al. (2019) reports an effect where the disgust manipulation might be considered as incidental, although strictly speaking their methodology works to suppress target-relevant disgust, rather than by inducing incidental disgust. Even then, the size of the

effect reported (*d* = .12) is in keeping with the upper bound of the amplification effect (*d* = .11) suggested by Landy and Goodwin (2015a) - although it is not clear that these effect sizes are readily comparable given the former is technically due to suppression.

However, the research does suggest Private Body Consciousness plays some part in moral judgements, showing participants who responded highly on this scale tended towards making harsher moral judgements (Johnson et al., 2016; Tracy et al., 2019); and there is some suggestion that individuals who are disposed towards high emotion sensitivity also make harsher moral judgements (e.g., Landy & Piazza, 2019). These findings are supportive towards claims for affect playing a causal role in moral judgement (e.g., Prinz, 2007a, 2009). These findings also suggest that the size of incidental disgust amplification effects may exceed the bounds hypothesized by Landy and Goodwin (2015a) once relevant moderators are considered. Yet although the (limited) post-review research cited suggests that Landy and Goodwin's estimate of the effect bounds may be accurate, it also suggests the effect found may be small to non-existent precisely because they are examining the effect of *incidental* disgust inductions. Increasing or suppressing target-relevant disgust may produce larger, and/or more robust, effects on moral judgements (e.g., Wisneski & Skitka, 2017). However, it remains the case that content-specificity confounds within the studies Landy and Goodwin (2015a) analyse may be suppressing an effect of disgust on judgements of impurity. The majority of post-review research cited, which focuses on correspondences between emotions and moral content, is relevant to addressing this concern.

Recent research relating to Cameron et al.'s (2015) review provides mixed evidence in relation to correspondences. In support of Cameron et al.'s (2015) position, a key concern is that different types of disgust appear to be associated with different types of (moral) content. There may be specific and exclusive links between 'impurity' and pathogenic or sexual disgust (measured by 'grossed out'), but there may be no such links between 'impurity' and *moral* disgust (measured by 'disgust') – which appears to be

more closely related to anger than core disgust. This conclusion finds support via research from Kollareth and Russell (2017; 2019), van Leeuwen et al. (2017), and Oaten et al. (2018). Additionally, there is also some concern with regard to how disgust is operationalised, as certain facets of this emotion (e.g., nausea, cf. Tracy et al., 2019) may be more closely linked with moral judgements than others (e.g., oral inhibition - including self-reported nausea, cf. Royzman et al., 2017).

Many of these concerns have been discussed in greater detail by Piazza et al. (2018) as part of their review regarding the role of disgust in moral cognition. They also echo many of the methodological concerns detailed by Cameron et al. (2015), particularly those concerning the co-occurrence of anger and disgust. However, Piazza et al. (2018) consider ANCOVA-based analysis methods more favourably, stating these methods "have consistently found clear and reliable mean differences in the relationship anger and disgust measures bear with moral content" (p. 28). Whilst Piazza et al. agree with the claim that disgust feelings seem to correspond with the presence of core elicitors, rather than the presence of moral content *per se* (cf. Oaten et al., 2018), they also note a number of ways in which disgust differs from anger in moral contexts. For example, the presence of mitigating circumstances tends to reduce anger, but not disgust; disgust may plausibly co-occur with moral approval, whereas anger and moral approval co-occurring is highly implausible; anger and disgust seem to track different aspects of sexual transgressions; and disgust seems largely unconcerned with the intentionality of actions, in contrast to anger. Importantly, Piazza et al.'s (2018) review identifies avenues for theoretical clarifications with regard to both how and why disgust may be associated with both impurity, and moral cognition moral generally.

For example, Piazza et al. (2018) proposes a conceptual dimension of 'disgustingness', defined as the recognition that certain events potentially elicit disgust for some people, which they suggest may be more important for the construal of 'impurity' than actually feeling disgusted. This concept is derivable from harm, but

dissociable from harm in practice. In support, they cite Wasserman et al. (2017), who report disgust as a key feature in discerning both violations of purity from those of harm, and pathogen-including violations from all other violations. Thus, whilst Piazza et al. conclude there seems to be little effect of core disgust experiences on moral judgements about others' actions, they do not rule out a role for disgust in morality, nor its apparent correspondences with impurity. In sum, the evidence favourable towards Cameron et al.'s (2015) review points towards a need for theoretical clarifications and revisions, rather than a lack of specific (and/or exclusive) correspondences per se. Future research examining the relationship between moral disgust and anger, as well as research exploring moral judgement in response to pathogen-free violations of sanctity (e.g. violations of religious codes or practices), would be particularly useful in assisting such clarifying revisions.

However, the majority of post-review research with findings opposed to Cameron et al.'s (2015) position benefit from addressing questions of correspondences more directly. Franchin et al.'s (2019) results suggest violations of purity elicit a specific (and different) emotional response to harmful violations. Landmann and Hess (2018) also report specific emotional responses in relation to harm and purity violations. Heerdink et al. (2019) show that an observer's expression of either anger, or disgust, in a context-ambiguous scenario biases the inferences made by participants with regard to whether a behaviour was harmful or impure. All these studies might be taken in support of specific correspondences between emotions and moral content. Furthermore, the results from Tracy et al.'s (2019) research provide evidence in favour of specific and exclusive correspondences between disgust and impurity, whilst suggesting that moral judgement may be constituted, at least in part, by elements of core disgust. Tracy et al.'s (2019) findings also benefit from bypassing the majority of methodological issues raised by critics, thus providing a convincing link between moral judgements and phenomenal aspects of disgust that Piazza et al. (2018, p. 7) would likely accept as evidence for a role of felt disgust in moral judgements.

On balance, the research which bears on Cameron et al.'s (2015) review seems to weaken their conclusions, rather than strengthen them. The evidence favourable towards their position suggests 'impurity' needs to be defined with greater precision, and that the (potentially) varying roles for disgust needs further investigation which accounts for different types of disgust (e.g., pathogen, sexual, moral). However, it does not rule out 'impurity' as a specific moral dimension which is rooted in some form of disgust, even when taking a favourable view of the Theory of Dyadic Morality (Schein & Gray, 2018) - as Piazza et al. (2018) seem to do. Furthermore, accepting there may be some merit with regard to ANCOVA-based analyses, following Piazza et al. (2018), puts several of the papers Cameron et al. (2015) review back into the evidence pile in favour of specific correspondences. Impurity and disgust may not be performing quite as described by Moral Foundations Theory (Graham et al., 2013), but it would seem 'characteristic associations' remain. If this is the case, then content-specificity confounds may indeed be suppressing the amplification effect size reported by Landy and Goodwin (2015).

The evidence against Cameron et al.'s (2015) position is suggestive of specific correspondences between emotions and moral content, although these are not always as predicted by Moral Foundations Theory (i.e., smiling at impurity - Franchin et al., 2019). However, this research does include studies showing evidence of an exclusive link between impurity and disgust whilst controlling for many potential confounds. Tracy et al.'s (2019) results are strongly supportive of Constructive Sentimentalism (Prinz, 2009), a theory which suggests moral judgement is constituted by emotions, and that moral disgust is constructed via recalibrations of non-moral disgust - an approach which may address many of the points highlighted by Piazza et al. (2018). Given the significance, and specificity, of Tracy et al.'s (2019) results, further investigations using this type of methodology are likely to be highly informative, and would benefit from a range of replication attempts. In particular, combining the methodologies of Tracy et al.

with those of Oaten et al. (2018), or Kollareth and Russell (2019), may be especially informative with regard to associations between impurity and disgust; for example, by examining the effects of antiemetic ingestion across dimensions of moral/non-moral, pathogen presence/absence, and stimulus strength.

Further research which investigates correspondences between emotions and moral content, whilst accounting for potential confounds, may also help clarify the validity of existing claims. In particular, claims regarding specificity/exclusivity may be examined using the experimental framework proposed by Cameron et al. (2015); and these claims can be tested across each of the three hypothesized relationships mentioned in Landy and Goodwin (2015a). For example, the claim that certain types of moral content *elicit* specific emotions may benefit from stronger support if favourable evidence were obtained when using methods which allow for an open-ended response. Similarly, claims that certain emotions may contribute to the *moralization* of content, and/or the *amplification* of moral judgements, may also receive better support from studies which employ a range of *both* emotion types *and* moral content; for example, studies which use all five types of 'basic' emotion between induction conditions, and using moral judgement scenarios which cover all theorized types of 'foundational' content as repeated measures. The results of these studies may assist in adjudicating between theories which advocate against specific correspondences (e.g., Dyadic Morality, Schein & Gray, 2018) and those which argue in favour of such 'characteristic' or 'constitutional' associations (Moral Foundations Theory, Graham et al., 2013; Constructive Sentimentalism, Prinz, 2009).

.

Chapter 7 - Testing Constructive Sentimentalism

Constructive Sentimentalism (CS; Prinz, 2009) argues that moral judgements are *constituted* by emotions, hypothesizing that judgements of immorality within different moral domains correspond to different emotions. On this approach, the emotion typically elicited in response to a moral violation is dependent on *both* the domain the action relates to, and the observers' relationship to its perpetrator. As such, CS shares some ground with Moral Foundations Theory (MFT; Graham et al., 2013) in arguing that (im)morality is about more than just 'harm', and in associating specific emotions with particular moral domains. However, CS also hypothesizes that what matters most for judgements of moral (i.e., 'good') actions is the relationship between agent and patient, whereas the moral domain the action relates to (if any) does not appear to factor in determining the emotional response. This focus on the agent-patient relationship means CS also shares some ground with the Theory of Dyadic Morality (TDM; Schein & Gray, 2018), a theory which claims (im)morality actually does revolve around 'harm', contra MFT. That CS appears to fit better with one theory when considering 'immoral' judgements, and possibly better with another when considering 'moral' judgements, is a juxtaposition of particular interest.

Constructive Sentimentalism suggests that the change in focus between immoral and moral actions is due to a relative asymmetry in morality generally. Immoral behaviour provides a much greater (evolutionary) potential threat than moral behaviour, and the overarching domain of immorality may be more important (and more developed) as a result. Furthermore, there is a much greater range of behaviour that may be considered 'good', but rarely considered moral (e.g., not stealing). Indeed, behaving morally (or at least not immorally) is considered the norm. In contrast, TDM would argue that there was no change in focus to begin with, as all (non-harm) moral foundations (or domains) are simply 'transformations or intermediaries of dyadic harm' (Schein & Gray,

2018). This may appear to be a more parsimonious position, particularly as it is not obvious why moral domains/foundations are hypothesized to be at least 'characteristically associated' with certain emotions for immoral actions, but then these emotion-domain associations seem to fade into relative irrelevance when considering moral actions.

The existence of associations between discrete emotions and specific moral domains is of particular importance to CS given its emphasis on emotion and the role emotions are argued to play in morality. However, the two most frequently hypothesized correspondences, those between anger and 'autonomy' (harm & fairness) violations, and disgust and 'divinity' (impurity/sanctity) violations, have reportedly been found wanting for evidence. Cameron, Lindquist and Gray (2015) reviewed a range of studies which investigated these associations and suggest there is little evidence to support these hypotheses – especially once methodological confounds, such as forced-choice responses, are considered. They find anger and disgust frequently tend to co-occur across both harmful and impure violations, which they claim provides evidence against specific and/or exclusive emotion-domain associations. Cameron et al. (2015) propose any 'loose' associations may be better accounted for by factors relating to 'core affect' (i.e., valence/arousal) and conceptual knowledge (e.g., contamination, disgust, and impurity are all conceptually related). However, a more recent set of studies (Tracy et al., 2019) has provided strong evidence in favour of CS's approach, particularly with regard to both emotion as a constituent of moral judgement and the existence of a specific and exclusive link between disgust and impurity violations. Importantly, Tracy et al.'s (2019) study design bypasses concerns about potential confounds relating to core affect and conceptual knowledge, which suggests Cameron et al.'s (2015) conclusions are open to challenge from studies which better control for these potential confounds (see Chapter 6 for review).

Hypotheses have also been advanced regarding associations between ‘positive’ emotions and ‘morally good’ actions on all approaches. Research following the Moral Foundations Theory approach (Graham et al., 2013) identifies ‘other-praising’ emotions of admiration, gratitude, and elevation as emotions which tend to arise in response to witnessing moral excellence (Algoe & Haidt, 2009). However, of these, only gratitude is identified by MFT as being characteristic of a particular domain (fairness/cheating) - admiration and elevation are not included. MFT suggests compassion (for victims) is characteristic for the harm/care foundation, and emotions of ‘group pride’ (loyalty/betrayal) and ‘respect’ (authority/subversion) are also identified in this manner, although notably - no positive emotion association is provided for the degradation/sanctity foundation. In contrast, CS hypothesizes admiration is elicited by witnessing others doing moral deeds, and gratitude (or gratification) when the beneficiary (or benefactor) is oneself. However, it does not propose a specific association for elevation, instead suggesting 'dignity' (read: self-worth) as the emotion typically elicited by doing good deeds for oneself. In further contrast, TDM proposes inspiration and elevation as typically arising in relation to moral heroes, and emotions of relief and happiness as relating to beneficiaries, but makes little mention of admiration or gratitude in these cases. Leaving disagreement over which emotions are involved aside; MFT proposes there are ‘characteristic’ emotion-domain associations for both moral and immoral actions, CS suggests domain associations are only present for immoral actions, and TDM argues there are no such domains - so any apparent associations must be explainable by other means.

7.1. The current study (Study 2)

Providing a comprehensive test of emotion elicitation as detailed by CS allows for further investigation of all the above claims regarding associations, as well as those CS claims depend on the observers’ relationship to the agent committing the violation.

Following the CAD model (Rozin et al., 1999) - which suggests 'community' (loyalty & authority) violations elicit contempt, 'autonomy' violations elicit anger, and 'divinity' violations elicit disgust - CS also proposes these emotions differ when oneself is the perpetrator of the violation. In such instances, 'autonomy' transgressions elicit guilt, 'divinity' violations elicit shame, and 'community' transgressions elicit a blend of guilt and shame (or embarrassment - following Rozin et al., 1999). However, CS further proposes these associations differ again when the violation is committed by a loved one, such that a blend of other-focused and self-focused emotions are elicited. In these instances, 'autonomy' violations may elicit a feeling of hurt (anger/guilt), we may be ashamed of those close to us who commit 'divinity' violations, and if a loved one were to transgress against 'community' then a blend of hurt/ashamed (perhaps disgrace) may be elicited. According to CS, what matters for emotion elicitation in response to immoral actions is defined by the (perceived) moral domain to which the action belongs *and* who is performing the action; whereas for moral actions, emotion elicitation is determined by the (dyadic) relationship between agent and patient, rather than by any moral domain to which it may relate. CS predicts that for immoral and moral actions, emotion responses are expected to be selected respectively in keeping with the following arrangements (Tables 7.1.1 & 7.1.2 are as detailed in Prinz, 2009, p.79-81).

Table 7.1.1 Immoral Actions

Domain/Actor	Stranger	Self	Loved One
Autonomy	Anger	Guilt	Hurt
Community	Contempt	Guilt/Shame	Hurt/Ashamed
Divinity	Disgust	Shame	Ashamed

N.B. Prinz argues contempt is a blend of anger and disgust.

Table 7.1.2 Moral Actions

Agent	Patient	Emotion
Other	Other	Admiration
Other	Self	Gratitude
Self	Other	Gratification
Self	Self	Dignity

N.B. The Other-Self and Self-Other actions are not tested directly in this study.

The primary aim of the study is to provide a comprehensive test of Constructive Sentimentalism's hypotheses regarding emotion elicitation patterns across a range of morally valenced events, and examine whether these patterns vary depending on the observer-perpetrator relationship (as shown in Tables 7.1.1 & 7.1.2). A secondary aim is to investigate whether the observers' type or level of attachment to the perpetrating agent affects ratings across a range of factors thought morally relevant, such as those relating to severity, accountability, and emotional intensity. In addressing these aims, the current study follows recommendations from Cameron, Lindquist and Gray (2015). The scenarios used depict a range of morally valenced events (immoral, neutral, moral), and cover a variety of moral transgression types. More importantly, the study avoids forced-choice emotion selection issues, offering up to 60 emotion response selections from which participants may choose up to 6 different responses. The study also takes up the challenge of offering an open response selection, just in case any emotion(s) elicited in participants could not be adequately conveyed using the other 59 options available. Cameron et al. (2015) report being unable to find any studies which offer open-ended responses when testing moral-emotion correspondences, so this might be the first study that provides a test of this strength.

7.2. Method

7.2.1. Design

The study used a mixed design to examine the effects of agency-attachment (broadly construed) on moral judgements. There are six between subject factors covering a variety of actor-attachments (Stranger, Token Group Member, Self, Idol, Friend, Family), three of which directly relate to the actors mentioned in Table 7.1. (stranger, self, family), and three others aimed at exploring types of actors which may be considered conceptually close to these categories (group member, idol, friend). There are ten within subject factors, split into three categories covering morally negative, positive, and neutral types of moral content. Morally negative events align with the three dimensions of the CAD model (Rozin, Lowery, Imada, & Haidt, 1999), covering violations of 'Community', 'Autonomy', and 'Divinity' -- approached from Constructive Sentimentalism (Prinz, 2009). These combine with the between subjects factor to test the emotion elicitation patterns hypothesized in Table 7.1.1. Positive events are also aligned along these dimensions, with the addition of a fourth 'Self-directed' factor to test self-perpetrated actions which benefit the self. These permit direct testing of the 'other-other' and 'self-self' combinations shown in Table 7.1.2. 'Counter-normative', 'Normative', and 'Neutral' actions are included to cover three varieties of non-moral stimuli. The dependent measures used include ratings of rightness/wrongness, goodness/badness, praiseworthiness/blameworthiness, reward/punishment, emotional intensity, and emotion elicitation (multiple-option categorical selections) given in response to scenarios depicting actions pre-categorised as relating to each of the ten within subject factors.

7.2.2. Participants

For the non-categorical measures, G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) indicated a sample size of 192 would be required to have 99% power for detecting a medium size interaction effect (0.2) at an alpha level of .01, and a sample size of 204-220 would have 99% power to detect a medium size main effect (0.2) within factors at an alpha level of .01.

The opportunity sample (*N* = 253) was recruited via adverts placed on the University research participation management system (SONA). Participants self-identified as speaking fluent English, being 18 or over, and being willing to read potentially offensive or upsetting content. Students were offered 0.5 Research Credits for completing the study. Of the 222 cases retained for analysis, respondents were 20 years old on average (*SD* = 7.56), and the majority (190) were female. The research for this project was submitted for ethics consideration under the reference PSYC 18/ 315 in the Department of Psychology and was approved under the procedures of the University of Roehampton's Ethics Committee on 24.10.18.

7.2.3. Materials

The majority of scenarios were created for the purposes of this study, such that there were two scenarios for each of the ten judgement dimensions. Each scenario briefly describes an event, often with minimal context, and highlights the action to be considered when rating. In the five most negative categories (i.e., non-moral or worse), one of each scenario pair was written to involve some form of 'mitigation' to reduce inferences about moral character which may influence responses (see Seidel & Prinz, 2013a). For balance, the three positive 'CAD' pairs consist of one 'moderate' scenario,

and a comparatively 'extreme' scenario - while for the self-directed scenarios, one ‘stops bad’ (quitting a bad habit) whilst the other ‘does good’ (mastering a skill).

However, some scenarios were drawn from the extant literature to better underpin materials depicting ‘autonomy’ and ‘divinity’ violations. The ‘WALLET’ and ‘plane crash’ ('CANNIBAL') scenarios used by Schnall et al. (2008) are included under ‘autonomy’ and ‘divinity’ respectively - following their use by Seidel and Prinz (2013a). Their use suggests both scenarios are considered as valid representations of their respective domains under Constructive Sentimentalism. Also, following their approach, the ‘plane crash’ scenario has been edited so that the instance of cannibalism is the main focus, with reference to child murder removed so as not to depict multiple moral violations. The 'CRUEL' and 'SEXBUY' scenarios are included as these are considered 'naturalistic' examples of each type of content (Gray & Keeney, 2015a), and are also in keeping with validated MFT scenarios. The former is validated as an example of 'harm' (Clifford et al., 2015), and the latter involves sex - which all discussed approaches consider as fitting under the 'divinity' domain (sanctity/degradation). A list of the actions depicted in the scenarios is detailed in Table 7.2; full scenario details are available on the pre-registration site for the study.

Questions were asked about each scenario, capturing the extent to which: (1) 'it was morally **right/wrong** for [the agent] to do this action?'; (2) 'it was **good/bad** for [the agent] to do this action?'; (3) should '[the agent] be **praised/blamed** for doing this action?'; and (4) should '[the agent] be **rewarded/punished** for doing this action?'. These questions are rated on 11-point bipolar scales, with every other point detailed such that the ends are labelled 'Extremely' (Q1 & Q2) or 'Greatly' (Q3 & Q4), all third points in are labelled 'Moderately', and the options either side of the mid-point are labelled 'Slightly' (Q1 & Q2) or 'Lightly' (Q3 & Q4).

Further questions were asked to gauge (5) 'What emotions are you experiencing as a result of [the agent] having done this action?' and (6) 'How intensely are you experiencing each of these emotions?'. Up to six emotions, from an alphabetically sorted list of sixty, could be selected in response to (5); there was also an 'Other' option, allowing for open-ended responses to be given via a follow up question, (5a) 'Please name or specify the nature of the other emotion(s) you are experiencing'. Only the emotions selected (or named) during question 5 (and 5a) were presented for question 6. This was rated on a 7-point scale, ranging from 'very little felt' to 'very much felt'; although question 6 was not presented if 'Nothing specific' was selected as the primary response to question 5.

7.2.4. Procedure

Participants were presented with a link to the study, which was administered on Qualtrics. Participants were asked to provide consent, and demographic information (age/sex). Instructions on scenario rating were displayed, followed by a list of the emotion response options available, and an example of the emotion response question with 'Nothing specific' selected as an instructional example. Participants were randomly allocated by Qualtrics to one of the six agency-attachment conditions. Those not in the 'Stranger' or 'Self' conditions were asked to provide a name for a person they thought matched a description provided; and all except those in the 'Self' condition were asked to specify the sex of the agent selected. All scenarios were presented in a random order, and started with 'Suppose [witness] found out that [the agent] ...', with agent specified by condition, and the witness being 'you' in all but the 'Self' condition where this was 'someone'. Questions 1 through 5 were all asked on a single page, with separate follow up pages for questions 5a and 6 as required.

7.2.5. Pre-registration

A priori power calculations, statements of hypotheses, planned analyses, and all study materials are available via the pre-registration site for this study - https://osf.io/eh7vz

Table 7.2. List of scenario actions by overarching *type* and domain

#	Scenario	Action
	IMMORAL	
	AUTONOMY (-)	
1	CRUEL	Talk to someone in a cruel and offensive manner.
11	WALLET	Keep the money from the wallet in order to have more money.
	COMMUNITY (-)	
2	LEAVEWORK	Regularly leave work early.
12	TESTIFY	Not testifying at the court [...] to minimize [...] risks to loved ones.
	DIVINITY (-)	
3	SEXBUY	Pay for sex
13	CANNIBAL	Eat the dead body of the child in order to stay alive.
	NON-MORAL	
	NORMS (-)	
4	LOUDMUSIC	Play music so loudly that other passengers can hear it.
14	BUMP	Not saying "excuse me" in order to save time [...]
	NEUTRAL	
10	BUSTRAIN	Take the bus in order to avoid waiting for a train.
20	BOOKREAD	Read an advanced physics textbook for fun.
	NORMS (+)	
5	POLITE	Have a polite conversation with the sales assistant
15	MAKEDRINK	Make drinks for fellow co-workers
	MORAL	
	AUTONOMY (+)	
6	KIDNEY	Donate a kidney
16	FEED	Buy food for a homeless person
	COMMUNITY (+)	
7	BLOODBANK	Donate blood.
17	CHARITY	Donate 25% of income to charity.
	DIVINITY (+)	
8	BEACH	Pick up litter from the beach.
18	PLANTING	Volunteer to help grow trees.
	SELF-DIRECTED	
9	SKILL	Spend time to master a life improving skill.
19	QUITHABIT	Quit an addictive and unhealthy habit.

7.3. Results

7.3.1. Response validity checks

A total of 253 responses were collected. Three cases were removed as the result of a participant completing the study twice, with the latter completion being removed each time. Three partial cases were also removed for failing to meet basic demographic criteria (e.g., not speaking fluent English). The remaining 247 responses were all above the improbable response speed threshold (i.e., eight minutes or greater).

Scenario-based validity checks identified 13 cases where non-moral scenarios were scored at or above the negative scale side mid-point (i.e., moderately or worse) for wrong, bad, blameworthy, or punishable. A further 69 cases where immoral scenarios were scored at or 'above' the positive scale side mid-point on the right-wrong dimension (i.e., at least moderately right) were identified, as were a further 9 cases where moral scenarios were scored at or above the negative scale side mid-point on the right-wrong dimension. The high number of responses identified by these checks warranted further investigation.

The 13 participants who scored either 'taking the bus instead of the train, or 'reading an advanced physics textbook for fun', as moderately negative in some regards were excluded; and (reversing the pre-registered analysis script order) a further 12 participants were excluded for rating any of the scenarios depicting 'morally positive' actions as at least moderately immoral. Inspection of the latter category showed half of these responses related to self-directed scenarios, and all positive scenarios (except autonomy-based ones) received at least one of these ratings. Although it may be plausible to contend that some of these ratings reflect genuinely held beliefs (e.g.,

religious prohibitions relating to blood/organ donation), it seems more likely that these responses are the result of careless or inattentive responding. As such, the exclusion of these 25 cases remains in keeping with the pre-registration plan.

However, inspection of the remaining 66 responses, from participants scoring 'immoral' actions as at least moderately positive, revealed two trends of interest. Firstly, events containing some form of mitigation (i.e., Wallet, Testify, Cannibal) received at least twice as many moderately positive responses as their unmitigated counterparts (i.e., Cruel, Leavework, Sexbuy). Secondly, the most positive response options appear to associate with attachment (especially over unmitigated scenarios), such that the majority of 'extremely right' (or just below) ratings across 'immoral' scenarios were given by participants rating the actions of themselves, idols, friends, or family -- rather than those of a stranger or group member. These trends, in combination with the relative frequency of such responses, suggest ratings on these scenarios are more likely to reflect genuinely held beliefs than careless or inattentive responding. As such, these cases were retained for analysis purposes – although this is a deviation from the pre-registered analysis plan.

7.3.2. Data Processing

The 222 responses considered to have passed validity checks were processed using the pre-registered scripts. These scripts unpacked the multiple response options for 'experienced emotion' (Q5) into multiple response sets; computed mean scores for scenario pairings (10) and scenario groupings (3) for each of the right/wrong, good/bad, praise/blame, reward/punishment, and emotion intensity ratings; and sorted emotion ratings into pre-defined 'family' types to investigate emotion response patterns.

7.3.3. Data Analysis

Data analysis was initially conducted by executing the pre-registered analysis script. A condition (6) by scenario type (10) ANOVA was conducted for each measure - right/wrong, good/bad, praise/blame, reward/punishment, and emotional intensity - with the aim of confirming scenario validity and investigating differences across these measures. Descriptive statistics across these measures over all conditions are reported for each scenario, domain pair (e.g., negative autonomy), and overarching moral valence type (e.g., immoral) in Table 7.3.

Measures of right/wrong (Figure 7.1) and good/bad (Figure 7.2) showed no significant differences between conditions. ANOVA showed a main effect ($p < .001$) of scenario type for each [Right/Wrong, $F(4.869,1046.867) = 966.357$; Good/Bad, $F(4.503,871.364) = 865.693$], confirming scenario validity (i.e., actions depicting immoral/neutral/moral events were discernibly rated as such). Subsequent pairwise comparisons (all mean differences $p < .001$, Bonferroni corrected) showed negative autonomy violations were rated as more wrong/bad (1.951/1.839) than negative divinity violations, which, in turn, were rated as more wrong/bad (.752/.691) than negative community violations. However, there was no significant difference in ratings between negative community violations and counter-normative violations. In contrast, morally positive scenarios were not rated as significantly different by type, although they were rated at least one scale point higher over each morally positive scenario type in comparison to morally neutral (i.e., non-moral) scenarios and scenarios depicting positive norms, with positive norms scoring higher than neutral scenarios (.745/.662).

Table 7.3. Descriptive Statistics. Mean ratings across all conditions.

#	Scenario	Right / Wrong	Good / Bad	Praise / Blame	Reward / Punish	Emotion Intensity
	IMMORAL	8.15 (1.28)	7.92 (1.49)	7.96 (1.14)	7.51 (1.07)	5.38 (1.09)
	AUTONOMY (-)	9.20 (1.56)	8.91 (1.77)	8.79 (1.40)	8.08 (1.40)	5.27 (1.31)
1	CRUEL	9.23 (1.74)	9.17 (1.71)	8.89 (1.69)	8.06 (1.69)	5.22 (1.44)
11	WALLET	9.17 (2.17)	8.66 (2.54)	8.69 (1.90)	8.10 (1.86)	5.33 (1.40)
	COMMUNITY (-)	7.26 (1.66)	7.09 (1.84)	7.36 (1.41)	7.10 (1.24)	5.10 (1.31)
2	LEAVEWORK	8.02 (1.95)	7.78 (2.20)	8.01 (2.02)	7.66 (1.85)	4.93 (1.46)
12	TESTIFY	6.48 (2.51)	6.37 (2.47)	6.68 (1.72)	6.52 (1.45)	5.23 (1.39)
	DIVINITY (-)	7.98 (2.12)	7.75 (2.18)	7.73 (1.85)	7.33 (1.66)	5.72 (1.17)
3	SEXBUY	8.42 (2.46)	8.41 (2.44)	8.25 (2.19)	7.65 (2.14)	5.54 (1.55)
13	CANNIBAL	7.54 (2.89)	7.09 (3.10)	7.21 (2.25)	7.01 (2.03)	5.86 (1.20)
	NON-MORAL	4.64 (1.05)	4.56 (.933)	5.15 (.922)	5.17 (.881)	4.78 (1.31)
	NORMS (-)	6.94 (1.47)	6.96 (1.32)	6.96 (1.13)	6.44 (.948)	4.61 (1.55)
4	LOUDMUSIC	7.62 (1.95)	7.49 (1.84)	7.48 (1.66)	6.73 (1.41)	4.68 (1.64)
14	BUMP	6.26 (1.69)	6.44 (1.46)	6.45 (1.16)	6.14 (.881)	4.52 (1.56)
	NEUTRAL	3.86 (1.74)	3.68 (1.57)	4.54 (1.46)	4.81 (1.43)	4.96 (1.49)
10	BUSTRAIN	3.84 (1.98)	3.78 (1.88)	4.78 (1.62)	5.02 (1.53)	4.80 (1.58)
20	BOOKREAD	3.89 (2.06)	3.59 (1.94)	4.29 (1.88)	4.59 (1.77)	5.09 (1.52)
	NORMS (+)	3.12 (1.46)	3.03 (1.37)	3.96 (1.46)	4.26 (1.46)	5.03 (1.38)
5	POLITE	2.94 (1.75)	2.98 (1.65)	4.34 (1.82)	4.61 (1.77)	5.14 (1.48)
15	MAKEDRINK	3.31 (1.76)	3.08 (1.57)	3.58 (1.61)	3.91 (1.63)	5.04 (1.43)
	MORAL	1.85 (.943)	1.82 (.974)	2.24 (1.15)	2.74 (1.40)	5.63 (1.03)
	AUTONOMY (+)	1.72 (1.04)	1.81 (1.24)	2.14 (1.33)	2.67 (1.58)	5.74 (1.12)
6	KIDNEY	1.82 (1.41)	1.96 (1.62)	1.86 (1.53)	2.17 (1.63)	5.89 (1.11)
16	FEED	1.61 (1.06)	1.66 (1.20)	2.42 (1.58)	3.17 (1.98)	5.64 (1.31)
	COMMUNITY (+)	1.87 (1.15)	1.84 (1.24)	2.15 (1.36)	2.75 (1.67)	5.70 (1.10)
7	BLOODBANK	1.74 (1.23)	1.81 (1.36)	2.27 (1.58)	2.90 (1.89)	5.66 (1.22)
17	CHARITY	2.00 (1.46)	1.87 (1.44)	2.02 (1.52)	2.61 (1.86)	5.76 (1.24)
	DIVINITY (+)	1.90 (1.19)	1.94 (1.23)	2.38 (1.38)	2.89 (1.63)	5.57 (1.21)
8	BEACH	1.95 (1.44)	1.98 (1.47)	2.51 (1.69)	2.99 (1.87)	5.61 (1.23)
18	PLANTING	1.85 (1.27)	1.91 (1.35)	2.25 (1.53)	2.78 (1.71)	5.57 (1.32)
	SELF-DIRECTED	1.93 (1.21)	1.69 (1.10)	2.28 (1.38)	2.66 (1.57)	5.58 (1.17)
9	SKILL	2.07 (1.45)	1.83 (1.30)	2.64 (1.72)	2.98 (1.84)	5.45 (1.37)
19	QUITHABIT	1.80 (1.36)	1.55 (1.25)	1.93 (1.51)	2.33 (1.71)	5.68 (1.28)

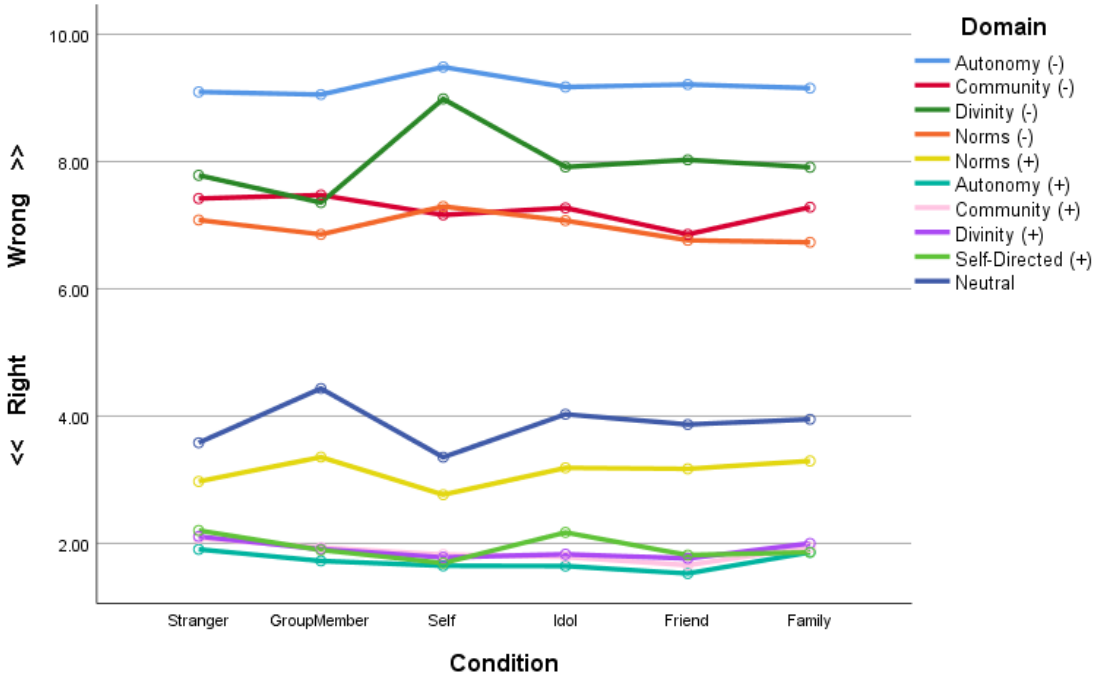


Figure 7.1. Ratings on the Right/Wrong dimension by Domain across Conditions

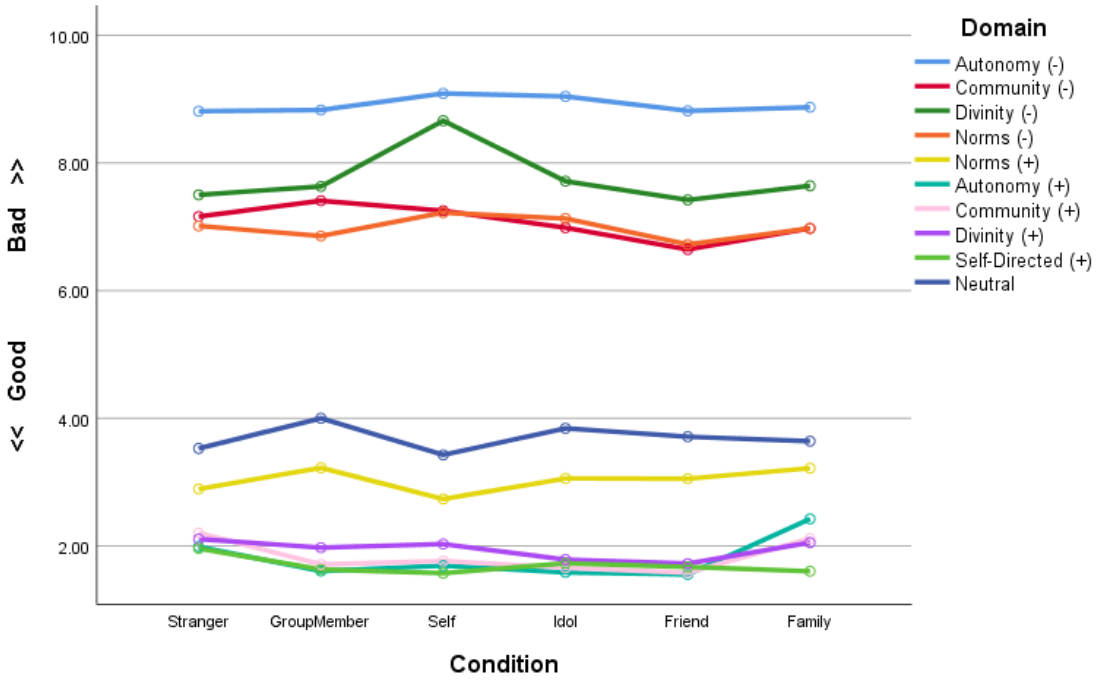


Figure 7.2. Ratings on the Good/Bad dimension by Domain across Conditions

Measures of praise/blame (Figure 7.3) and reward/punishment (Figure 7.4) also showed a similar pattern of differences between scenario types, with multivariate tests showing a main effect ($p < .001$) of scenario type for each [Praise/Blame, $F(4.377,941.027) = 888.523$; Reward/Punish, $F(3.475,747.132) = 557.025$]. Once again, morally positive scenarios were not rated as significantly different by type, although they were rated as being at least 1.5 scale points more praiseworthy and rewardable than morally neutral scenarios and scenarios depicting positive norms. Subsequent pairwise comparisons (all mean differences $p < .001$, Bonferroni corrected) showed negative autonomy violations were rated as more blameworthy/punishable (1.443/.986) than negative divinity violations, although these, in turn, were not rated significantly differently to negative community violations. However, negative community violations were rated as more blameworthy (.37, $p = .01$) and punishable (.637) than counter-normative transgressions, in contrast to ratings given over measures of wrongness and badness.

However, there were also differences across conditions on measures of praise/blame and reward/punishment. On the latter measure, Bonferroni adjusted pairwise comparisons revealed the only significant between subjects effect [$F(5,215) = 3.42$, $p = .005$] was between the 'self' and 'friend' conditions, with the self ($M = 5.178$) being considered more deserving ($p = .002$) of reward/punishment than friends ($M = 4.554$). In contrast, praise/blame scores frequently differed by condition [$F(5,215) = 4.18$, $p = .001$]. Bonferroni adjusted pairwise comparisons showed participants in the 'self' condition ($M = 5.244$) rated scenarios as deserving more praise/blame than strangers ($MD = .462$, $p = .05$), group members ($MD = .48$, $p = .032$), friends ($MD = .673$, $p < .001$), and family members ($MD = .47$, $p = .037$). Furthermore, these differences appear to be asymmetrically driven, whereby self-perpetrated actions tended to be considered slightly more blameworthy, and slightly less praiseworthy, than the same immoral or moral actions (respectively) perpetrated by other people - suggesting individuals may considers themselves as being more morally accountable for their own actions than others' are for theirs. Exploration of differences by condition across overarching category

types showed morally positive scenarios in the 'self' condition were rated closer to the scale mid-point (*M* = 3.12, *SD* = 1.77) and morally negative scenarios were rated further towards the end of the scale (*M* = 7.93, *SD* = .861) in comparison to all other conditions. However, a one-way ANOVA run to examine these trends showed they did not meet the threshold for significance.

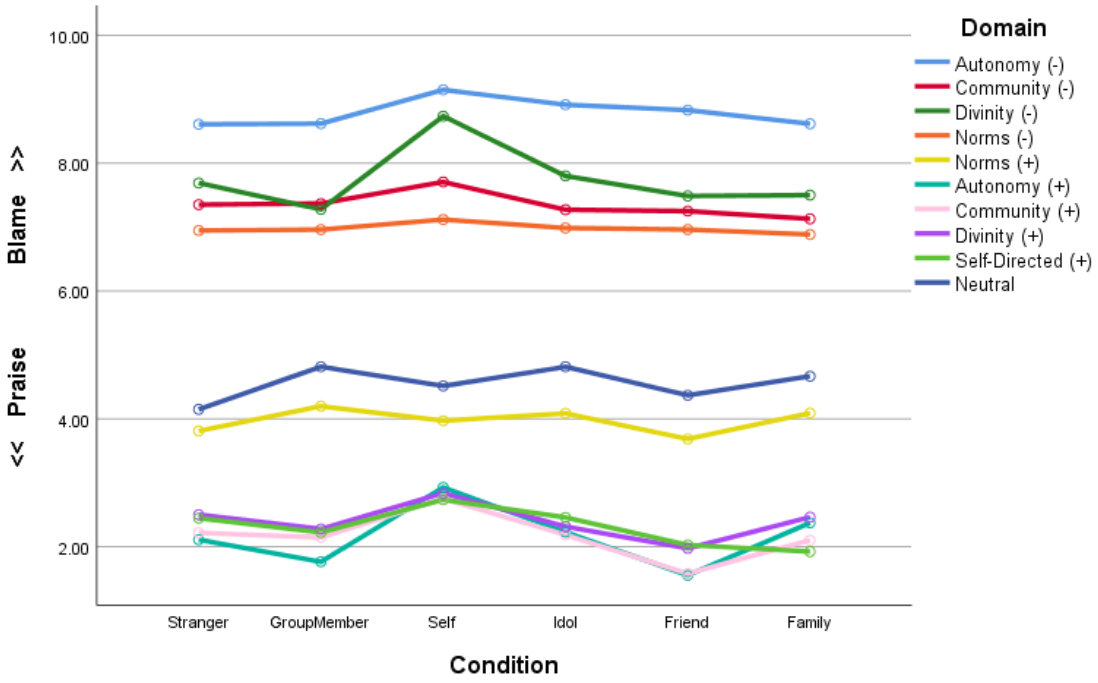


Figure 7.3. Ratings on the Praise/Blame dimension by Domain across Conditions

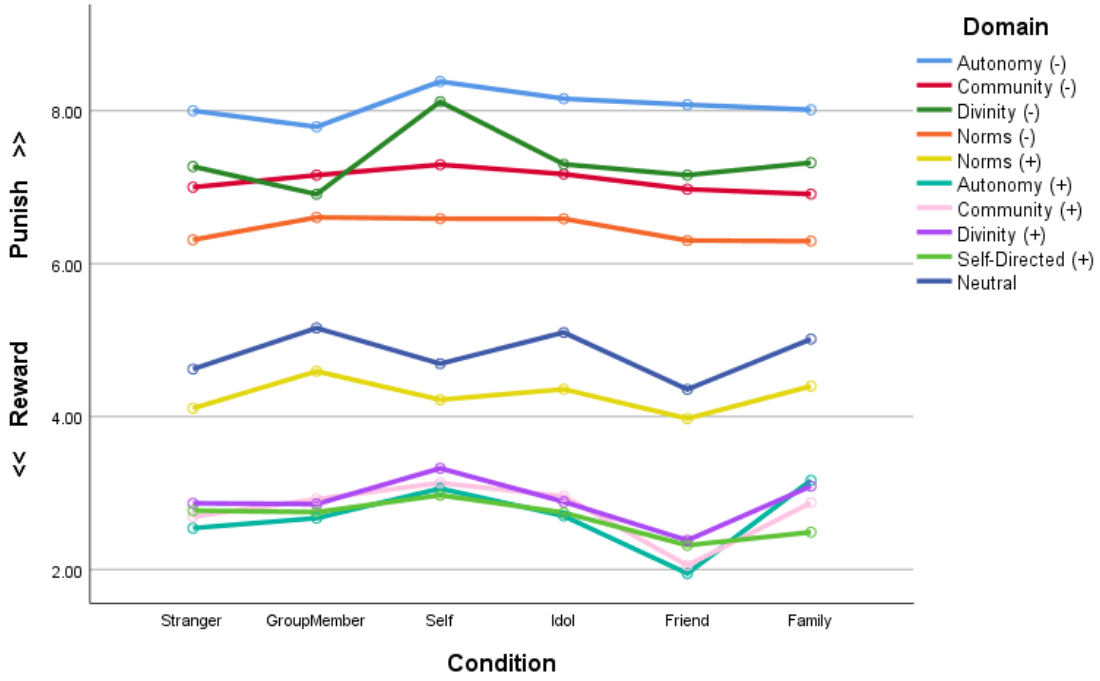


Figure 7.4. Ratings on the Reward/Punish dimension by Domain and Condition

Ratings for emotion intensity did not differ significantly between conditions, although ratings were highest in the 'self', 'family' and 'friends' conditions; and there were some differences in intensity by scenario type [$F(9,88 = 11.086, p < .001$]. Bonferroni adjusted pairwise comparisons showed scenarios within each overarching category tended to garner similar responses. The primary exception being that negative divinity violations elicited higher intensity ratings than negative autonomy or community violations, and higher intensity ratings than normative or non-moral scenarios, such that only morally positive scenarios scored similarly to negative divinity violations with regard to emotional intensity. However, the few other exceptions do not follow the same trends as the other dependent variables. Negative autonomy ratings were not significantly more intense than those for negative community, non-moral, positive norm, positive community or positive divinity scenarios. Counter-normative scenarios were less intense than all scenario types except negative community, and positive autonomy scenarios were slightly more intense than positive divinity scenarios. Simplifying these results by overarching category type, further exploration showed morally positive scenarios (*M* =

5.63, *SD* = 1.03) tended to elicit higher intensity ratings than morally negative scenarios (*M* = 5.38, *SD* = 1.09), which in turn tended to elicit higher intensity ratings than non-moral scenarios (*M* = 4.78, *SD* = 1.31). Simplifying further, the results show scenarios containing morally relevant content elicited more intense emotional responses than those not containing moral content.

There was only one consistent multivariate interaction between conditions and scenario type, which was for self-perpetrated divinity violations ('self' x 'divinity'). In contrast to the other five conditions, where autonomy violations were rated as consistently 'worse' than divinity violations, ratings for self-perpetrated divinity violations were comparatively closer to self-perpetrated autonomy violations. The extent of this difference in ratings shows that self-perpetrated divinity violations have a ratings profile more similar to those provided for autonomy violations (regardless of perpetrator) than to other-perpetrated divinity violations (i.e., violating ones' own purity is akin to violating someone else's autonomy). This interaction was found, in each case using Roy's Largest Root, across measures of wrongness [$F(9,211) = 2.501$, $p = .01$], badness [$F(9,211) = 2.09$, $p = .032$], blameworthiness [$F(9,211) = 4.376$, $p < .001$], and punishability [$F(9,211) = 3.205$, $p = .001$]. However, the pattern of interaction differed over emotional intensity [$F(9,92) = 2.357$, $p = .019$], with examination showing that emotional intensity ratings tended to cluster together more in the 'self' and 'family' conditions, as well as clustering together more over negative divinity scenarios and positive moral domain scenarios than other types.

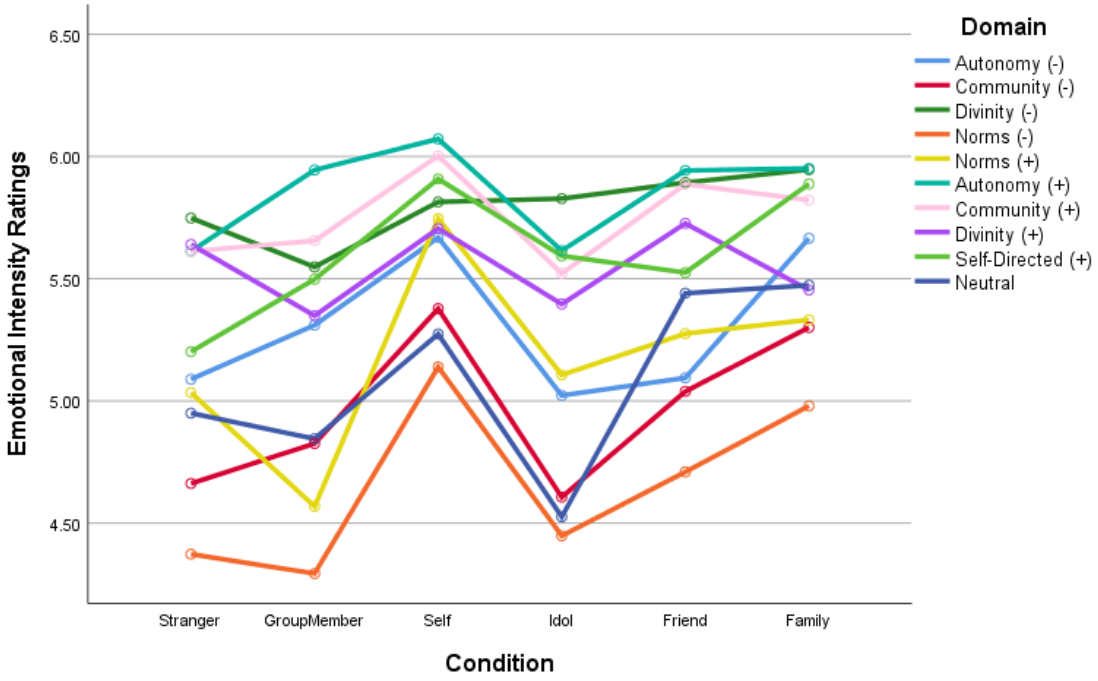


Figure 7.5. Ratings of Emotional Intensity by Domain across Conditions

Examination of categorical emotion selections in response to each scenario showed moral scenarios performed closest to expectations, with admiration the modal selection across these scenarios. Anger and disgust were also selected comparatively more often in response to autonomy and divinity scenarios respectively, although often less frequently than disappointment and concern. However, the results were not supportive of links between contempt and community violations, nor of specific emotions being elicited in response to violations committed by the self or loved ones; although the results do suggest that the relationship between observer and perpetrator may be a factor in determining which emotions are elicited. Table 7.4 shows the average number of emotion selections, and the three most frequent emotion selections, made in response to each scenario across all conditions; more detailed breakdowns of these selections are available on the pre-registration site for the study.

Table 7.4. Most frequent emotion selections across all conditions.

#	Scenario	Avg. # Selected	Top Selection	Second Selection	Third Selection
IMMORAL					
	AUTONOMY (-)				
1	CRUEL	3.17 (2.10)	Disappointment	Anger	Concern
11	WALLET	3.29 (2.20)	Disappointment	Concern	Anger
	COMMUNITY (-)				
2	LEAVEWORK	2.52 (2.21)	Disappointment	Nothing	Concern
12	TESTIFY	3.31 (2.20)	Concern	Anxious	Disappointment
	DIVINITY (-)				
3	SEXBUY	3.23 (2.33)	Disappointment	Disgust	Concern
13	CANNIBAL	3.92 (2.12)	Disgust	Gross	Concern
NON-MORAL					
	NORMS (-)				
4	LOUDMUSIC	2.23 (2.23)	Nothing	Disappointment	Embarrassment
14	BUMP	1.55 (2.19)	Nothing	Disappointment	N/A
	NEUTRAL				
10	BUSTRAIN	1.32 (2.10)	Nothing	N/A	N/A
20	BOOKREAD	2.09 (2.21)	Nothing	Amazed	Admiration
	NORMS (+)				
5	POLITE	2.21 (2.32)	Nothing	Happy	Admiration
15	MAKEDRINK	2.27 (2.14)	Nothing	Happy	Appreciation
MORAL					
	AUTONOMY (+)				
6	KIDNEY	3.96 (1.95)	Admiration	Amazed	Pride
16	FEED	3.36 (2.03)	Admiration	Happy	Pride
	COMMUNITY (+)				
7	BLOODBANK	3.42 (2.09)	Admiration	Happy	Pride
17	CHARITY	3.83 (1.94)	Admiration	Amazed	Happy
	DIVINITY (+)				
8	BEACH	3.27 (2.13)	Admiration	Pride	Appreciation
18	PLANTING	3.25 (2.11)	Admiration	Amazed	Happy
	SELF-DIRECTED				
9	SKILL	3.27 (2.12)	Admiration	Happy	Pride
19	QUITHABIT	3.55 (1.99)	Admiration	Pride	Happy

Given the number and range of emotion selections available, the varying associations between these emotions and moral domains, and potential variance in selection choices across conditions, responses were also sorted into pre-specified emotion-type groupings which collapsed emotions into familial types. For example, disgust and shame are both hypothesized to be associated with 'divinity' violations and so were grouped accordingly. The results showed a uniform pattern for morally positive scenarios. Emotions hypothesized as belonging to the 'moral' group (e.g., admiration, gratitude) were selected more frequently than those of any other grouping, and emotions in this grouping also received the highest intensity ratings. Morally positive scenarios also generated a high number of emotion responses grouped under 'positive' or 'general' categories. In contrast, 'general' (e.g. concern) or 'negative' (e.g. disappointment) emotions were selected more frequently across morally negative scenarios than those associated with a particular (immoral) domain - although emotions associated with 'autonomy' were selected more frequently than those associated with 'community' or 'divinity' in response to autonomy scenarios, and emotions associated with 'divinity' were selected more frequently than those associated with 'community' or 'autonomy' in response to divinity scenarios. General negative emotions aside, the results over immoral scenarios show some support for the hypothesized associations between emotions and the moral domains of 'autonomy' and 'divinity'. Details of familial groupings, the average and overall number of familial selections made, and the average emotional intensity of these selections, are reported on the preregistration site.

Emotion elicitation patterns for immoral autonomy and divinity scenarios in each condition, which are of particular theoretical interest, were also examined in detail. The results in each condition are similar to those reported in Table 7.4., although there were a few differences of note between conditions with regard to guilt, embarrassment, and shame. In particular, guilt was elicited most frequently in the 'self' condition, and both embarrassment and shame were selected relatively more frequently in the 'friend' and 'family' (i.e., 'loved one') conditions. These results suggest support for one key

hypotheses - the emotion elicited by a moral transgression depends on the actor-observer relationship. However, the results are less clear with regard to the intersection of actors and moral domains. Disgust was typically selected more frequently than shame in the 'self' condition for both autonomy and divinity scenarios, and shame was typically elicited with similar frequency to anger and disgust in the 'friends' and 'family' group. As such, the data shows no clear support for the hypothesized actor-domain-emotion relationships - although this seems partly due to the relative dominance of the emotion-domain selections. Putting aside emotions of disappointment and concern, the most charitable interpretation of results as they relate to the hypotheses for immoral scenarios across actors and domains is shown in Table 7.5. Further details of emotion selections by condition for autonomy and divinity scenarios are available on the pre-registration site.

Table 7.5. Immoral Actions

Domain/Actor	Stranger	Self	Loved One
Autonomy	Anger	Guilt / Disgust	Anger / Shame
Community	N/A	N/A	N/A
Divinity	Disgust	Disgust	Disgust / Shame

Statistics for emotion elicitation are not reported as the high number of emotion response options combined with the infrequent selection of many of these options undermines the assumptions underlying chi-squared in many cases. However, the most apparent general trends for emotion elicitation across conditions are summarised as follows. Guilt was selected more frequently in the 'self' condition for all immoral scenarios and negative norm scenarios. Anxiety was selected more frequently in the

'self' condition for all immoral scenarios and negative norm scenarios, with the exception of the CRUEL scenario. Embarrassment was selected more frequently in the 'self' condition for WALLET, LEAVEWORK, SEXBUY, and LOUDMUSIC. Disgust was selected more frequently for strangers (than for friends or family members) in response to the CANNIBAL scenario. Confidence was selected more frequently in the 'self' condition in response to self-directed moral scenarios, with a similar trend present for the BOOKREAD scenario. Admiration was selected more frequently in non-self conditions for all moral scenarios, although this trend was somewhat weaker for the PLANTING scenario. These trends for guilt, anxiety, confidence, and admiration, remained present when exploring emotion elicitation across all scenarios by condition. This exploration also showed differences over other emotions when comparing the 'self' condition with others. Amazement, respect, shock, and surprise were all selected less frequently in the 'self' condition; whereas exhilarated, happy, humiliated, misery, and satisfaction were all selected more frequently in the 'self' condition. The analyses from which these trends are drawn, exploratory analysis of non-moral scenarios, and all other analyses, can be found on the Open Science Framework pre-registration site for this study - https://osf.io/eh7vz

7.3.4. Exploratory Analyses

The emotion elicitation patterns across non-moral scenarios were suggestive of individual differences regarding whether certain events are morally construed, as some participants responded to these by selecting (morally relevant) emotions. For each non-moral scenario, responses were separated into a dichotomous grouping dependent on whether participants selected an emotional response or the (modal) ‘nothing specific’ option. Independent *t*-tests were run to examine whether participants which selected emotional responses to non-moral scenarios rated these differently to those not selecting an emotional response. For both counter-normative scenarios, those

responding with emotions rated events as more wrong, bad, blameworthy, and deserving of punishment than those who felt ‘nothing specific’. Corollaries were also found for neutral and positive norm scenarios, such that participants reporting emotion elicitation in response to these scenarios rated them as more right, good, praiseworthy, and deserving of reward than those reporting ‘nothing specific’. These results show an association between emotions being elicited in response to an event and an event being construed as having greater moral import. Across all instances, this difference was around one scale point closer to the respective end of the scale, with differences in ratings of right/wrong and good/bad somewhat greater than those of praise/blame and reward/punishment. The results of these tests are shown in Table 7.6.

7.3.5. Details of 'Other' emotion response selection

Although participants were provided with the means to provide open-ended responses to emotion elicitation questions, this was only used 67 times in total by all respondents. Of the selections entered, several of these (named in parentheses) seem similar to available selections such as anger (annoyed, irritated), fear (agitated, scared), interest (inquisitive), inspired (impressed, motivated), and possible variants of compassion-empathy-sympathy (helpful, understanding). Other selections included conflicted, confusion, dubious, loyal, and helpless. Apologetic, and regretful, accomplished, and rewarded, were also selected, as were amused, indifferent, uninterested, and unbothered.

Table 7.6. Exploratory *t*-tests between emotive and non-emotive responses to non-moral scenarios

#	SCENARIO	Right / Wrong	Good / Bad	Praise / Blame	Reward / Punish
NORMS (-)					
4	LOUDMUSIC				
	Emotive (150)	7.97 (2.01)	7.84 (1.91)	7.76 (1.76)	6.91 (1.60)
	Non-Emotive (72)	6.88 (1.57)	6.75 (1.44)	6.90 (1.24)	6.36 (.775)
	t(220) =	4.074**	4.294**	3.716**	2.743**
14	BUMP				
	Emotive (93)	6.70 (1.85)	6.80 (1.67)	6.84 (1.39)	6.30 (1.09)
	Non-Emotive (129)	5.95 (1.49)	6.18 (1.24)	6.16 (.855)	6.03 (.672)
	t(220) =	3.352**	3.173**	4.484**	2.276*
NEUTRAL					
10	BUSTRAIN				
	Emotive (76)	2.88 (1.72)	3.00 (1.64)	4.11 (1.69)	4.42 (1.69)
	Non-Emotive (146)	4.34 (1.93)	4.18 (1.89)	5.14 (1.47)	5.34 (1.35)
	t(220) =	5.522**	4.657**	4.724**	4.391**
20	BOOKREAD				
	Emotive (134)	3.31 (1.97)	2.98 (1.75)	3.69 (1.90)	4.13 (1.85)
	Non-Emotive (88)	4.78 (1.87)	4.51 (1.86)	5.22 (1.43)	5.28 (1.40)
	t() =	(192.62) 5.640**	(178) 6.614**	(220) 6.456**	(220) 4.970**
NORMS (+)					
5	POLITE				
	Emotive (135)	2.43 (1.47)	2.61 (1.45)	3.87 (1.90)	4.27 (1.89)
	Non-Emotive (87)	3.72 (1.85)	3.54 (1.78)	5.07 (1.40)	5.14 (1.41)
	t(220) =	5.767**	4.236**	5.039**	3.658**
15	MAKEDRINK				
	Emotive (155)	2.99 (1.76)	2.70 (1.44)	3.28 (1.58)	3.63 (1.59)
	Non-Emotive (67)	4.03 (1.53)	3.94 (1.54)	4.27 (1.48)	4.57 (1.53)
	t() =	(143.57) 4.424**	(118.18) 5.614**	(132.71) 4.484**	(130.05) 4.157**

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

7.4. Discussion

The primary aim of the study was to provide a comprehensive test of Constructive Sentimentalism's hypotheses regarding emotion elicitation patterns across a range of morally valenced events, and examine whether these patterns vary depending on the observer-perpetrator relationship. A secondary aim was to investigate whether the observers' type or level of attachment to the perpetrating agent affects ratings across a range of factors thought morally relevant, such as those relating to severity, accountability, and emotional intensity. The moral valence of the scenarios used was validated by participant ratings, with immoral, non-moral, and moral groupings readily apparent over measures of right/wrong, good/bad, praise/blame, and reward/punishment. There was also a trend for morally charged scenarios to be rated as more emotionally intense than non-moral scenarios, although the selection of 'nothing specific' for felt emotion was modal across all non-moral scenarios. These findings provide general support for the most basic form of the elicitation hypothesis, in that morally charged events tend to elicit emotion in observers more often (and more intensely) than non-moral events.

Focusing on the secondary aim first, the results showed any differences between conditions on measures of right/wrong, good/bad, and general emotional intensity were non-significant. This suggests that the relationship between the observer and perpetrating agent may be unrelated to judgements of severity – actions done by strangers are just as right/good or wrong/bad as those done by family – although there was a slight tendency for immoral actions to be rated as highly 'right' in conditions representing higher levels of attachment (i.e., rarely for strangers or group members). However, there were some differences across ratings of reward/punishment, and praise/blame, by condition. Friends were deemed to be generally less deserving of reward/punishment than oneself, and the same applies for consideration of

praise/blame. This difference was also found in other conditions, such that actions done by oneself were also considered more praiseworthy or blameworthy than those done by strangers, group members, or family members. That the majority of differences across agent-attachment condition are centred around praise and blame is particularly noteworthy considering proposals that emotions of (dis)approbation (i.e., blame) are constitutive of moral judgements, and the observer-agent relationship is a key contextual factor in determining emotion elicitation (Prinz, 2009).

With regard to the other key factor, moral content, there were fairly consistent differences in the ratings given to immoral scenarios. Autonomy scenarios were rated as more wrong, and bad, than divinity scenarios; and divinity scenarios were more wrong, and bad, than community scenarios – but community scenarios did not differ significantly from scenarios depicting negative norm violations on these dimensions. Similarly, for ratings of blame and punishment, autonomy scenarios scored higher than divinity scenarios. However, on these dimensions divinity scenarios were rated similarly to community scenarios, but community scenarios received higher ratings than scenarios covering negative norms. This might be taken (suggestively) as supporting the contention that the community domain is derived from the two primary domains (autonomy and divinity), and it may be pertinent to note that divinity scenarios were rated as more emotionally intense that autonomy or community scenarios – also suggesting support for this being a primary domain. However, that the highest ratings were otherwise consistently given to autonomy scenarios may be taken to support the notion that concerns in this domain (especially 'harm') are perhaps the most morally salient - and are in keeping with research which suggests violations of divinity (i.e., impurity) are less severe (Gray & Keeney, 2015a). It is of particular interest that the only consistent interaction effect was found at a point of theoretical contention – specifically concerning self-directed wrongs and how such wrongs relate to the concept of impurity (for examples see Chakroff et al., 2013; Dungan et al., 2017). The results of this study show that, in contrast to other-perpetrated divinity violations, self-perpetrated divinity

violations are rated much more like violations of autonomy. Further investigation of such 'self x impurity' interactions may provide a useful line of enquiry for future research.

7.4.1. Morally Positive Scenarios

Turning to the primary aim, the results over morally positive scenarios were broadly in line with expectations. Any differences between any of the four morally positive scenario pairs were non-significant, meaning all types of morally positive action were rated as similarly right, good, praiseworthy, rewardable, and emotionally intense. Also, the results for emotion elicitation within familial categories were similar across positive moral scenarios. Emotions hypothesized as relating to morality were the modal selections for all these scenarios, being selected around 50% more often than emotions considered merely positive; and the intensity of emotional response was always higher (albeit not significantly) for the moral family emotions than for general positive emotions. Furthermore, many of the individual emotions selected are ones featured in the hypotheses. Across all conditions combined, admiration was the modal selection for all positive moral scenarios. Pride, respect, compassion, inspiration gratitude, and appreciation were also frequently selected in response to each morally positive scenario – with happiness and amazement the most common 'non-moral' emotions selected. The apparent lack of significant differences across morally positive scenarios may provide some support for suggestions that moral domains exert no particular influence on emotions elicited in response to morally positive events. There were also occasional differences in the prevalence of certain positive emotions (e.g., happiness, satisfaction), and moral emotions (e.g., admiration) when comparing the 'self' condition with others, supporting notions that emotion elicitation may depend on who is performing the action.

The results across conditions for individual emotions fit, to some extent, with the predictions of Constructive Sentimentalism. The majority of scenario by condition

combinations depict actions done by other agents to other patients, to which admiration is the predicted response. This emotion was modal across all morally positive scenarios, and was selected substantially less often in the 'self' condition. Gratitude is predicted to arise in actions done by another agent which benefit oneself, and although this combination was not tested directly, the scenarios which elicited gratitude most frequently are particularly suggestive. The BEACH, PLANTING and BLOODBANK scenarios all conceivably link to some form of benefit to oneself from another, whereas the FEED, CHARITY, SKILL, and QUITHABIT scenarios do not readily confer any such benefits. On average, the former three scenarios elicited gratitude almost twice as frequently as the latter four scenarios.

However, neither dignity nor elevation was selected particularly frequently in any scenario by condition combination, including the 'self' by self-directed morally positive combination which is hypothesized to elicit 'dignity' in particular. For all morally positive scenarios, pride, respect, and inspiration were each selected more often than either dignity or elevation. Both respect and inspiration were selected slightly more often in non-self conditions, suggesting neither of these could substitute for dignity. Pride may have some potential in this regard, as family members elicited pride more often than strangers, and this may scale loosely with attachment (i.e., more attached = more pride). Alternatively, at best, the hypothesized emotion elicited by 'self-self' actions (dignity) might be substituted with a feeling of confidence. In comparison to the other conditions, confidence was selected around twice as often within the ‘self’ condition in response to the self-directed morally positive scenarios, as well as tending to be selected comparatively more often within the ‘self’ condition generally.

7.4.2. Non-moral Scenarios

The results for scenarios classed as non-moral were similar to (im)moral scenarios in pattern, but not in strength, and diverge with regard to moral valence. Although ‘nothing specific’ was the modal selection across conditions for all these scenarios, the next most selected emotions selected in response to ‘positive norms’ and ‘neutral’ scenarios were similar to those selected for moral scenarios – albeit with ‘general’ or ‘positive’ emotions being selected somewhat more often than ‘moral’ emotions. These scenarios were also rated similarly across non-categorical measures, such that although ‘positive norms’ always scored higher than ‘neutral’ scenarios on these measures, any differences were non-significant. In contrast, ratings given to scenarios depicting ‘negative norm’ violations differed substantially from ‘neutral’ scenarios on all measures except emotional intensity. Over the two negative norm scenarios, disappointment was the most frequent selection after 'nothing'. However, it is noteworthy that the unmitigated negative norm scenario, LOUDMUSIC, also elicited similar emotions to the immoral scenarios, albeit also less frequently, and less intensely. That positive and negative norms seem to elicit a similar-but-weaker response to their moral and immoral counterparts is suggestive of individual differences with regard to moral construal – such that some participants may be considering such actions as not merely normative. Exploratory analyses of non-moral scenarios showed that participants who selected an emotion, rather than 'nothing specific', in response tended to rate these scenarios as closer to the respective ends of the scale across all non-emotion measures. This finding further suggests that individual differences in moral construal may be driven, at least in part, by emotions.

7.4.3. Immoral Scenarios

The results for emotion elicitation over immoral scenarios are open to interpretation with regard to the extent they may be taken as supporting Constructive Sentimentalism (Prinz, 2009) and Moral Foundations Theory (Graham et al., 2013). Violations of community-based ethics are hypothesized to elicit contempt, yet contempt was not selected by >10% of participants for any scenario. However, this may be due to issues both with eliciting contempt and in capturing verbal reports of this emotion (Matsumoto & Ekman, 2004), as well as potential issues with the scenarios used. The LEAVEWORK scenario was the most un-emotive of the immoral scenarios, with only disappointment selected more frequently than 'nothing'; and the TESTIFY scenario, in hindsight, may have contained some form of moral dilemma vis not testifying in order to minimize potential risks to loved ones – placing community ethics in conflict with a stronger 'protect those you care about' rule. Thus, although any link between 'community' and contempt is not apparent in the data, there are good methodological reasons which explain why this may be the case. These aside, an absence of such a link may also be explained by making small tweaks to theory. For example, it may be that the community domain, being a derived domain, is comparatively weaker and/or less salient than 'autonomy' or 'divinity'. This fits with the results showing community scenarios were rated as less wrong and bad than other immoral scenarios, and rated as similarly wrong and bad to negative norm violations, but were comparable with other immoral scenarios with regard to measure of blame and punishment. However, future research on the hypothesized community-contempt link may be best left until more commonly hypothesized links have been more thoroughly investigated.

Results for the CRUEL and WALLET scenarios provide some degree of support for the hypothesis. Across conditions, disappointment was the modal selection for both scenarios. Given that moral wrongs, almost by definition, involve a negative violation of expectation (e.g., a norm violation), the elicitation of disappointment is perhaps

unsurprising - and its modal response position unproblematic. Anger and concern were the next most frequently selected emotions, with anger chosen somewhat more often than concern for CRUEL, and just slightly less than concern for the (mitigated) WALLET scenario. The selection of concern may be more open to interpretation than disappointment, but may also be classed as definitional on Constructive Sentimentalism's approach – emotions represent concerns (Prinz, 2004b). Therefore, of the emotions hypothesized as morally relevant, anger was the most frequent choice in response to violations of autonomy. A similar response pattern was found for the SEXBUY scenario, with disgust (as hypothesized) a close second behind disappointment, and gross a close fourth behind concern. However, for the CANNIBAL scenario, disgust was the modal selection, followed by gross, and then concern. Sick and repulsed were also selected relatively frequently, such that combining selection numbers for these two emotions would put the combination into third place just behind gross. These findings are echoed in analysis of emotion families, which show that of the emotions theorised as associated with autonomy and divinity violations were the most frequently selected *moral* emotions (i.e., not just 'negative' or 'general'). Although there are several caveats to be discussed, these results may also provide prima facie support for the association between disgust and violations of divinity.

Emotion responses across attachment conditions do fit with theoretical predictions to some extent, but also suggest some revision of theory may be beneficial. Guilt was selected significantly more often in the 'self' condition, but was elicited across all negative scenarios to a similar extent within this condition. There was no obvious association of guilt with autonomy violations, and guilt was elicited to a similar extent by both immoral scenarios and those depicting negative norm violations. Embarrassment was also selected significantly more often in the 'self' condition in comparison to all other conditions except family, but similarly, there was no apparent association between embarrassment and any specific moral domain. Also, any differences with regard to selecting shame across conditions were non-significant; although the data hints that this

may be found to correlate with attachment levels (i.e., more attached = more shame) in future studies. That shame and embarrassment are selected comparatively often between ‘self’ and ‘family’ conditions may be taken to provide some support for changes to elicitation patterns hypothesized across attachment conditions, but there is no obvious support for any of guilt, embarrassment, or shame, as being exclusively associated with violations of a particular moral domain.

Support for domain associations is not readily apparent even if allowing familial emotion associations to be combined for the ‘self’ condition. Combining (self-)anger and guilt, and (self-)disgust with shame, would make each combination modal for autonomy and divinity scenarios respectively. However, doing this fairly also brings the opposing combination to a level almost equal for autonomy scenarios, although not for SEXBUY, or for CANNIBAL if disgust related emotions are included. This is in contrast to the results across conditions where hypothesized emotions tended to be selected more often (by at least a 3:2 ratio) for both autonomy and divinity scenarios. Furthermore, combining emotion selections of embarrassment and disgrace, both of which could be argued as relating to community violations, would make this combination modal for the WALLET and SEXBUY scenarios. Thus, although the emotions hypothesized as relevant to self-committed moral violations were selected frequently in this condition, there was no apparent correspondence between these emotions and moral domains.

7.4.4. Summary

Overall, the extent to which these findings may be taken in support of particular hypotheses depends on the interpretation of results. Morally charged events do seem to elicit emotions more frequently, and more intensely, than non-moral events; although finding support for this most basic form of the elicitation hypothesis is expected on all the theories discussed. Similarly, the results provide some support for the ‘primacy of harm’

hypothesis, as autonomy scenarios were rated as more wrong, bad, blameworthy, and deserving of punishment, than other immoral scenarios – although again, this hypothesis is not controversial. Furthermore, there is some evidence that the relationship between agent and patient affects emotional responses to events, but this might also be expected (or at least not ruled out) by the theories under discussion. However, apart from these few general points of agreement, the remaining findings seem to favour different aspects of different theories.

Morally positive events were all rated similarly across non-categorical measures, in contrast to immoral events, which suggest support for claims that 'foundational content' is not particularly relevant for morally positive events (Prinz, 2009), and fits with the help-harm dichotomy proposed by TDM (Schein & Gray, 2018). Some differences between these ratings may have been expected on an MFT-based approach, although as these scenarios have not been validated, it is possible to argue these all relate to a single foundation (e.g., care) – although doing so would seem to preclude MFT's claims of positive emotion-domain associations. However, single foundation claims may also explain why there is no apparent corollary for a 'primacy of care' in the results - although the scenarios differ to some extent with regard to agency, such that scenarios with clearly identifiable beneficiaries might have been expected (extrapolating from TDM) to receive higher ratings than those without.

Categorical results for emotion show that 'other-praising' emotions (Algoe & Haidt, 2009) were frequently elicited in response to morally positive events, and tend towards the predictions of CS. Admiration was the modal response across all morally positive scenarios, and selected substantially less frequently in the 'self' condition, both of which match with predictions from CS. However, a harsh critic might suggest that the apparent strength of this pattern might be influenced by admiration being first on the list of emotions. Gratitude was also selected to some extent, and the elicitation pattern of this particular emotion is tentatively in keeping with CS's hypotheses - although

appreciation was also selected quite frequently across all morally positive scenarios. This could be considered a form of gratitude, which may weaken claims that this emotion is specific to morally positive events providing some self-benefit – although it could also plausibly be argued that appreciation is merely a positive corollary of disappointment. Inspiration was also selected quite frequently across morally positive scenarios, which fits well with predictions made via TDM (see Gray & Wegner, 2011). However, although pride and respect were also selected with moderate frequency, there was no evidence to support these emotions as being characteristic of a particular foundation – as proposed by MFT. Lastly, that emotions such as elevation (TDM & MFT) and dignity (CS) were not frequently selected might be explained by difficulty in eliciting relatively complex emotional states from simple moral scenarios. This issue might also be explained by language use, in that neither word is commonly used by people describing their emotional states (e.g., "I feel elevated", "I feel dignified"), and future studies may benefit from using different terminology for these emotions (e.g., "uplifted" instead of "elevated"). Overall, the results for morally positive scenarios fit better with predictions of CS and/or TDM than those of MFT.

For immoral scenarios, leaving aside issues with community violations, and contempt, the results are similar to those found in other research (see Cameron et al., 2015 for review). Violations of autonomy elicited anger more frequently than they did disgust, and violations of divinity elicited disgust more frequently than they did anger, which fits with the associations predicted by both CS and MFT. That this association appears despite a lack of forced choice responses is promising for advocates of these theories, and provides some challenge the conclusions of Cameron et al. (2015) regarding specificity of emotions. Objections that these emotions were not modal selections can be readily overcome, although further investigation using different scenarios may provide better support for specificity. Although both divinity scenarios are not under dispute with regard to moral content per se (i.e., they are examples of 'impurity' violations), it could be argued that the content of these scenarios elicit non-

moral forms of disgust because they focus on sex and the consumption of human remains. This might be countered by arguing that (moral) disgust was selected more often than gross (non-moral disgust), but replication using impure scenarios which do not contain core/pathogen or sexual disgust elicitors may provide better evidence in this regard.

With regard to self-committed moral wrongs, there was no apparent association between emotions and moral domains as predicted by Constructive Sentimentalism. This may be due, at least in part, to a differing ecological setup within the 'self' condition. Having someone find out about self-committed moral wrongs may promote more socially appropriate emotions (i.e., embarrassment, shame) than those which may be elicited in a less social setup (i.e., suppose you…), and that anxiety was selected substantially more often in the 'self' condition is also suggestive in this regard. Alternatively, some revision of theory may be required, although it is an open question as to whether such revision might contain emotion-domain associations. It may be that emotion elicitation (partly) depends on severity, such that guilt and embarrassment tend to be elicited by relatively moderate transgressions, whereas shame and disgrace may be elicited by more severe moral violations. However, it may be that self-committed divinity violations also elicit disgust, rather than shame, because committing such a violation would not be in keeping with the self-perceived 'natural order'. The refrain "I was not myself when I did that" shows how typically other-directed emotions may turn inward, and this may be particularly common with regard to disgust. Apart from the hypothesized link with natural order violations, there may also be a conceptual link with contamination. For example, "something bad took over" or something along the lines of "I want to remove whatever part of me that was responsible for doing that". The results are slightly suggestive toward this latter possibility, as it is notable that disgust was generally selected more often than anger within the 'self' condition.

7.4.5. Conclusions

The findings of this study fit well with previous literature investigating emotions across immoral scenarios, such as the research reviewed by Cameron et al. (2015). Although Cameron et al. suggest that, at best, correspondences between discrete emotions and specific moral content are 'loose', the results of this study are similar to those reviewed despite being collected under open-response conditions. That a stronger test produces similar evidence may be encouraging to proponents of such correspondences, although further research would be useful for establishing the precise nature of any such relationships. For example, replication using scenarios covering different varieties of harm (e.g., physical vs. non-physical), and scenarios containing different types of disgust elicitors (e.g., pathogen vs. sexual vs. moral), may provide a stronger and more informative test. More nuanced examination of disgust-impurity associations may also be useful in combination with further investigation of how these interact with self-committed (impure) wrongs. Transgressions at this intersection appeared to have a ratings profile more similar to violations of autonomy, but an emotion elicitation profile closer to violations of divinity. Additional exploration in this area may help explain why the hypothesized emotion-domain associations for self-committed violations (and derivatives of these) did not materialise clearly.

Overall, the results seem slightly more favourable towards Constructive Sentimentalism than other theories. Explanations for the appearance of ‘loose’ correspondences over immoral scenarios aside, results across morally positive scenarios were much closer to predictions. Admiration was the modal response to all these scenarios, and notably lower in the ‘self’ condition, which would be expected under Constructive Sentimentalism. Also, no specific (positive) emotion-domain associations were apparent in the data, contra Moral Foundations Theory - although the hypothesized emotions were selected with moderate frequency. Additionally, despite some scenarios likely being much more ‘dyadic’ than others, ratings across positive

scenarios did not seem to differ, and inspiration was only selected with moderate frequency, in contrast to what might be expected under the Theory of Dyadic Morality. Furthermore, there is also some suggestion that emotion elicitation varies depending on the relationship between the observer and the agent, which neither MFT nor TDM address sufficiently. Moreover, that the only differences in ratings across conditions were focused on praise and blame also fits better with Constructive Sentimentalism, which emphasizes these dimensions over considerations such as ‘dyadicness’ or harmfulness (vis TDM). Finally, that compassion, empathy, and sympathy, all victim focused-emotions, were not elicited with greater frequency lends further support to Prinz's claims that it is agent-focused emotions which matter for moral judgement. Thus, although Constructive Sentimentalism may benefit from further testing, and theoretical refinement, it seems best able to account for the pattern of results obtained in this study.

Chapter 8 - Sound Morality Extended – Effects of emotion induction on judgements across Moral Foundations

Constructive Sentimentalism (CS; Prinz, 2009) argues that moral judgements are *constituted* by emotions, such that moral emotions are constructed from non-moral emotions. CS also argues in favour of correspondences between specific emotions and certain types of moral content, similarly to Moral Foundations Theory (MFT; Graham et al., 2013) which hypothesizes 'characteristic associations' between emotions and moral content; and both these theories advance claims about the roles of emotion in moral judgements. Firstly, moral violations are hypothesized to *elicit* emotions, and although debate continues over why this may be the case, the claim itself seems generally accepted. Secondly, the experience of emotion extraneous to the moral event is hypothesized to *amplify* moral judgements – along with a corollary hypothesis that suppressing the physiological underpinnings of emotions (e.g., nausea-disgust) suppresses moral judgements. Lastly, the experience of extraneous emotion is hypothesized to influence *moralization*, such that novel or relatively benign events may come to be regarded as being more morally relevant through combination with emotions. These three hypotheses, particularly under Constructive Sentimentalism, are argued to operate *specifically*. Violations of harm elicit anger, and are judged more harshly and/or moralized as harmful if anger is induced; impure violations elicit disgust and are judged more harshly and/or moralized as impure if disgust is induced. Stronger versions of these hypotheses argue that these links are also *exclusive*, such that each emotion only operates within its respective moral domain – inducing anger should have no effect on judgements of impurity, and inducing disgust should have no effect on judgements of harm.

Recent reviews have not been favourable to these hypotheses. Firstly, Cameron, Lindquist and Gray (2015) reviewed a range of studies relating to claims of specificity/exclusivity. They conclude that there is no good evidence for exclusive links between emotions and moral content, and challenge claims of specificity through detailing how any apparent appearance of this might be explained by a combination of common emotional components and methodological confounds. Secondly, Landy and Goodwin (2015a) conducted a meta-analysis to investigate the extent to which incidental disgust amplifies moral judgement. They conclude any amplification effect for disgust is likely to be small (d = .11), and not specific to violations of purity. Landy and Goodwin further suggest the effect may be non-existent once accounting for both publication bias and for confounds additional to those described by Cameron et al. (2015) - although they do suggest there is preliminary support for a small moralization effect of disgust on neutral actions. Taken together, these reviews provide a strong challenge to some of Constructive Sentimentalism's main hypotheses.

However, there are potential issues with both these reviews (for detail see 'On Emotion Specificity'). Cameron et al.'s argument relies on methodological issues they identify acting to confound results, as well as the premise that there are morally salient actions which relate to *only one* moral foundation. These notwithstanding, several of the studies reviewed show some support for 'weak' specificity - anger is elicited relatively more often than disgust over harmful violations, and disgust relatively more often than anger over impurity violations. Additionally, research conducted after this review has also shown some support for specificity (Franchin et al., 2019; Landmann & Hess, 2018; Heerdink et al., 2018), and even exclusivity (Tracy et al., 2019). Furthermore, if Cameron et al.'s (2015) conclusions regarding exclusivity are correct, then the amplification effect size calculated by Landy and Goodwin may be understated given they treat sadness and fear induction as control conditions. In contrast, if Cameron et al. are incorrect about specificity, then the number of multiple-content violations used in studies reviewed by Landy and Goodwin may be acting as a confound, resulting in both

overestimating the effect of disgust on non-purity transgressions and underestimating the effect for purity violations. Finding common ground to test these assertions may provide a particularly informative line of inquiry.

Of the studies examined by Landy and Goodwin, the largest effect size in favour of disgust-impurity associations is included via Eskine, Kacinik and Prinz (2011) who report consuming a bitter beverage amplified participants' moral judgements. However, this study is identified by Cameron et al. (2015) as failing to control for 'core affect' (e.g., valence) and 'conceptual content' (e.g., the taste could evoke disgust *or* offence). Additionally, with regard to moral content, Landy and Goodwin (2015a) report the effect in this study is not exclusive to purity violations. Furthermore, a large-scale direct replication of the original study (Ghelfi et al., 2020) both (re-)estimates the effect size as being within the bounds proposed by Landy and Goodwin (2015a), and contrasts with the original in showing a similar amplification effect for sweet beverages. This finding further strengthens the case against claims of specificity/exclusivity.

The next largest favourable effect size, and the largest in favour of exclusivity, is contributed by Seidel and Prinz (2013a). Using a two factor between subjects design, Seidel and Prinz (2013a) report aural induction of anger, but not disgust, amplified judgements relating to autonomy; and aural induction of disgust, but not anger, amplified judgements regarding purity. However, Cameron et al., (2015) argue this study suffers from methodological confounds. Specifically, disgust is induced by listening to the sound of an emetic event, but two of the three purity scenarios involve oral consumption - a conceptual similarity which may have influenced the results. Cameron et al. (2015) proposes that "if such [emotion-content] correspondences do exist, they can be found using the following experimental framework that is inspired by constructionism: a 3 (Core affect: negative, neutral, positive) × 3 (Conceptual knowledge: anger, disgust, unrelated) × 3 (Judgment type: harm, purity, non-moral) design with core affect and conceptual knowledge manipulated between subjects and judgment type manipulated within

subjects." (p.17). Furthermore, they suggest the inclusion of other emotions to examine differences relating to emotions which vary in terms of arousal and/or valence (e.g., happiness, sadness, fear).

8.1. The current study (Study 3)

Seidel and Prinz's (2013a) research is identified by Cameron et al. (2015) as that which most challenges their conclusions regarding emotion specificity, and Seidel and Prinz's methodology contributes the largest effect size in favour of exclusivity to Landy and Goodwin's (2015a) meta-analysis. Extending Seidel and Prinz's (2013a) study design to fit the framework proposed by Cameron et al. (2015), whilst accounting for potential moderators and confounds identified by Landy and Goodwin (2015a), thus offers fertile common ground for testing the various hypotheses relating to amplification effects, specificity/exclusivity, and moralization.

The primary aim of the study is to examine claims of emotion specificity and/or exclusivity across moral foundations through testing whether emotion induction amplifies moral judgements and/or moralizes non-moral judgements. In doing so, this study extends on 'Sound Morality' by examining moral content as a within-subjects factor, as well as by using a wider range of moral content - covering six types relating to Moral Foundations Theory (Graham et al., 2013), as well as 'counter-normative', and 'non-moral' content. This study also extends along the range of induced emotions to fit with Cameron et al.'s use of 'core affect' between subjects in their experimental framework. These include four negatively valenced conditions - anger, disgust, fear, and sadness - in addition to a neutral (control) condition, and a positively valenced (happy) condition. However, their proposed between-subject factors for conceptual knowledge are not included in this study as the priming of conceptual knowledge is largely avoided by inducing emotions using sound.

The secondary aim of the study is to examine and account for 'Private Body Consciousness'. This is argued act as a moderator on amplification effects (cf. Schnall et al., 2015, Johnson et al., 2016), whereby those reporting greater sensitivity to bodily states (i.e., higher interoceptive awareness) tend to experience an amplification effect whereas those reporting lower sensitivity to bodily concomitants of emotion do not. This study also extends on work in this area through including the Multidimensional Assessment of Interoceptive Awareness (MAIA - Mehling et al., 2012) to provide a comparison with Private Body Consciousness and allow for exploration of how different dimensions of interoceptive awareness might relate to moral judgements.

The tertiary aim of the study is in seeking to explore 'offence-at-materials' confounds (experimenter effects) which Landy and Goodwin (2015a) suggest may affect the results via displaced affect. They argue it is possible that any amplification effects may stem from moral disapproval towards the researcher or research process, arising as a result of having undergone a negative emotion induction method. Moral disapproval may be enhanced as "the experimenters are knowingly doing some small harm to their participants" (Landy & Goodwin, 2015a, p30), and it is this, rather than the emotion induction, which produces any amplification effects.

There are multiple hypotheses for the study (Table 8.1), which are reflective of different theories and the multitude of findings in the extant literature. All theories under consideration predict induction of either anger or disgust should result in increased ratings of wrongness. Moral Foundations Theory (Graham et al., 2013) in its stronger form predicts any increase in ratings should be greatest in response to scenarios depicting violations of ‘care-harm’ or 'fairness-cheating' when inducing anger, and greatest in response to ‘sanctity-degradation’ violations when inducing disgust. There may also be some influence of anger induction over violations of 'loyalty-betrayal' (via

traitor directed rage), and uniquely, fear induction may influence judgements relating to 'authority-subversion' violations. A weaker form of the theory predicts anger induction should increase wrongness ratings over violations of the 'individualizing foundations' (e.g., 'harm', 'cheating'), whereas disgust induction should increase ratings for 'binding foundations' violations (e.g., 'betrayal', 'subversion', 'degradation').

Constructive Sentimentalism (Prinz, 2009) makes the same predictions as Moral Foundations Theory with regard to anger and disgust; and both theories suggest scenarios covering the ‘fairness-cheating’ foundation may garner similar responses to those covering ‘care-harm’, which together cover violations against persons/violations of autonomy. However, this theory suggests there may be some effect of anger or disgust induction across ‘loyalty-betrayal’ and ‘authority-subversion’ foundations because violations of the natural order of persons (community violations) are related to contempt, and contempt is a blend of anger and disgust (Prinz, 2009). This differs from Moral Foundations Theory, which details no particular associations between anger and 'authority-subversion' violations, and makes potentially differing predictions about 'loyalty-betrayal' violations when comparing the stronger, anger-favouring, version with the weaker one which favours disgust in relation to induction influence.

In contrast to both these theories, the Theory of Dyadic Morality (Schein & Gray, 2018) predicts induction of anger or disgust should result in similarly increased ratings across all moral foundations, given the associations of these emotions with ‘harmful agents’ (Gray & Wegner, 2011). This theory also suggests any increase in wrongness ratings should be particularly apparent in response to scenarios depicting violations of sanctity which involve ‘oral activities’ (e.g., eating the flesh of a deceased relative as part of a group funeral rite), as the method of inducing disgust involves sounds depicting another ‘oral activity’ (vomiting) which is hypothesized to prime related conceptual knowledge regarding the body (Cameron et al., 2015). Furthermore, in citing Cheng et al., (2013), Dyadic Morality suggests all emotion induction conditions should have similar

effects on moral judgements through shared dimensions of affective arousal. However, in citing Cameron et al. (2015), it allows that such effects may be reduced in the sadness condition (lower arousal), and may differ in the happiness condition (positive valence). Fear is considered a particularly useful comparison condition as it shares characteristics of anger and disgust, in that all are high arousal negatively valenced emotions. It may also help differentiate between the motivation tendencies of these emotions, in that both fear and disgust are commonly avoidance orientated whereas anger is generally approach orientated (Cameron et al., 2015).

Simplifying these positions to align with the experimental design and planned analyses provides the following working hypotheses. Ratings of moral wrongness provided by participants induced into an emotional state will differ from those of participants not induced into an emotional state (***amplification/suppression effects***), and this effect will also be present over ratings of counter-normative and non-moral scenarios (***moralization effect***). The effect is expected to be most apparent following the induction of emotions associated with moral judgements (i.e., disgust or anger), but less apparent or absent following the induction of an un-associated emotion (i.e., fear) which is similar in terms of valence/arousal (***emotion specificity***) - although the effect may also be present, potentially in the opposite direction, following inductions of sadness or happiness. It may further be the case that any amplification effects are only apparent in response to certain types of moral content (***emotion exclusivity***), such that anger induction may only lead to increased wrongness ratings for some categories (most notably 'harm/care') and not others (e.g., sanctity/degradation), whereas disgust induction may show the reverse pattern with increased wrongness ratings only apparent over sanctity/degradation scenarios. The emergence of any amplification/suppression effects should not be dependent on whether participants have been offended by the study materials (***experimenter effects***), although the effect may be limited to participants who are more sensitive to aspects of emotional experience related to private body consciousness and/or interoceptive awareness (***PBC/MAIA moderation effects***).

Table 8.1. Hypothesized effects according to different theoretical positions

EMOTION INDUCTION	HARM	FAIRNESS	LIBERTY	LOYALTY	AUTHORITY	SANCTITY	COUNTER-NORM	NON-MORAL	COMBINED
^ANGER^	MFT CS	MFT CS		MFT				MFT CS TDM	ALL
^DISGUST^				MFT CS	MFT CS	MFT CS		MFT CS TDM	ALL
^FEAR^					MFT				TDM
^SADNESSv									ANY
^HAPPINESSv									ANY

^ = Amplification, v = Suppression. All theories expect anger and disgust to result in amplification, MFT and CS suggest this effect may be stronger, or exclusive, over certain types of scenarios. TDM suggests any effect present for anger or disgust is likely to be present for fear given the shared dimensions of affect between these emotions. Differences due to sadness or happiness can be accommodated by any of these theories, with literature suggesting effect may vary in direction for these emotions.

8.2. Method

8.2.1. Design

The study uses a mixed design to examine the effects of emotion induction on moral judgements. There are six between subject factors covering the specific emotion induced - chosen following Cameron et al.'s (2015) recommendations. 'Anger' and 'Disgust' are the main conditions of interest, with controls included in the form of 'Fear' (another high arousal negatively valenced emotion), 'Sadness' (a negatively valenced low arousal emotion), 'Happiness' (positively valenced), and 'Control' (neutral 'core affect') conditions. There are eight within subject factors covering violations of the six 'foundations' specified by Moral Foundations Theory (Care/Harm, Fairness/Cheating, Liberty/Oppression, Loyalty/Betrayal, Authority/Subversion, Sanctity/Degradation) with two 'control' factors consisting of 'Counter-normative' and 'Non-moral' actions. Measures of private body consciousness and interoceptive awareness are included as potential moderators acting on the primary dependent measure - ratings of ‘wrongness’. Measures are also included to check the efficacy of emotion induction, as well as items aimed at examining (moral) disapproval directed towards the experimenter.

8.2.2. Participants

G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) was used to inform planned sample size design with regard to the effect sizes reported in literature. For a repeated measures ANOVA, and with the interaction effect of primary interest, a minimum *N* of 84 is required to detect an effect of *f* = .35 reported by Seidel and Prinz (2013a) with relative certainty (99% power at an alpha of .01), and a sample of 288 would have 99%

power to detect an interaction effect half this size assuming an alpha of .01. The initial target sample size was based on rounding the latter figure (i.e., *N* = 300).

However, Landy and Goodwin (2015a) suggest an effect size of *d* = .11 (equivalent to *f* = .055) is an upper bound with regard to disgust amplifying moral judgement – an effect small enough to require substantially more participants in order to provide sufficient power to the study. Alternative analysis techniques, which treat each scenario separately, might be used to achieve sufficient power for detecting an interaction effect of this size (minimum *N* = 396), but the study would otherwise require an N of 1248 to provide 80% power to detect an effect of this size assuming an alpha of .05 - slightly smaller than the *N* = 1299 per condition for a basic comparison of fully independent groups noted by Landy and Goodwin (2015a).

This wide range of estimated sample sizes follows from the difference between Landy and Goodwin's (2015a) overall effect size estimate and the effect size they report for Seidel and Prinz (2013a), *d* = .92 for disgust amplifying severity judgements of purity violations. On the assumptions that Landy and Goodwin's meta-analysis may understate the size of the effect, and that Seidel and Prinz's (2013a) results may overstate the effect size, the target sample size was increased with the aim of gaining at least twice Seidel and Prinz's (2013a) number of participants per cell (*N* = ~600). A sample of this size would provide 97% power to detect an interaction effect size of *f* = .1 assuming an alpha of .05, or 90% power assuming an alpha of .01.

The opportunity sample was recruited via adverts placed on the University research participation management system (SONA), adverts to research students, and adverts placed online (Facebook and Twitter). Participants would have self-identified as speaking fluent English, being 18 or over, and being willing to read potentially offensive or upsetting content. They would also have been advised against taking part if they had

any form of epilepsy, schizophrenia, any form of sensitivity to sound (e.g., hyperacusis), any type of phobia regarding the body or bodily sensations (e.g., emetophobia), or any form of eating disorder (e.g., bulimia). Students were offered 0.5 Research Credits for completing the study. The research for this project was submitted for ethics consideration under the reference PSYC 18/ 315 in the Department of Psychology and was approved under the procedures of the University of Roehampton's Ethics Committee on 24.10.18.

8.2.3. Materials

All scenarios used are identical to those reported in Landy and Bartels (2018), with seven scenarios in each content category. Accordingly, violations of each moral foundation have been validated as representative of such, and groupings have been 'pre-normed' on measures of wrongness so as to minimise floor and ceiling effects (5.23 ± 0.03 on a 9-point scale). Furthermore, the range of content across their impure scenarios provides a way to investigate claims regarding the activation of conceptual knowledge, which is difficult to avoid when inducing disgust. Two of the scenarios directly evoke oral concepts, whereas two scenarios do not contain any obvious (non-moral) disgust elicitor. This provides a means of quantifying any effect of disgust-related conceptual activation on moral judgements of impurity. Ratings for all scenarios are taken on a 9-point scale (following Landy & Bartels, 2018), with the end points labelled 'Perfectly okay' and 'Extremely wrong' (following Seidel & Prinz, 2013).

All 'sound' stimuli have previously been validated for emotion induction purposes. The sound in the anger condition was the first track from 'Inner Mind Mystique' (Takushi, 1996), and the disgust condition sound was that of an emetic event -- both of which are reported as effective by Seidel and Prinz (2013a, 2013b). The sound in the fear condition was 'Threnody to the Victims of Hiroshima' by Krzysztof Penderecki

(2012/1960), and the happiness condition sound was Edvard Greig's 'Morning Mood' (1993/1875) - both of which are reported as effective by Prinz and Seidel (2013b). John Dowland's 'Semper Dowland, semper Dolens' was selected for the sadness condition based on research by Kreutz, Ott, Teichmann, Osawa, and Vaitl (2008); whilst for the 'sound' in the control condition, participants were told they would be listening to the (silent) composition 4'33" by John Cage.

A self-report measure of feelings was used to investigate the success of emotion induction. This consists of 18-items covering each of the emotion induction conditions, with ratings running on a 7-point scale ranging from 'very little felt' to 'very much felt'. Items have been included to cover 'Anger', as well as 'Annoyed' and 'Irritated' (following Seidel & Prinz, 2013a); 'Grossed Out' for 'Disgusted', with 'Revolted' also included to allow for different proxy measures of disgust. 'Fearful' is covered in conjunction with 'Afraid' and 'Anxious'; 'Sad' in conjunction with 'Heavy-hearted' and 'Gloomy'; and 'Happy' in conjunction with 'Uplifted' and 'Cheerful' (following Seidel & Prinz, 2013b). 'Calm', 'Relaxed', and 'Peaceful' are included as neutral items. Scores for participants self-reported emotion state are formed from their respective three-item composites.

A self-report measure of Private Body Consciousness (PBC - Miller, Murphy, & Buss, 1981), consisting of 5 items scored on a 6-point scale from 'disagree strongly' to 'agree strongly', was included following Schnall, Haidt, Clore, and Jordan (2008). The Multidimensional Assessment of Interoceptive Awareness (MAIA - Mehling et al., 2012) was also included to provide a comprehensive measure of interoceptive ability. This 32-item measure is rated on a 6-point scale running from 'Never' to 'Always'.

Following Landy and Goodwin's (2015a) recommendation, an exploratory measure was used to investigate how participants felt towards the experimenter as a

result of completing the study. Ratings were given on a 7-point scale ranging from ‘Much more negatively’ to ‘Much more positively’. Responses given on the negative side of the scale prompted two follow up questions. The first asked about whether the cause of negativity was the scenarios, the sound, or something else; the second asked for intensity ratings of negative emotions felt towards the experimenter. Ratings were given on an 8-point scale, ranging from ‘not at all felt’ to ‘very much felt’, over the six items related to ‘Anger’ and ‘Disgust’ from the manipulation check items, as well as items for ‘Contemptuous’ and ‘General negativity’.

8.2.4. Procedure

Participants were presented with a link to the study, which was administered on Qualtrics. Participants were asked to provide consent, and demographic information (i.e., age/sex/fluency in English). Instructions were presented requesting participants undertake the study in a private setting with a stable internet connection, and to wear headphones if these were available to them. A test question was used to allow participants to check their sound and volume settings, and to serve as a validity check, asking which one of four instruments they could hear being played (Guitar*, Piano, Flute, Trumpet). Participants were randomly allocated by Qualtrics to one of the six conditions, given a brief description of the sound they would be listening to, and then asked to 'take at least a minute to focus on and become familiar with these sounds before continuing the study' -- with the 'next' button unavailable for 60s. Scenarios were allocated (list-wise, as reported in Landy & Bartels, 2018) into blocks, 1 of 8 items and 3 of 16 items, each respectively containing 1 or 2 scenarios relating to each of 8 types of violation, such that participants would typically rate 16-items per page. The order of blocks, and of the items within blocks, were both randomised by Qualtrics. Once all 56 scenarios had been rated, participants were asked to self-report on their feelings via the MAIA and PBC (items on each scale were listed in a random order). They were also asked how

they felt towards the study researcher, whether they experienced any technical difficulties during the study (e.g., with sound), and asked to confirm that they continued to listen to the sound requested for the duration of the study.

8.2.5. Pre-registration

A priori power calculations, statements of hypotheses, planned analyses, and all study materials are available via the pre-registration site for this study - https://osf.io/u8n4w/

8.3. Results

8.3.1. Response validity checks

A total of 569 responses were collected. Nine cases were removed as the result of participants completing the study more than once, with the latter completion(s) being removed each time. Five partial responses were also removed where participants had failed to meet basic demographic criteria (e.g., not speaking fluent English) and had thus been unable to proceed with the study.

43 cases were removed because participants stated they did not follow the instruction to listen to the sound throughout entirety of the study. A further 87 cases which would have been removed for meeting similar criteria (e.g., technical issues, incorrectly identifying a test sound) which may have affected their response to the instruction check question were retained to form an additional control group for the study. Planned validity checks based on response times and ratings of non-moral items were discarded as times may have been affected by technical issues and ratings of non-moral items showed considerable variation - removing cases meeting these criteria would have cut the sample size even further. The only other change to the pre-registered plan was initially extending the point at which data collection would be stopped in order to allow for a larger sample to be collected, with the aim of reaching double number of participants per cell in the original study by Seidel and Prinz (2013a). However, this extension was cut short due to the onset of a global pandemic, as any responses after this point may have been confounded as a result. Having such a salient pathogenic presence in the environment may have affected any responses linked to disgust given the link between this emotion and pathogen avoidance mechanisms.

8.3.2. Scale Reliability

The 512 responses considered to have passed validity checks were processed using the pre-registered scripts. These computed scores for each of the scales, ran reliability checks, and produced analyses relevant for addressing the hypotheses.

Scales relating to each of the judgement categories showed moderate reliability; harm (α = .71), fairness (α = .69), liberty (α = .7), loyalty (α = .57), authority (α = .76), sanctity (α =.75), counter-normative (α =.6), non-moral (α =.89). Sub-scales of the sanctity scenarios relating to potential confounds displayed lower reliability (α =.63) than their combination. Reliability scores for all other measures tended to be similar or better (MAIA, α = .89; PBC, α = .73), particularly those designed to check that emotion induction had achieved the desired effects; anger (α = .9), disgust (α = .95), fear (α = .9), sadness (α = .9), happiness (α = .95), and neutral (α = .95).

8.3.3. Emotion Induction Checks

Separate ANOVA's were run to examine responses across six categories of emotion induction checking items between conditions (7), and for measures of both private body consciousness and interoceptive awareness. As expected, there were no significant differences between conditions with regard to scores on measures of interoceptive awareness [$F(6,504) = 1.791$, $p = .099$, $\eta p2 = .021$], or private body consciousness [$F(6,504) = 1.471$, $p = .186$, $\eta p2 = .017$]. Analysis of emotion induction checks were performed within each condition, with Bonferroni adjusted post hoc tests suggesting emotion induction had been successful in at least one condition. However, the overall pattern of results suggests induction success was limited in most conditions.

Disgust induction was the only condition where the target emotion fully met expectations. Ratings of disgust were higher than those of any other emotion ratings within the disgust condition, as well as higher than ratings of disgust across any other condition. Ratings of fear were higher in the fear condition than in any other condition, although the ratings were not significantly different from those in the anger condition (i.e., angry music may have also induced some fear), and participants in the fear condition actually gave higher ratings for measures of anger than for fear. Fearful music did induce fear to a greater extent than most of the other music, but fear inducing music seems to have induced anger to a greater extent than it induced fear. Indeed, ratings of anger in the fear condition were comparable to ratings of anger in the anger condition, and neither were significantly different from ratings of anger in the disgust condition. Thus, although ratings of anger were higher than those of any other emotion within the anger induction condition, both fear and disgust inductions elicited anger to a similar extent.

Ratings of happiness were higher when inducing happiness than when inducing anger, disgust, or fear. However, ratings of happiness in this induction condition were not significantly different from happiness ratings in the sadness condition, and happiness ratings were (non-significantly) higher in both control conditions than the happiness condition. Neutral emotion ratings were higher in both control conditions than neutral ratings in the anger and fear conditions, as well as higher than disgust ratings in the second control condition, although neutral emotion ratings in the happiness and sadness conditions were comparable to those in control conditions. Ratings for items relating to sadness showed no differences between conditions, and both happiness and neutral emotions received higher ratings within the sadness condition. Descriptive statistics for emotion ratings across conditions are shown in Table 8.2, and illustrated in Figure 8.1, with test values and mean differences provided in Table 8.3. Correlations between emotion ratings are provided in Table 8.4.

Table 8.2. Means and Standard Deviations of Emotion Family Ratings by Emotion Induction Condition

	Anger Induction	Disgust Induction	Fear Induction	Sadness Induction	Happiness Induction	Control	Control 2	All Groups Average
Anger Ratings	**3.7 (1.71)**	*3.11 (1.57)*	*3.71 (1.98)*	2.5 (1.8)	2.19 (1.51)	2.19 (1.55)	2.04 (1.38)	2.72 (1.83)
Disgust Ratings	2.72 (1.83)	**4.44 (2.14)**	3.05 (1.77)	2 (1.69)	1.85 (1.39)	2.09 (1.62)	2.24 (1.57)	2.56 (1.88)
Fear Ratings	*2.76 (2)*	1.87 (1.24)	**3.17 (2.04)**	2.16 (1.72)	1.79 (1.09)	2.03 (1.44)	1.77 (1.17)	2.19 (1.62)
Sadness Ratings	2.49 (1.54)	2.05 (1.59)	2.54 (1.85)	**2.48 (1.88)**	1.99 (1.44)	2.44 (1.78)	1.89 (1.28)	2.26 (1.64)
Happy Ratings	2.32 (1.42)	2.75 (1.57)	2.24 (1.49)	*3.03 (1.79)*	**3.35 (1.58)**	*3.42 (1.78)*	*3.71 (1.75)*	3.03 (1.71)
Neutral Ratings	2.73 (1.91)	*3.39 (1.93)*	2.49 (1.67)	*4.31 (2.09)*	*4.62 (1.73)*	**4.28 (1.95)**	**4.51 (1.91)**	3.82 (2.05)

Emphasis denotes comparable emotion ratings across conditions (i.e., target specificity - read across). Underline highlights where alternative inductions have been rated as higher within the target condition (i.e., induction efficacy - read down). **Bold** indicates intersections of induction conditions and target emotions (read across and down).

Data shown in Table 8.2. above maps directly to data shown in Figure 8.1. below.

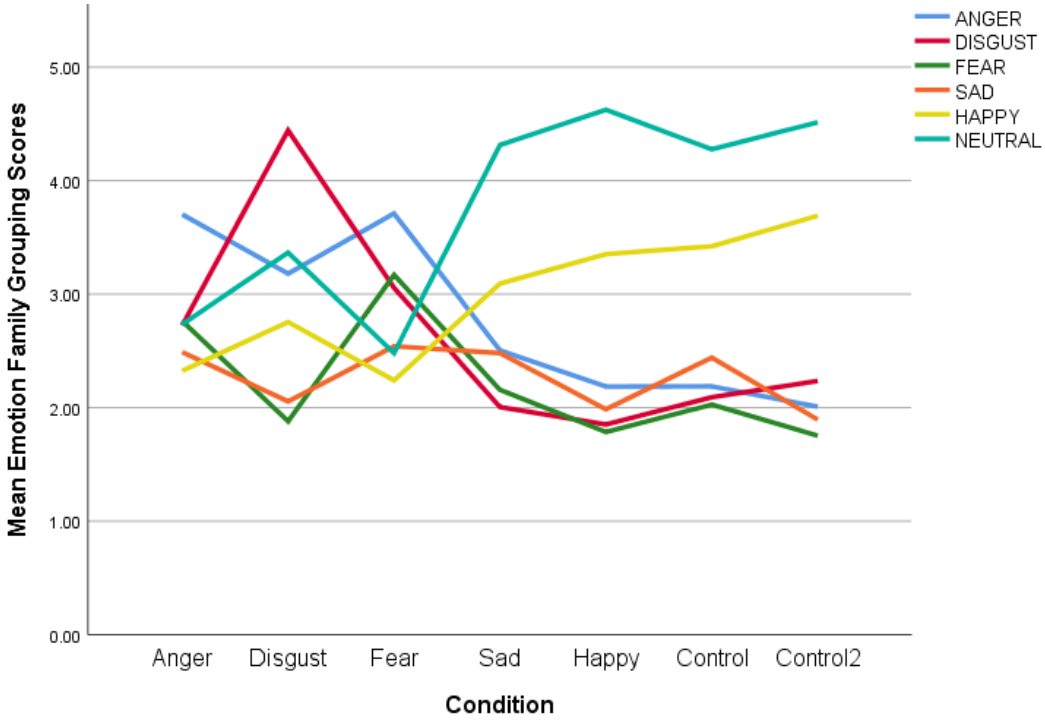


Figure 8.1. Mean Emotion Family Ratings across Emotion Induction Conditions

Table 8.3. Test Results and Mean Differences for Emotion Family Ratings across Emotion Induction Conditions

ALL *F*(6,500)	Anger Induction	Disgust Induction	Fear Induction	Sadness Induction	Happiness Induction	Control	Control 2
Anger Ratings	***F* = 14.799**** ***ηp2* = .151**	*n.s.*	*n.s.*	1.197**	1.517**	1.514**	1.66**
Disgust Ratings	1.718**	***F* = 17.951**** ***ηp2* = .177**	1.388**	2.438**	2.59**	2.351**	2.207**
Fear Ratings	*n.s.*	1.304**	***F* = 8.212**** ***ηp2* = .09**	1.103**	1.384**	1.143**	1.402**
Sadness Ratings	*n.s.*	*n.s.*	*n.s.*	***F* = 2.155*** ***ηp2* = .025**	*n.s.*	*n.s.*	*n.s.*
Happy Ratings	1.028**	*n.s.*	1.112**	*n.s.*	***F* = 8.668**** ***ηp2* = .093**	*n.s.*	*n.s.*

Significant at $p < .05$*, ** $p < .01$. Bonferroni adjusted for mean differences.

Mean differences relate to Table 8.1. read across (i.e., target specificity).

Table 8.4. Correlations of Emotion Family Ratings

Ratings	Anger	Disgust	Fear	Sadness	Happiness	Neutral
Anger	**-**	.568	.573	.497	-.327	-.510
Disgust		**-**	.332	.280	-.214	-.289
Fear			**-**	.716	-.224	-.339
Sadness				-	-.278	-.370
Happiness					**-**	.706
Neutral						**-**

All correlations were significant, $p < .01$ (2-tailed).

8.3.4. Examining Experimenter Effects

Another ANOVA was run to examine Landy and Goodwin's (2015) contention that participants may experience negative affect towards the experimenter as a result of the induction methodology. This is suggested to act as a confound in studies on moral judgement which induce negative affect, such that significant results for experimenter effects would need to be accounted for in subsequent analyses. The results of this analysis showed differences between groups [$F(6,504) = 5.462$, $p < .001$, $\eta p2 = .061$], although planned post-hoc analyses showed a limited pattern of variation. There was a marginal difference between participants in the happiness and anger conditions, with those in the latter category feeling more negatively towards the experimenter ($MD = -.505$, $p = .038$). There were also differences between those that experienced fear induction and those in the sad (MD = -.704, $p = .001$), happy ($MD = -.781$, $p < .001$), and control induction conditions ($MD = -.637$ / $-.538$, $p = .002$ / $.021$), with those in the fear condition feeling more negatively. Ratings towards the experimenter in the disgust condition tended towards those in the anger and fear conditions, although there were no statistically significant differences between disgust induction and any other conditions.

In broader terms, those in negatively valenced, high arousal induction conditions (i.e., anger, disgust, fear) tended to feel more negatively towards the experimenter than participants in other induction conditions (i.e., sadness, happiness, neutral). When asked, those that felt less favourably ($n = 113$) attributed this rating to the sound (55%) more often than to the scenarios (35%), or to something else (10%) - although 'something else' was often used to mean both the sound and the scenarios. However, a similar number of participants ($n = 112$) felt more positively towards the experimenter as a result of taking part in the study (the reasons for which were not captured), and the majority of participants reported no change in this regard ($n = 285$).

Further exploration of this measure showed it had no significant influence on ratings of wrongness for any type of scenario, nor on ratings of emotion checking items when re-running the analysis with experimenter effect as a covariate, nor on any of the subsequent planned analyses when entered as a covariate. Correlations with emotion checking items were generally reflective of the results of the ANOVA, although there were notable differences in the pattern of emotion associations. The experiencing of high valence emotion (i.e., anger, disgust, fear) was associated with greater negativity towards the experimenter, and happy and neutral emotions were associated with greater positivity towards the experimenter. However, whereas experimenter effects in the disgust condition were non-significant, the actual experiencing of disgust (i.e., rated disgust) was correlated with experimenter negativity to a similar extent as experiencing anger. Furthermore, whilst those in the fear condition reported greater negativity towards the experimenter, the correlation of experiencing fear with experimenter negativity was under half the size of that reported for experienced anger or disgust. Experimenter effects also showed some correlation with measures of private body consciousness and interoceptive awareness, whereby scoring higher on these measures was associated with more positive ratings towards the experimenter. Statistics for experimenter effects are provided in Tables 8.5 and 8.6.

Table 8.5. Means and Standard Deviations of Experimenter Effects by Condition

Overall	Anger	Disgust	Fear	Sadness	Happiness	Control	Control 2
4.07 (.994)	3.87 (.991)	3.92 (1.15)	3.59 (.844)	4.29 (1.07)	4.37 (.951)	4.23 (.956)	4.13 (.823)

Table 8.6. Correlations of Experimenter Effects with Emotion Ratings, PBC, MAIA

Anger	Disgust	Fear	Sadness	Happy	Neutral	PBC	MAIA
-.309**	-.311**	-.138**	-.086 *n.s.*	.225**	.310**	.101*	.141**

Correlations significant at $p < .05$* or $p < .01$** (2-tailed)

8.3.5. Testing for amplification effects

A mixed ANOVA was run to explore whether there was any effect on wrongness within judgements (8) between conditions (7). The result of primary interest to the hypotheses was non-significant. There was no main effect found between conditions [$F(1,6) = .256$, $p = .957$, $\eta p2 = .003$], nor was there an interaction effect with judgement type found [$F(24.607,2066.977) = .933$, $p = .558$, $\eta p2 = .011$]. Participants listening to music (of any kind) did not judge the wrongness of the scenarios significantly more or less severely than those in the control group(s) who did not listen to music. There was no evidence in favour of any direct amplification effect present in the planned analyses.

There was a main effect of judgement category [$F(4.101,2066.977) = 2023$, $p < .001$, $\eta p2 = .801$]. Examining this effect using *t*-tests showed, as expected, there were differences (all $p < .001$ following Bonferroni adjustments) between all scenarios containing moral content and those that were merely counter-normative ($M = 5.96$, $SD = 1.08$), with these being rated as less wrong than moral violations (min-max mean difference = .497-1.51), but more wrong than non-moral scenarios ($M = 1.62$, $SD = 1.26$). Unexpectedly, scenarios depicting violations of sanctity ($M = 7.47$, $SD = 1.29$) were rated as more wrong than any other type of violation, including those of harm ($M = 7.11$, $SD = 1.14$) and liberty ($M = 7.11$, $SD = 1.11$), which scored similarly. Both of these were rated as more wrong than violations of loyalty ($M = 6.76$, $SD = 1.07$) and authority ($M = 6.88$, $SD = 1.13$), which also scored similarly, whilst fairness violations were rated as less wrong than all other types of moral violation ($M = 6.46$, $SD = 1.16$). Significant mean differences between different types of scenarios are shown in Table 8.7, with overall results summarised in Figure 8.2. Correlations between wrongness ratings over different types of moral scenario ranged between .343 and .675, with associations between sanctity and other scenarios appearing slightly weaker than associations between any other scenario type pairing.

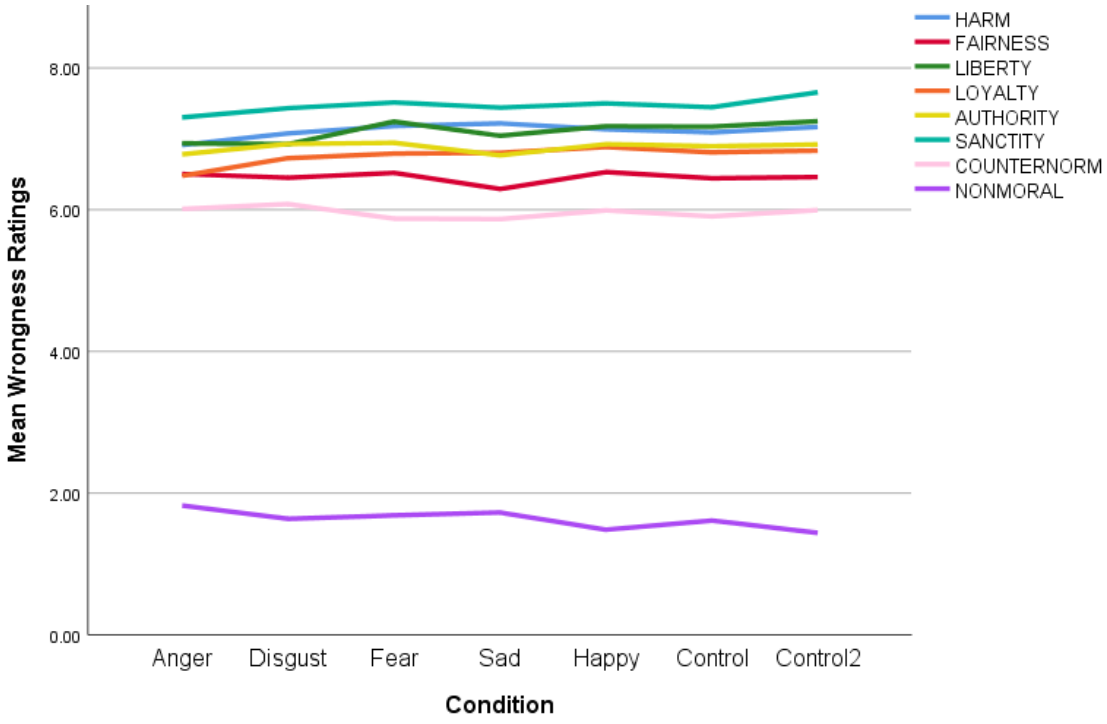


Figure 8.2. Mean wrongness ratings for judgement type across conditions.

Table 8.7. Mean differences between Judgements across all conditions combined.

	HARM	FAIR	LIB	LOY	AUTH	SANCT	NORM	NON-M
HARM		-.654	n.s.	-.351	-.232	.359	-1.151	-5.482
FAIR	.654		.649	.304	.423	1.103	-.497	-4.827
LIB	n.s.	-.649		-.346	-.227	.364	-1.146	-5.477
LOYAL	.351	-.304	.346		n.s.	.709	-.801	-5.131
AUTH	.232	-.423	.227	n.s.		.590	-.920	-5.250
SANCT	-.359	-1.013	-.364	-.709	-.590		-1.510	-5.840
NORM	1.151	.497	1.146	.801	.920	1.510		-4.330
NON-M	5.482	4.827	5.477	5.131	5.250	5.840	4.330	

Mean difference values all significant at $p < .001$ following Bonferroni adjustments.

8.3.6. Examining potential moderators of the effect

Any potential moderating effects of interoceptive awareness and private body consciousness were investigated by assigning participants to high or low scoring groups (drawing on Schnall et al., 2008), based on the median sample score for each measure, and re-running the analyses with these high and low scoring categories as a second between subjects factor (i.e., 2x7x8, where 2 is high/low for PBC or MAIA scores).

The 3-way interaction effect when including private body consciousness was non-significant. There was also no interaction effect between condition and private body consciousness category [$F(1,6) = .469$, $p = .832$, $\eta p2 = .006$]. The trend for participants with higher private body consciousness scores to provide higher wrongness ratings failed to achieve statistical significance [$F(1,497) = 2.963$, $p = .086$, $\eta p2 = .006$], and wrongness ratings in the fear induction condition showed the opposite pattern, with those categorised as having low-PBC providing higher wrongness ratings that those in the high-PBC group (Figure 8.3., Table 8.8). However, an interaction effect with judgement type was apparent across ratings of private body consciousness [$F(4.150,2062.592) = 3.092$, $p = .014$, $\eta p2 = .006$]. Subsequent independent *t*-tests to examine this interaction effect showed that private body consciousness category had an effect on wrongness ratings for judgements of liberty, loyalty, authority, and (in the opposite direction for) non-moral judgements, but this effect was non-significant over ratings for judgements of harm, fairness, sanctity, and counter-normative violations (see Figure 8.4., Table 8.9).

Table 8.8. Mean wrongness ratings across condition by PBC category

	Anger	Disgust	Fear	Sadness	Happiness	Control	Control 2
LOW PBC	6.078	6.062	6.293	6.028	6.126	6.088	6.121
HIGH PBC	6.122	6.286	6.147	6.232	6.261	6.274	6.314
t *p*	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*

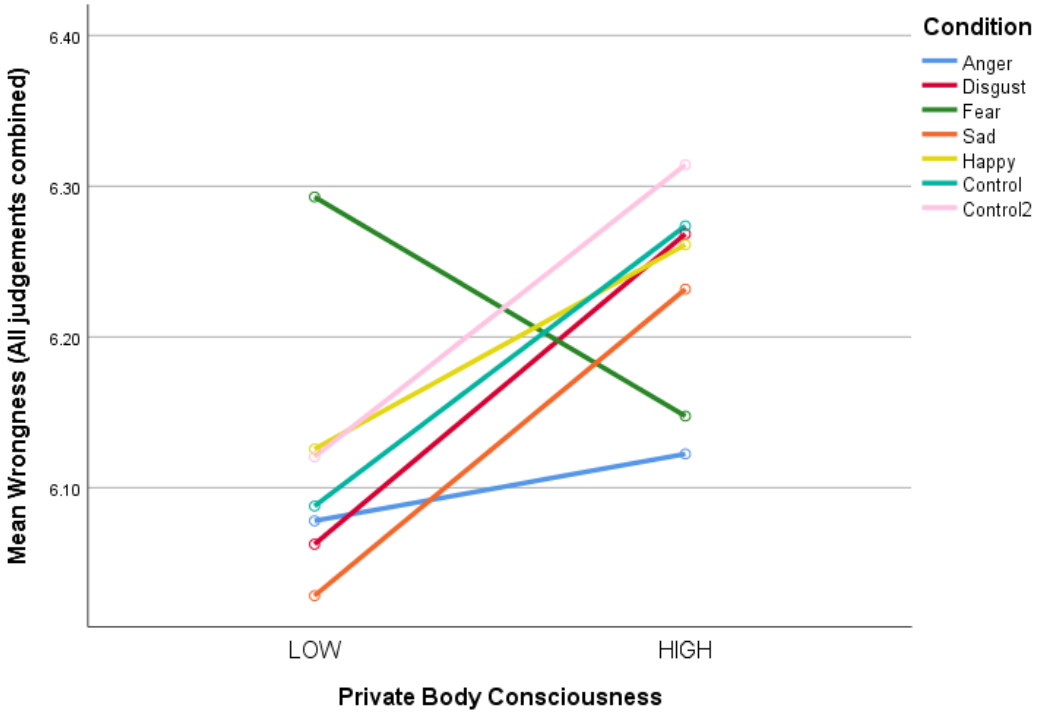


Figure 8.3. Mean wrongness ratings across conditions by PBC category.

Table 8.9. Mean wrongness ratings across judgements by PBC category

	HARM	FAIR	LIB	LOY	AUTH	SANCT	NORM	NON-M
PBC LOW	7.046	6.382	6.993	6.680	6.763	7.418	5.916	1.711
PBC HIGH	7.194	6.529	7.256	6.843	7.005	7.533	5.997	1.492
t *p*	*n.s.*	*n.s.*	-2.870 .004	-2.023 .044	-2.622 .009	*n.s.*	*n.s.*	2.202 .028

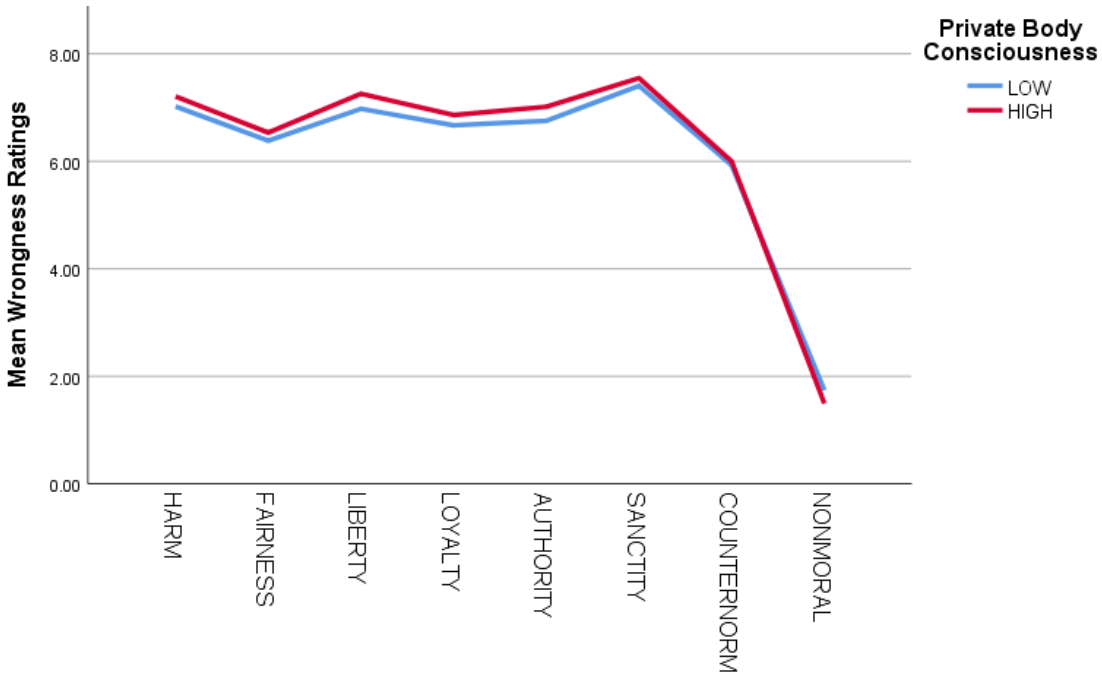


Figure 8.4. Mean wrongness ratings across judgements by PBC category.

The 3-way interaction effect when including interoceptive awareness was also non-significant. Further inspection showed, once again, there was no interaction effect between condition and interoceptive awareness categories [$F(1,6) = .568$, $p = .756$, $\eta p2 = .007$]. However, there was a main effect for interoceptive awareness [$F(1,497) = 9.798$, $p = 0.02$, $\eta p2 = .019$] whereby those scoring 'high' on this measure tended to rate scenarios as being more wrong overall than those scoring 'low' ($MD = .213$), although independent *t*-tests by condition showed the difference in wrongness ratings between high- and low-MAIA groups was particularly pronounced within the anger condition (see Figure 8.5., Table 8.10). The interaction effect between interoceptive awareness and judgement was short of the significance threshold [$F(4.1,2037.892) = 2.297$, $p = .055$, $\eta p2 = .005$]. Examination using independent *t*-tests showed scenarios with most forms of potentially moralized content (i.e., including counter-normative scenarios) were rated as more wrong by participants with higher-than-median interoceptive awareness, although scenarios in the sanctity and non-moral categories were rated as similarly wrong by both above- and below-median groups (see Figure 8.6., Table 8.11).

Table 8.10. Mean wrongness ratings across condition by MAIA category

	Anger	Disgust	Fear	Sadness	Happiness	Control	Control 2
LOW MAIA	5.899	6.034	6.178	6.040	6.120	6.104	6.128
HIGH MAIA	6.403	6.256	6.252	6.223	6.286	6.281	6.291
t *p*	-2.689 .009	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*

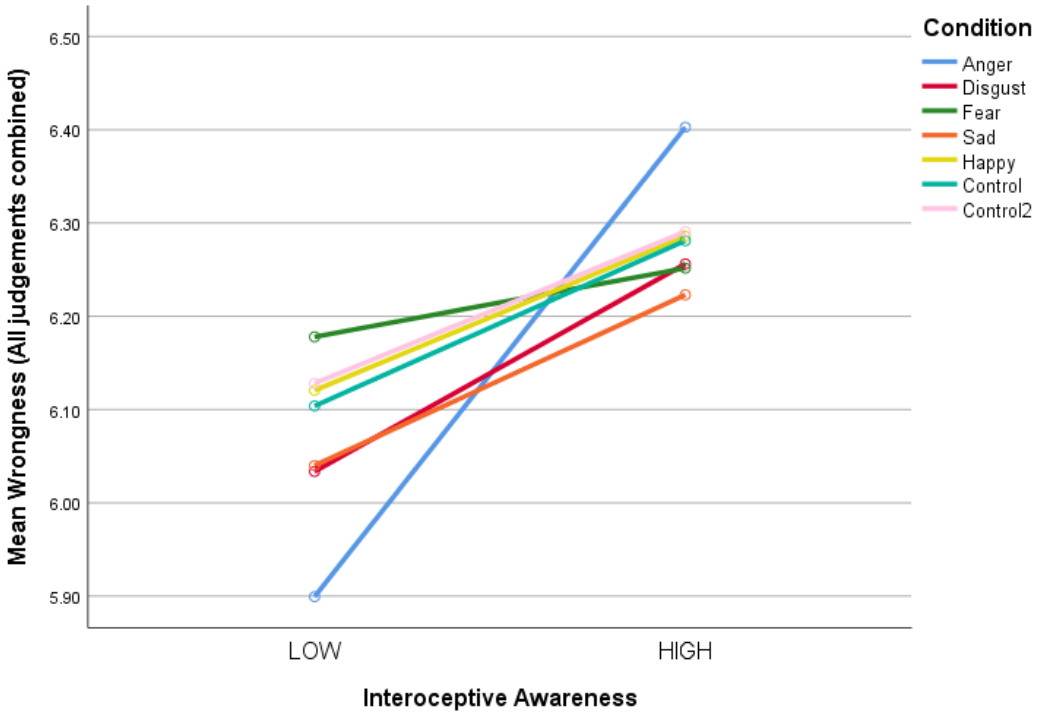


Figure 8.5. Mean wrongness ratings across conditions by MAIA category.

Table 8.11. Mean wrongness ratings across judgements by MAIA category

	HARM	FAIR	LIB	LOY	AUTH	SANCT	NORM	NON-M
LOW MAIA	7.009	6.313	6.984	6.617	6.735	7.437	5.831	1.649
HIGH MAIA	7.242	6.617	7.243	6.911	7.044	7.519	6.100	1.599
t *p*	-2.635 .009	-2.935 .003	-2.697 .007	-3.159 .002	-3.083 .002	*n.s.*	-2.619 .009	*n.s.*

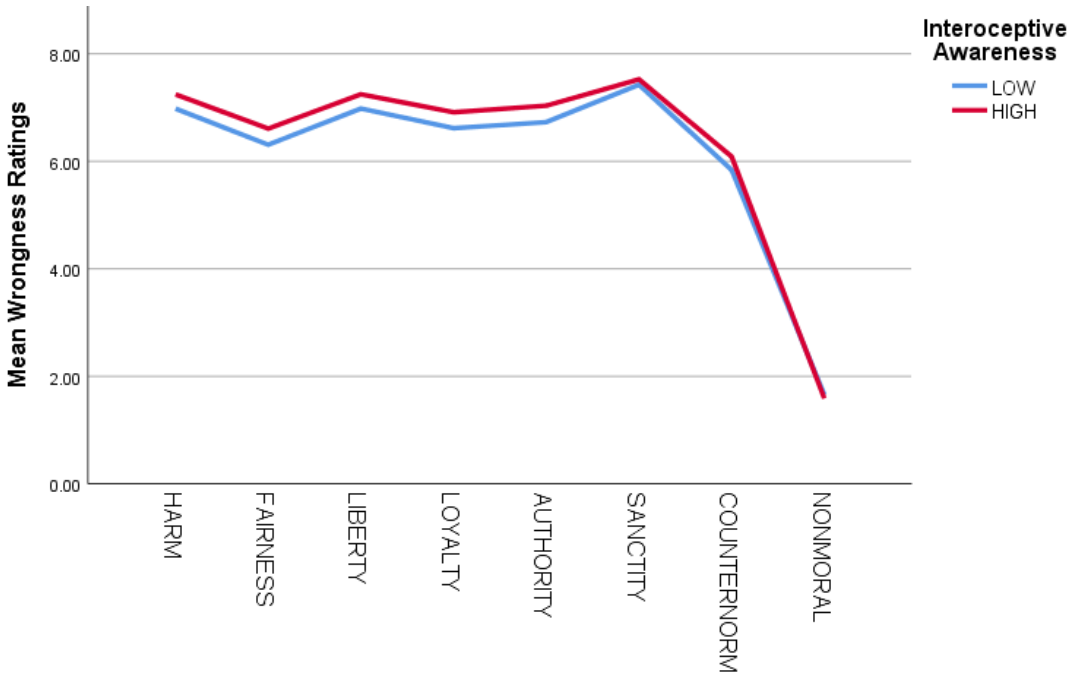


Figure 8.6. Mean wrongness ratings across judgements by MAIA category.

Re-running these analyses as ANCOVAs, using scales scores rather than median-based groupings for both PBC and MAIA, suggested that results may have been partially influenced by, but not dependent on, dichotomising from a continuous variable. ANCOVA's confirmed both measures had significant effects on wrongness ratings when entered as covariates [PBC: $F(1,503) = 7.27$, $p = .007$, $\eta p2 = .014$]; MAIA: $F(1,503) = 6.851$, $p = .009$, $\eta p2 = .013$], and judgement type continued to interact with private body consciousness as a covariate [$F(4.169,2097.222) = 6.950$, $p < .001$, $\eta p2 = .014$]. Parameter estimates showed the effect for PBC appeared present across most scenario types ($p < .05$) with the exception of sanctity ($p = .079$) and counter-normative ($p = .189$) scenarios. In contrast, the effect of MAIA became non-significant over harm and loyalty scenarios, and remained non-significant over sanctity and non-moral scenarios. However, scores on both measures returned similar correlations with overall ratings of wrongness - interoceptive awareness ($r = .121$, $p = .006$), private body consciousness ($r = .123$, $p = .005$) - suggesting both may exert similar overall influence in this regard.

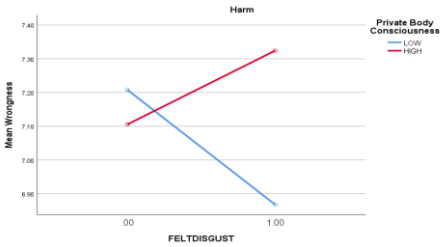
8.3.7. Exploratory Analyses

For the purposes of amplification effects, it seems what matters most is the actual experiencing of an emotion - "...the more clearly participants are experiencing disgust, the more directly this feeling is taken as input to moral judgments" (Schnall et al., 2008, p. 1105). As the results of emotion induction were less than ideal, data were re-analysed to focus on emotion elicitation responses. Participants were allocated to one of two conditions based on median responses to each of the six emotion measures, such that an emotion (e.g., anger) was categorised as being felt (above median) or not felt (median-and-below) regardless of induction condition. A series of mixed 2x2x8 ANOVA's, 'felt emotion' (felt/not felt) x PBC or MAIA (high/low) x judgement type (8), were run for each of the six emotion measures. Any two-way interactions between judgement type and PBC or MAIA are already reported above, whereas any two-way interactions between judgement type and felt emotion are addressed after the three-way analyses.

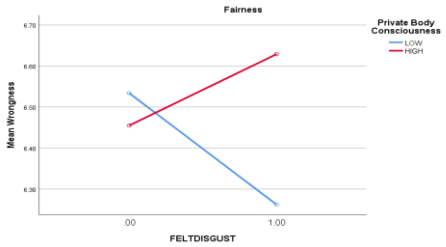
When running the analyses with PBC groupings, a between groups effect was found for felt emotion when analysing both happiness [$F(1,504) = 6.881$, $p = .009$, $\eta p2 = .013$] and sadness [$F(1,504) = 4.387$, $p = .037$, $\eta p2 = .009$]. Those categorised as feeling happy ($M = 6.079$), or sad ($M = 6.103$), provided lower overall wrongness ratings (within their respective analyses) than those categorised as not feeling happy ($M = 6.256$), or not feeling sad ($M = 6.243$). There remained a trend toward a between groups effect for PBC category in each iteration of the analysis, with high-PBC participants giving higher wrongness scores overall, although this only cleared the significance threshold when analysing happiness as the felt emotion [$F(1,504) = 4.366$, $p = .037$, $\eta p2 = .009$]. In this analysis, participants categorised as high-PBC provided higher wrongness ratings ($M = 6.238$) than participants classed as low-PBC ($M = 6.097$).

The results further showed a two-way interaction between private body consciousness and felt disgust [$F(1,503) = 6.482$, $p = .011$, $\eta p2 = .013$]. High-PBC participants who felt disgust gave higher wrongness ratings ($M = 6.324$) than low-PBC participants who felt disgust ($M = 6.024$), whereas high-PBC participants who did not feel disgust gave similar wrongness ratings ($M = 6.173$) to low-PBC participants who did not feel disgust ($M = 6.215$). There was also a 3-way interaction effect [$F(4.256,2140.525) = 2.394$, $p = .045$, $\eta p2 = .005$]. Examination using *t*-tests showed the PBC x felt disgust interaction effect was as described for all scenario types, and running in the opposite direction for non-moral scenarios (i.e., lower wrongness ratings), although the difference in ratings between high- and low-PBC participants who felt disgust did not meet the significance threshold over non-moral scenarios ($p = .08$) or those relating to sanctity ($p = .07$).

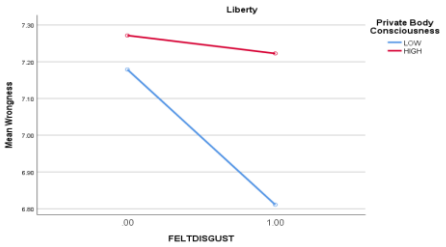
Figure 8.7 shows the felt disgust x PBC x judgement type interaction, as well as other effects present in these analyses. Plots for liberty, loyalty, authority, and non-moral scenarios are illustrative of interactions between PBC and judgement type (from Table 8.9), with plots for liberty and non-moral scenarios also reflective of felt disgust x judgement type interactions (reported following all three-way analyses). That the felt disgust x PBC interaction was short of significant over sanctity and non-moral scenarios might be explained via the main effect for judgement type (sanctity) and the felt disgust x judgement interaction (non-moral, detailed following all three-way analysis). Plots for the felt disgust x PBC interaction are shown across Figures 8.8, 8.9, and 8.10.



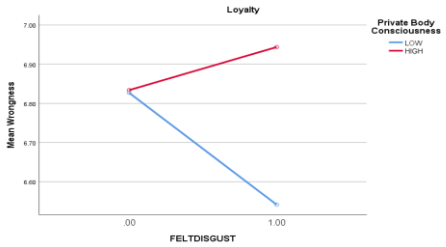
HARM



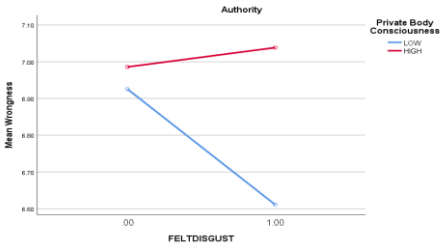
FAIRNESS



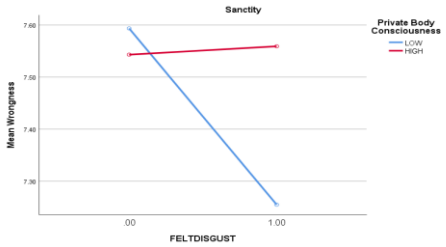
LIBERTY



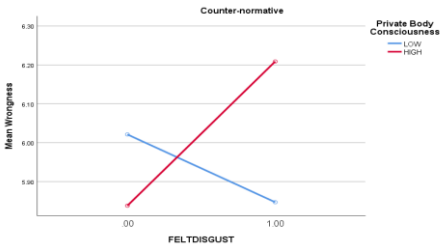
LOYALTY



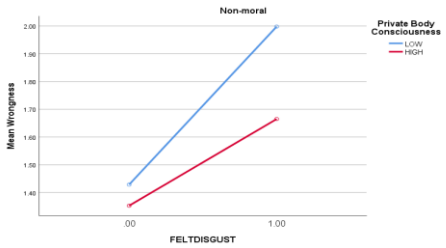
AUTHORITY



SANCTITY



COUNTERNORMATIVE



NON-MORAL

Figure 8.7. Mean wrongness (Y-axis) by felt disgust (left-side-low, right-side-high) and private body consciousness (blue-low, red-high) for each scenario type.

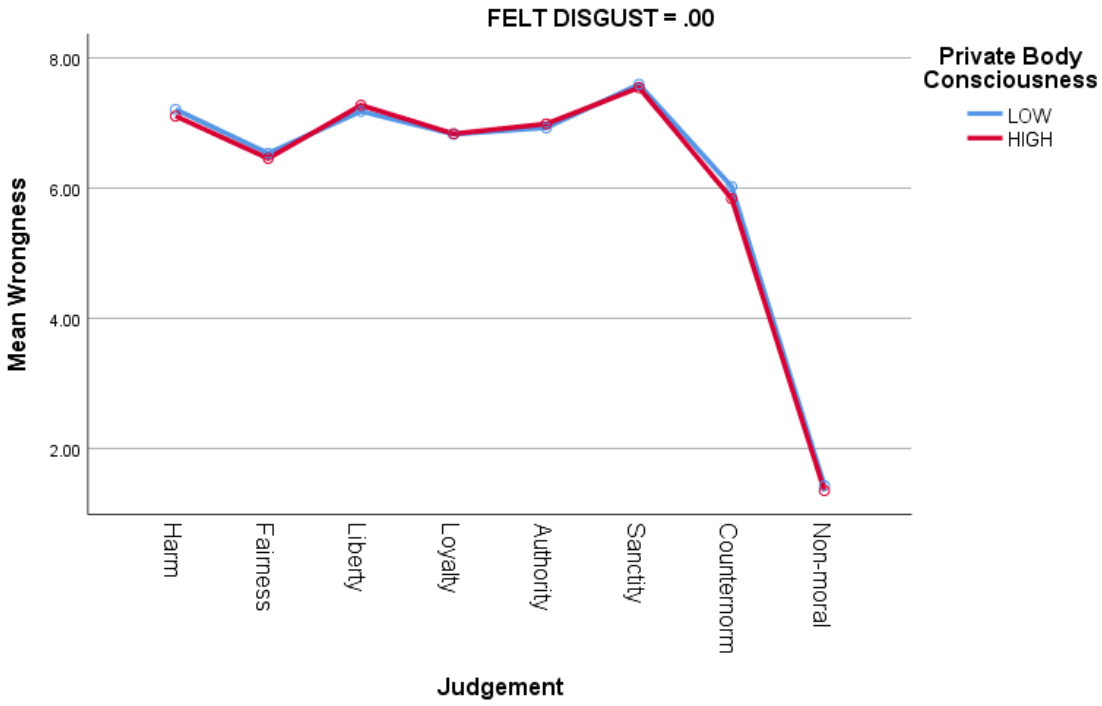


Figure 8.8. Mean wrongness ratings across judgements by PBC (not felt disgust)

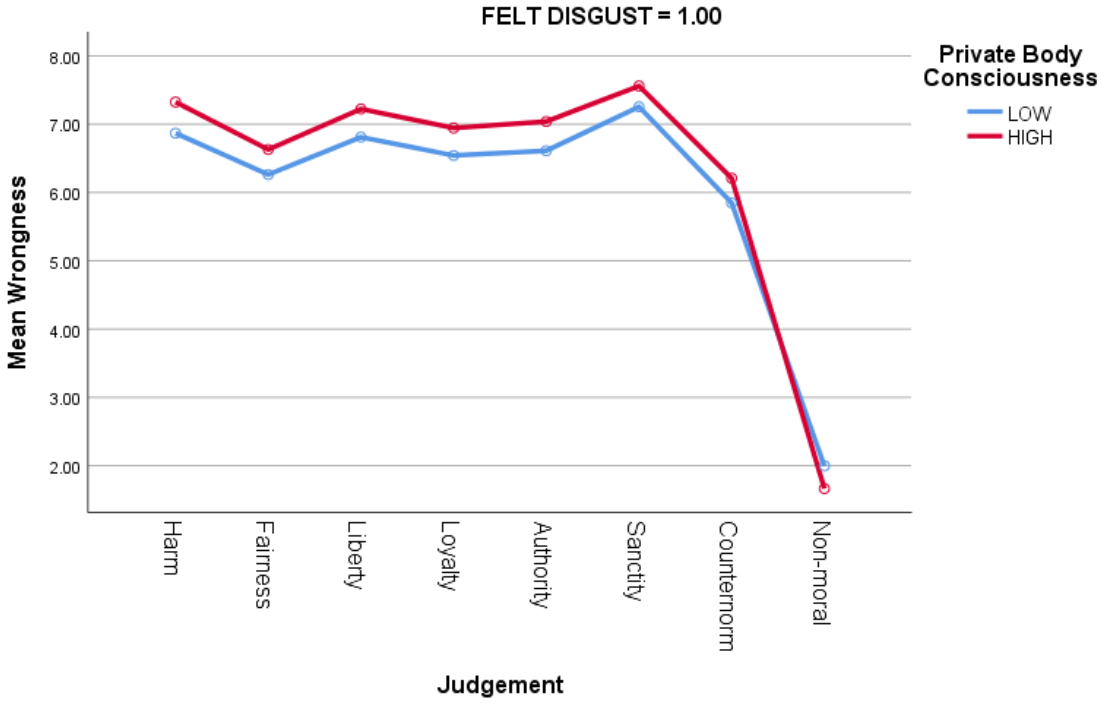


Figure 8.9. Mean wrongness ratings across judgements by PBC (felt disgust)

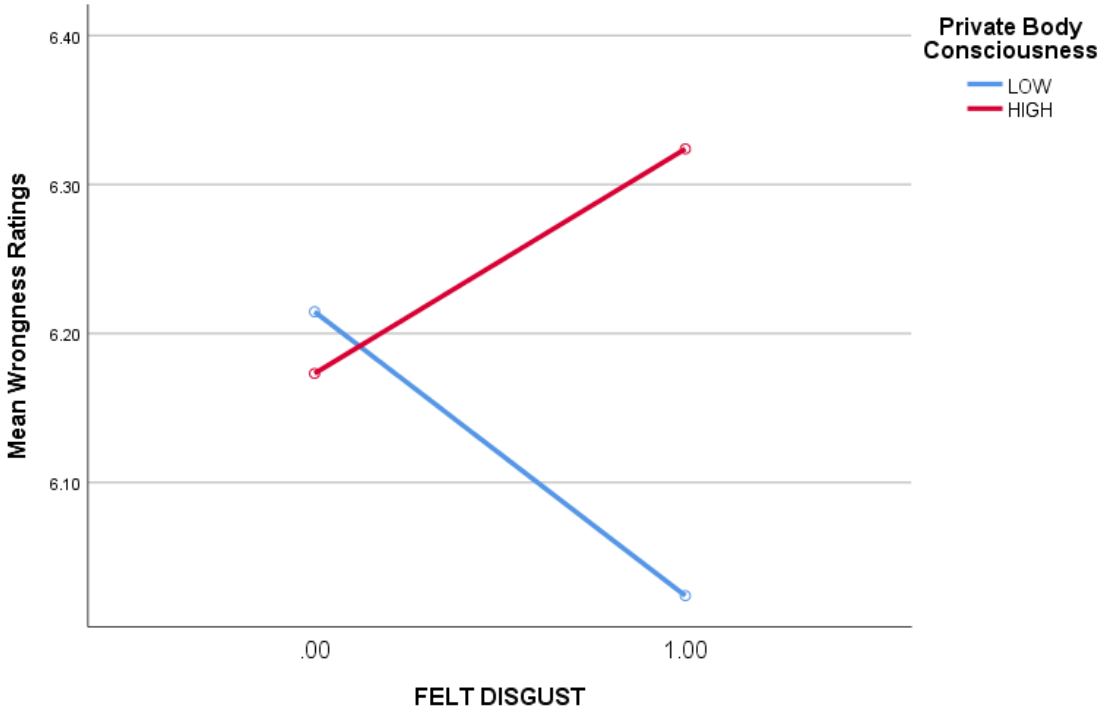


Figure 8.10. Mean wrongness ratings across PBC and felt disgust categories

When analysing with MAIA, the previously detailed main effect was present in each iteration, with high-MAIA participants providing more extreme wrongness ratings than low-MAIA participants. Analysing with MAIA instead of PBC similarly found a main effect between groups for happiness [$F(1,504) = 9.92$, $p = .002$, $\eta p2 = .019$], with those in the felt group providing lower wrongness ratings than those in the not felt group. However, the main effect found when analysing felt sadness with PBC fell short of the significance threshold [$F(1,504) = 3.817$, $p = .051$, $\eta p2 = .008$]. There were also no interactions between interoceptive awareness categories and felt emotion groups, although the felt disgust group was closest to the significance threshold [$F(1,503) = 1.767$, $p = .184$, $\eta p2 = .004$]. Examination of interoceptive awareness by felt disgust showed the same trend as for the felt disgust x private body consciousness interaction effect over the majority of scenarios, although this trend was notably absent over sanctity and non-moral scenarios (similarly to interoceptive awareness by judgement type). In further contrast to including PBC, the only significant three-way interaction (unpacked across Figures 8.11 - 8.13) was in the happiness analysis [$F(4.226,2129.839) = 2.404$, $p = .044$, $\eta p2 = .005$].

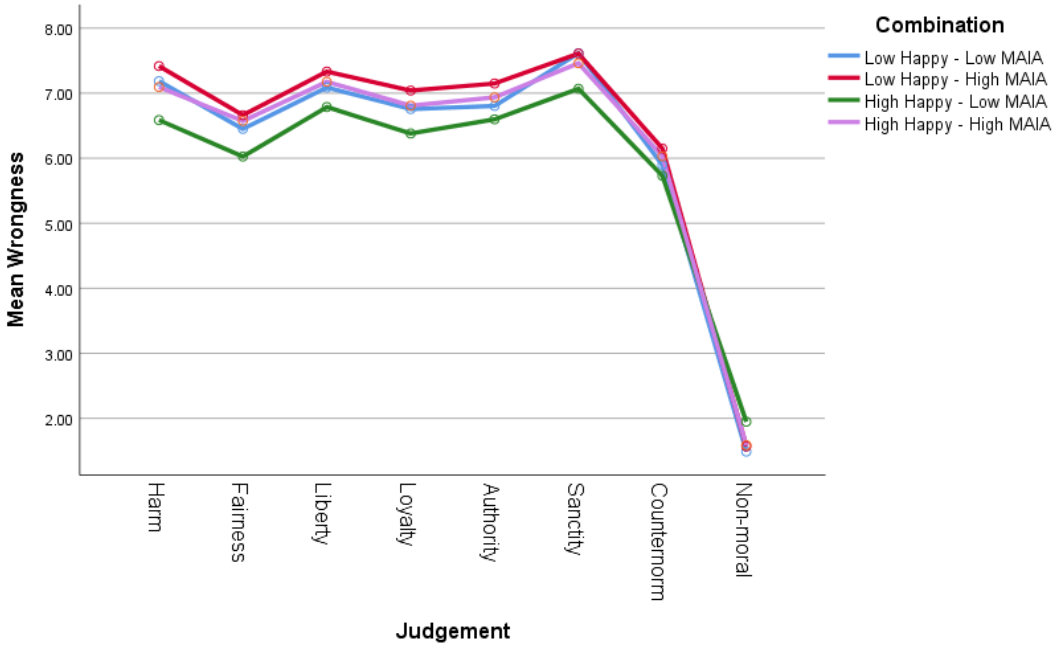


Figure 8.11. Mean wrongness across judgement type by felt happy x MAIA

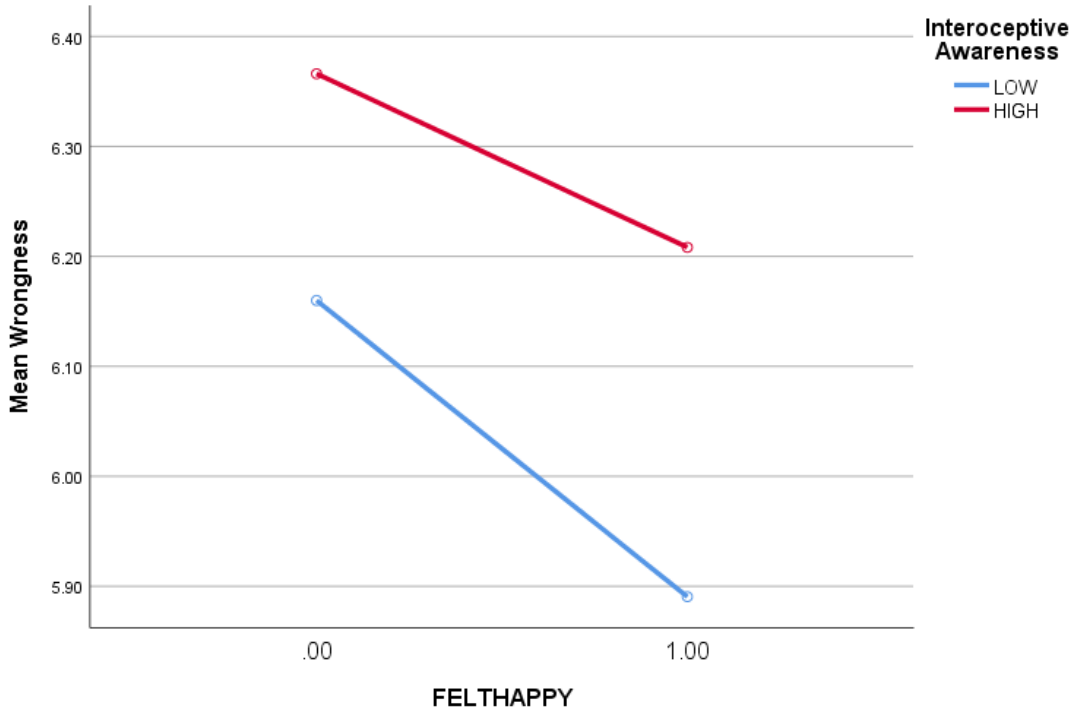
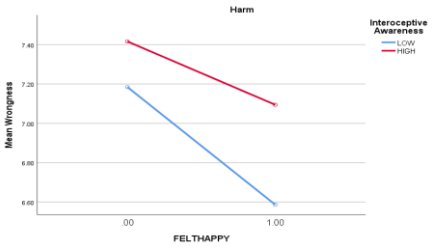


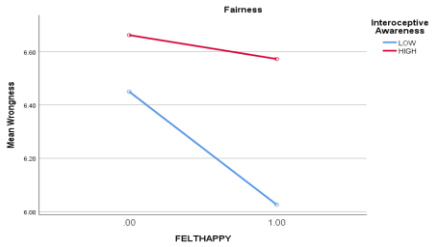
Figure 8.12. Mean wrongness ratings across MAIA and felt happy categories

The three-way interaction effect was notably less uniform than the three-way interaction effect found for felt disgust, as the interoceptive awareness x felt happy interaction effect was non-significant overall [$F(1,504) = .676$, $p = .411$, $\eta p2 = .001$]. Further examination via a 4x8 ANOVA and independent *t*-tests showed that wrongness ratings for low-MAIA participants who felt happy ($M = 5.891$) were significantly lower than those of high-MAIA participants who felt happy ($MD = -.318$, $p = .012$). However, wrongness ratings for happy-low-MAIA participants were also lower than ratings from both low-MAIA participants who did not feel happy ($MD = -.269$, $p = .037$), and high-MAIA participants who did not feel happy ($MD = -.476$, $p < .001$).

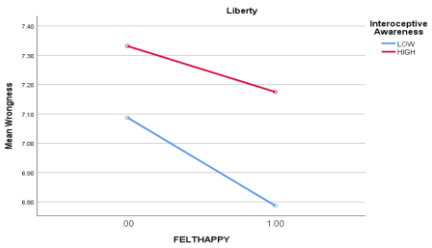
Figure 8.13 shows plots for felt happy x MAIA x judgement type, with significant differences in wrongness ratings found between the following plot points. Above-median happy participants with high-MAIA scores (red-right) rated all types of scenarios as more wrong than above median happy participants with low-MAIA scores (blue-right), although this difference was short of the significance threshold for sanctity and counter-normative scenarios. Participants with lower happiness ratings and high-MAIA scores (red-left) rated liberty, loyalty, authority, and counter-normative scenarios as more wrong than those with lower happiness ratings and low-MAIA scores (blue-left). The only significant difference in wrongness ratings between high-MAIA participants (red slopes) by happiness was for harm scenarios, whereas differences between low-MAIA participants (blue slopes) were found over harm, fairness, loyalty, sanctity and non-moral scenarios. For each type of scenario, wrongness ratings given by low-MAIA participants who did not feel happy (blue-left) were comparable with ratings given by happy-high-MAIA participants (red-right), whereas the largest differences in ratings were between happy-low-MAIA participants (blue-right) and high-MAIA participants who did not feel happy (red-left). The difference between high- and low-MAIA categories (red-blue) was almost always significant (from Table 8.11), whereas the difference between low- and high-felt happiness (left-right) was only significant over scenarios of harm, loyalty, and sanctity. Comparisons of these relationships are detailed in Table 8.12.



HARM

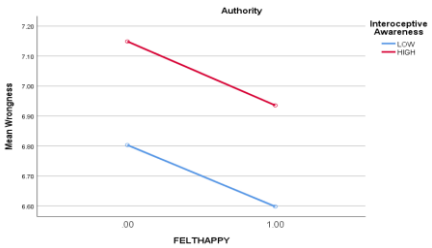


FAIRNESS

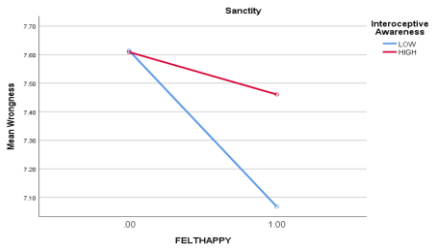


LIBERTY

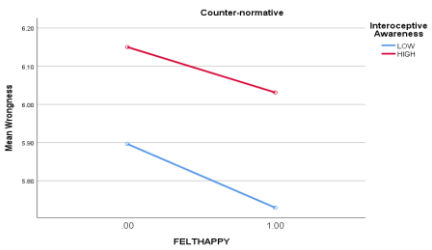
LOYALTY



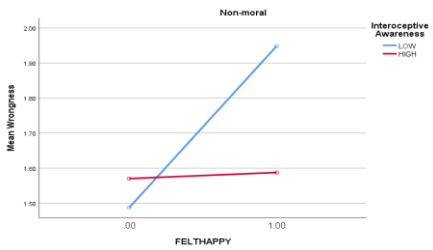
AUTHORITY



SANCTITY



COUNTER-NORMATIVE



NON-MORAL

Figure 8.13. Mean wrongness (Y-axis) by felt happy (left-side-low, right-side-high) and interoceptive awareness (blue-low, red-high) for each scenario type (below).

Table 8.12. Mean differences over felt happy x MAIA by judgement type

	HARM	FAIRNESS	LIBERTY	LOYALTY	AUTHORITY	SANCTITY	COUNTER-NORM	NON-MORAL	COMBINED
NH **FH**	.384 < .001	.179 = .085	.165 = .111	.231 = .015	.141 = .163	.305 = .011	.087 = .370	-.208 = .065	.214* = .002
HM **LM**	.265 = .009	.300 = .003	.264 = .007	.297 = .002	.306 = .002	.101 = .376	.249 =.009	-.066 = .558	.262* < .001
FH HM **FH LM**	.508 = .005	.546 = .003	.388 = .027	.428 = .006	.338 = .040	.393 = .064	.301 =.056	-.361 =.048	.318** = .012
NH HM **NH LM**	.232 = .060	.212 = .101	.244 = .035	.284 = .018	.346 = .008	-.005 = .971	.254 = .042	.083 = .573	.206** = .123
NH HM **FH HM**	.322 = .014	.090 = .491	.157 = .246	.231 = .060	.213 = .105	.148 =.357	.119 = .380	-.017 =.915	.158** = .558
NH LM **FH LM**	.598 = .001	.424 = .021	.301 = .079	.376 = .012	.205 = .195	.546 = .006	.167 = .247	-.461 = .008	.269** = .037
FH HM **NH LM**	.090 = .456	-.122 = .327	.087 = .481	.053 = .661	.132 = .300	-.153 = .275	.135 = .281	.100 = .446	.048** = 1.00
NH HM **FH LM**	.830 < .001	.636 = .001	.545 = .003	.660 < .001	.551 = .001	.541 = .011	.421 = .007	-.378 = .056	.476** < .001

The group/group-pair with higher wrongness ratings in each comparison is listed on top. Mean differences detailed above *p*-values; non-significant relationships are grey-scaled. Scenario stats from *t*-tests. Combined via 2x2x8* or 4x8** with Bonferroni adjustments. **NH** = Not Happy, **FH** = Felt Happy, **LM** = Low MAIA, **HM** = High MAIA.

Inspection of these effects suggests any differences in wrongness ratings over liberty, authority, and counter-normative scenarios are more strongly influenced by interoceptive awareness - with any effect from feeling happy being non-significant here. Fairness and non-moral scenarios show differences by MAIA when participants felt happy, and differences by happiness when they had low-MAIA scores. However, there were also two-way interaction effects showing differences by happiness over scenarios of harm, loyalty, and sanctity. For loyalty scenarios, differences by MAIA were significant for both happy and not happy participants, as were differences by happiness for low-MAIA participants. Sanctity scenarios show differences by happiness for low-MAIA participants, although differences by MAIA when participants felt happy were short of significant - whereas harm scenarios show differences by MAIA when participants felt happy, as well as differences by happiness for both high- and low-MAIA participants.

In combination, fairness scenarios show evidence for a two-way, felt happy x MAIA, interaction effect - as do non-moral scenarios (in the opposite direction). Sanctity scenarios also show a strong trend toward this effect. Scenarios relating to harm and loyalty show a partial trend toward interaction, but appear to be more strongly influenced by happiness and MAIA ratings respectively. Liberty scenarios also showed a slight trend toward interaction, although there was no evidence for this effect over counter-normative scenarios or those relating to authority violations. It is worth re-stating that differences between not-happy-high-MAIA participants and happy-low-MAIA participants were almost always significant, whereas any differences between not-happy-low-MAIA participants and happy-high-MAIA participants were always non-significant.

It is further notable that, whilst the felt disgust x MAIA results trended towards those of felt disgust x PBC, felt happy x MAIA results markedly diverge from those of felt happy x PBC. Differences by PBC tended to be greater when participants did not feel happy and smaller when they felt happy (albeit non-significantly).

None of the other three-way analyses returned a significant result for the three-way interaction. However, there were two-way interaction effects for felt emotion on judgement type present across every iteration of the exploratory analyses - with the neutral emotion iteration being the notable exception [F(4.134,2095.959) = .561, p = .697, $\eta p2$ = .001]. Statistics are reported from 2x8 ANOVA's to avoid any minor discrepancies between figures over PBC and MAIA iterations. The two-way interaction effect between felt emotion and judgement type was significant for felt anger [F(4.206,2136.452) = 6.912, p < .001, $\eta p2$ = .013], felt disgust [F(4.213,2127.554) = 7.441, p < .001, $\eta p2$ = .015], felt fear [F(4.172,2115.419) = 2.489, p = .039, $\eta p2$ = .005], felt sadness [F(4.136,2092.943) = 2.517, p = 0.38, $\eta p2$ = .005], and felt happiness [F(4.184,2117.098) = 4.446, p = .001, $\eta p2$ = .009].

Examining these interactions with independent t-tests showed occasional differences in wrongness ratings over different types of scenarios for participants feeling (versus not feeling) certain emotions. Participants who felt a negatively valenced emotion consistently rated scenarios depicting violations of liberty as less wrong. This difference was significant for felt anger [t(508) = 3.316, p = .001], felt disgust [t(505) = 2.459, p = .014], felt fear [t(507) = 2.588, p = .01], and felt sadness [t(506) = 2.418, p = .016], but not felt happiness [t(393.223) = 1.595, p = .111]. However, participants who felt happiness gave lower wrongness ratings over scenarios depicting violations of harm [t(419.564) = 3.719, p < .001], loyalty [t(506) = 2.431, p = .015], and sanctity [t(396.378) = 2.567, p = .011]. Participants who felt sadness gave lower wrongness ratings over scenarios depicting violations of authority [t(506) = 2.188, p = .029], and sanctity [t(506) = 2.112, p = .035]. Also, participants who felt anger gave lower wrongness ratings for violations of loyalty [t(506) = 1.993, p = .047]. Any differences by felt emotion were non-significant over violations of fairness and counter-normative scenarios. Interestingly, the trend for non-moral scenarios to be rated as more wrong by participants feeling an emotion was only significant when participants felt anger [t(450.974) = -3.637, p < .001] or felt disgust [t(397.092) = -4.188, p < .001].

8.3.8. Exploratory analysis variation checks

Further analysis iterations were run to investigate whether certain significant results found during the exploratory analyses may have been dependent on aspects of data processing. For example, the criteria for being included in the felt disgust group was relatively low (*median* = 1.67), although participants were relatively evenly distributed across felt disgust x PBC categories and there was no significant difference between felt disgust groups over PBC scores. In contrast, the criteria for being included in the felt happy group was closer to the scale mid-point (*median* = 3), but the distribution of participants across felt happy x MAIA categories was more unequal, and participants who felt happy also happened to provide higher MAIA scores than those who did not feel happy.

Investigation showed the results for felt disgust appear relatively robust. The pattern of results appeared confirmatory when re-running the analysis using ANCOVA, and when varying the inclusion criteria for the felt disgust category. The felt disgust x PBC (moderated amplification) effect remained present if the criteria were increased so that inclusion required a rating of more than two points over the felt disgust measure, and also remained if the criteria was weakened such that participants were included if they reported any non-minimum rating for any of the felt disgust items. The interaction effect remained present when comparing participants who felt *any* disgust with those who felt none. This was also the case for the felt disgust x non-moral (moralisation) effect, which remained significant each time.

Examination of a similar iteration over the felt happiness analysis showed slightly greater variation when altering the inclusion criteria for the felt happy category. The felt happy x MAIA (moderated suppression) effect became more apparent across the majority of scenarios when changing the inclusion criteria for feeling happy from above-median to median-and-above happiness ratings. This reallocation made the felt happy x MAIA group sizes more equal, although there remained a between groups

difference on MAIA scores. However, re-running the analyses using ANCOVA once again appeared confirmatory. Parameter estimates were broadly reflective of the results reported for felt happiness X MAIA x judgement type, such that the reported results do not appear to be unduly influenced by dichotomizing the interoceptive awareness measure. If anything, changes made to either the inclusion criteria or analysis method both seem to make the reported felt happy x MAIA interaction more apparent in the data.

8.3.9. Summary of Exploratory Analyses

Focusing on the emotions which participants report experiencing after having rated the scenarios may have potentially compensated for any issues with induction methodology - the analysis is based on whether the participant is actually feeling a particular emotion, rather than mere allocation to an emotion induction condition.

Starting with main effects, the exploratory analyses showed a direct suppression effect for happiness, with participants who felt happy (above median happiness) rating moral scenarios as less wrong than participants who did not (below median happiness). A similar suppression effect was also apparent when comparing groups based on median sadness ratings, with participants who felt sad also tending to rate moral scenarios as less wrong than those who did not feel sad. There was also a main effect for interoceptive awareness, and a (non-significant) trend towards this for private body consciousness, whereby participants with high scores on this measure consistently provided more extreme wrongness ratings (i.e., more wrong for immoral, less wrong for non-moral) than those with low scores. It is also worth recalling that the planned analyses showed effects by judgement type, as this may provide partial explanation for interactions with sanctity scenarios typically failing to cross the significance threshold.

The two-way interactions between felt emotion and judgement type suggested support for the moralization hypothesis, and some degree of specificity in this regard, as

the trend for non-moral scenarios to be rated as more wrong by participants feeling an emotion was only significant when participants felt anger or felt disgust (i.e., moral emotions). However, the two-way interactions also suggested that feeling any negatively valenced emotion resulted in reduced wrongness ratings over scenarios depicting violations of liberty - although this may somehow be a feature of such violations. Additionally, feeling happy led to lower wrongness ratings over harm scenarios, feeling sad led to lower ratings for authority scenarios, and feeling either of these emotions led to lower ratings for sanctity scenarios. These are reflective of the main effects for these emotions, although the appearance of specificity in these cases in unexpected. Furthermore, wrongness ratings of loyalty violations were lower when participants felt happy, but were also lower when they felt angry - which is more challenging to explain, as the effect for feeling angry was expected to run in the opposite direction.

The exploratory analyses also showed evidence for a PBC-moderated amplification effect of disgust on moral judgements (i.e., a two-way between-groups interaction). Participants in the above-median private body consciousness group rated the majority of scenario types as being significantly more wrong than those in the below-median group, but only when they also reported feeling disgust. Private body consciousness had no significant influence on wrongness ratings for participants who did not feel disgust. This interaction effect was specific to the felt disgust analysis, although it was not exclusive to any particular scenario type.

Examination of the three-way interaction effect shows the relative consistency of the two-way interaction between felt disgust and private body consciousness. That this effect did not reach statistical significance over sanctity and non-moral scenarios might be explained with reference other effects. Sanctity violations were rated as more wrong than any other type of violation, and the ratings given by low-disgust-low-PBC participants are notably (although not significantly) deviant when compared with other scenarios in this regard (see Figure 8.7), such that this might be explained by the main effect for judgement type. Similarly, the two-way interaction between felt disgust and

judgement type, showing feeling disgust resulted in higher wrongness ratings for non-moral scenarios, may provide some explanation for why the interaction effect was non-significant for this type of scenario.

The exploratory analyses further showed evidence for a MAIA-moderated suppression effect of happiness, apparently exclusive to certain types of moral judgements. A two-way between-groups interaction (felt happy x MAIA) was apparent over scenarios of harm, fairness, and sanctity, as well as over loyalty and non-moral scenarios to some extent, but this interaction is not present over liberty, authority, or counter-normative scenarios (see Figure 8.13). Across all scenario types, low-MAIA participants who felt happy provided significantly lower (higher for non-moral) wrongness ratings than those in any other happiness x MAIA combination (see Figure 8.11). The interaction effects show that - for fairness and non-moral scenarios - any differences between high- and low-MAIA participants were only significant when participants also felt happy.

This interaction was also apparent over sanctity scenarios, although the difference in ratings between happy-high- and happy-low-MAIA participants was short of significant. However, the effect of happiness on sanctity scenarios appears largely dependent on interaction with MAIA. The difference by happiness for high-MAIA participants was non-significant, but the difference by happiness for low-MAIA participants was larger than for most other scenario types - and there was virtually no difference by MAIA for low-happy participants.

Loyalty scenarios show a partial trend toward interaction, with a larger difference by MAIA when participants felt happy over not happy, although there was also a difference by happiness for low-MAIA participants. The ratings over loyalty scenarios thus appear less reliant on interaction effects, although the difference by happiness for high-MAIA participants was only bordering on significant, suggesting the influence of MAIA was greater than that of happiness for ratings over loyalty scenarios.

Harm scenarios also show the interaction pattern, with the difference by MAIA significant when participants felt happy but only bordering on significant when they did not. However, harm scenarios further showed differences by happiness for both high- and low-MAIA participants, suggesting the influence of happiness was greater than the influence of MAIA for these scenarios. The ratings over harm scenarios are reflective of influences from both happiness and interoceptive awareness, although the interaction effect still appears to exert some influence. The mean difference between ratings from high- and low-MAIA participants who felt happy was more than double the mean difference between ratings of those who did not feel happy, and the difference between happy-low-MAIA and not-happy-high-MAIA participants over harm scenarios was higher than for any other scenario type.

Table 8.13 summarizes the results of the exploratory analyses with regard to the study hypotheses detailed in Table 8.1.

Table 8.13. Summary of effects found during exploratory analyses.

FELT EMOTION	HARM	FAIRNESS	LIBERTY	LOYALTY	AUTHORITY	SANCTITY	COUNTER-NORM	NON-MORAL	COMBINED
ANGER	*n.s.*	*n.s.*	**v**	**v**	*n.s.*	*n.s.*	*n.s.*	**^**	*n.s.*
DISGUST	**# ^**	**# ^**	**v** **# ^**	**# ^**	**# ^**	*n.s.*	**# ^**	**^**	**# ^**
FEAR	*n.s.*	*n.s.*	**v**	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*
SADNESS	*n.s.*	*n.s.*	**v**	*n.s.*	**v**	**v**	*n.s.*	*n.s.*	**MAIN v**
HAPPINESS	**v** **# v**	**# v**	*n.s.*	**v** **# v**	*n.s.*	**v** **# v**	*n.s.*	**# v**	**MAIN v**

^ = Amplification, v = Suppression, # = Interaction. Simple effects underscored.

8.4. Discussion

The aims of the study were to examine claims of emotion specificity and/or exclusivity across moral foundations through testing whether emotion induction amplifies (or suppresses) moral judgements and moralizes non-moral judgements. Unfortunately, the study design and methodology failed to sufficiently induce the target affective state in the majority of conditions, with disgust induction the only condition where post-test emotion ratings were fully distinct. The planned analyses failed to find any evidence of either direct or moderated amplification (or suppression) effects; participants gave similar ratings of wrongness across all conditions. However, the presence of interoceptive effects on moral judgement lends some support to the broad claim that emotion is involved in the formation of such judgements. Regardless of condition, and whether or not any particular emotion was present, participants reporting greater sensitivity to bodily associates of emotions reported the majority of scenario types as being slightly (but significantly) more wrong than those reporting lower sensitivity.

A similar (but non-significant) trend was also apparent for private body consciousness, with those scoring higher also tending to provide harsher moral judgements, and both measures correlate to a similar extent with wrongness ratings. The effects are small and more consistently apparent when measured using the more detailed of the two instruments, and the pattern of results seems noteworthy. Private body consciousness did not seem to affect ratings of 'harm', 'fairness' or 'sanctity' violations, and interoceptive awareness did not appear to affect ratings of 'sanctity' violations. The main effect is broadly supportive towards claims made by Johnson et al. (2016) over those of Schnall et al. (2015). The effect was not dependent on any emotion being elicited, and appeared weaker or absent in relation to the moral foundation(s) it is most commonly hypothesized to act on (i.e., harm/fairness, and sanctity). The effect found here was a main effect (as in Johnson et al., 2016), rather than an interaction.

The results also show, in support of Landy and Goodwin's (2015a) assertion, that the induction of high arousal negatively valenced emotions and/or just reading the content of certain moral scenarios may lead participants to harbour negative affect towards the experimenter. The results further suggest it is the type of affect that participants actually experience, rather than the type of affect they are being induced toward, which matters most with regard to generating negative affect toward the experimenter. There was also a small correlation between experimenter effects and measures of private body consciousness/interoceptive awareness, showing a relationship between feeling more positively towards the experimenter and higher scores on interoceptive measures. However, examination showed any experimenter effects on other measures were non-significant, which runs counter to Landy and Goodwin's (2015a) suggestion that such experimenter-directed disapprobation may act as a confound on wrongness ratings. Induction condition had a moderately sized effect on experimenter ratings, but how participants felt towards the experimenter did not appear to significantly influence wrongness ratings in any way.

The general pattern of emotion induction results might be taken as favouring Gray and Schein's (2018) approach to emotion, as ratings were readily distinguishable by valence (i.e., happy/neutral vs. 'negative'), and also with regard to arousal level (i.e., high - anger/fear/disgust vs. sadness - low). This provides some impetus to further explore Cameron et al.'s (2015) contention that including 'fear' as a condition and/or measure in such experiments can act as an important control when investigating 'anger'. The apparent overlap of these two emotions in the results is also of interest as the induction method is relatively 'content free', and this absence of content may explain some of the overlap (i.e., through shared valence/arousal levels). However, this explanation would also suggest that inducing *only* the target emotion may require the addition of 'conceptual knowledge' (i.e., including 'content'). Yet as this is argued to act as a confound on wrongness ratings, particularly with regard to disgust (see Schein &

Gray, 2018), it further complicates investigations in this area. If the induction is content free, then the affect may be classified as fear instead of anger - confounding the results, but if conceptual knowledge is used to anchor the affect as relating to anger, then this places much stronger demands on the sample size. Using the full design suggested by Cameron et al., (2015), with conceptual knowledge added as another between subjects factor, would require a substantially larger sample size to sufficiently power adding even the simplest version of this factor.

Although the results suggest emotion induction achieved a degree of success, there would be room to challenge the study findings even if the planned analyses had found the effects of interest to the hypotheses. As administering emotion induction checks at a pre-test stage are argued to nullify the effects of interest (see Schnall et al., 2015), there is no way of telling if the post-test results are solely reflective of the induction method or whether these are confounded by having read the scenarios. For example, it is unclear whether rated anger in the fear condition is a result of the induction (i.e., the music also elicits anger), or results from reading about moral violations (either of 'harm' or generally), or relates to negative affect directed toward the experimenter (e.g., the participant feels they have suffered some amount of 'harm' by taking part) which may arise via response to either the sound and scenarios (or both).

Post-test emotion checks only show emotions reportedly felt after taking the main part of the experiment and cannot comment on the providence of these emotions, which may vary by condition. For example, anger and fear ratings in the anger and fear conditions may reflect their respective induction, but anger ratings in the fear condition may reflect experimenter directed disapprobation, whereas fear ratings in the anger condition may reflect a content-free induction method. Similarly, disgust ratings given in non-disgust conditions could be reflective of induction method, scenario content, experimenter disapprobation, or any combination of these - and the providence of any emotion may also vary at the participant level.

The exploratory analyses evade this issue, focussing on the emotions participants reported feeling - regardless of their providence, but do so at the expense of direct comparison between emotions. These analyses were run on a specific emotion for each iteration and take no account of any other emotions which participants may have reported feeling. For example, participants classified as feeling disgust may (or may not) also have felt anger and/or any other emotion, such that direct comparison is prevented by the potential for participants to belong to more than one 'felt emotion' group. The number of analyses run, and potential issues with repeated sampling of the same data set, also provides greater scope for errors within the analyses - particularly false positives. As such, the results of the exploratory analyses are to be treated tentatively.

The exploratory analyses showed suppression effects for sadness and happiness, several simple interaction effects between emotions and judgement types - including moralization effects for anger and disgust, a PBC-moderated amplification effect specific to disgust, and a MAIA-moderated suppression effect specific to happiness. The majority of these findings fit within the extant literature, and the novel MAIA-moderated suppression effect for happiness might be explained as an opposing effect to the PBC-moderated effect for disgust - although there were also a few findings which do not fit well with particular hypotheses.

The suppression effect of sadness on moral judgements was small, only cleared the significance threshold when analysing with private body consciousness, and appeared to influence ratings of liberty, authority, and sanctity scenarios to a greater extent than other types. This fits with the findings of Schnall et al. (Experiment 4, 2008), who show participants feeling sad provided lower wrongness ratings than those feeling disgust, with a trend toward also being scored lower than when participants' emotional state was 'neutral'. However, it conflicts with findings of Cheng et al. (2013), who report

sadness as leading to greater moral condemnation than neutral emotions, and this effect being moderated by private body consciousness.

In further contrast to Cheng et al. (2013), fear had no apparent effect on any type of judgement, except those concerning liberty - although this difference was in the opposite direction, and seems to be more a feature of these scenarios. Participants feeling any negatively valenced emotion (i.e., anger, disgust, fear, or sadness) provided lower wrongness ratings for liberty scenarios, with ratings from those feeling happy also trending in this direction, such that feeling any emotion appears to influence ratings of liberty violations. This finding runs counter to claims that the influence of emotion on moral judgement is based on arousal (Cheng et al., 2013; Cameron et al., 2015), as the results for feeling fear were expected to be similar to results for feeling anger or disgust.

The results for anger were also unexpected, showing no overall effect on moral judgements. The apparent absence of an effect of anger on moral judgements contrasts with results from Cheng et al. (2013), but also with the findings of Seidel and Prinz (2013a,b). The apparent effect on liberty scenarios might be explained with reference to this particular type of content, although the apparent effect of anger on loyalty scenarios is more challenging to explain. Moral Foundations Theory does predict anger may influence judgements of loyalty scenarios, but that anger should serve to amplify condemnation of such violations - whereas the effect found in this study was suppression. However, the effect of anger on non-moral judgements was indicative of a moderately sized moralization effect. This fits with other findings in literature (e.g., Rottman et al., 2017), as well as all theoretical positions mentioned in the introduction.

A similar moralization effect was found for disgust, which fits with the findings of Landy and Goodwin's (2015a) meta-analysis, although the effect size found here is larger than that suggested for moralization (via disgust) by Landy and Goodwin. The

two-way interaction between felt disgust and private body consciousness supports claims made by Schnall et al. (2015) over those of Johnson et al. (2016) - a notable difference from the results of the planned analyses. The effect size here is much smaller than that reported by Schnall et al (2008), but larger than Landy and Goodwin's (2015a) estimate for (non-moderated) amplification effects arising via disgust. The results show that, when feeling disgust, participants with higher private body consciousness scores provided higher ratings of wrongness than participants with lower scores on this measure - whereas there were no significant differences between ratings from high- and low-PBC participants when disgust was not felt. Private body consciousness moderated the amplification effect of feeling disgust on moral judgement. The differences in ratings between disgusted high- and low-PBC participants were significant over the majority of scenario types - sanctity being the notable exception - and remained relatively consistent during analysis variation checks.

Effects relating to happiness also appear to conflict with Cheng et al.'s (2013) results, here showing a reduction in ratings of wrongness when participants felt happy. However, the results do fit with those reported by Seidel and Prinz (2013b), who show participants feeling happy scored scenarios as less wrong than participants feeling anger - although those feeling happy did not provide significantly lower ratings than participants in the control condition. The effect of happiness in this study appears to be moderated by interoceptive awareness, and runs in the opposite direction to the effect of disgust. The exploratory analyses reported suggest this interaction effect is primarily apparent over scenarios of harm, fairness, and sanctity - although slight variations to the analysis suggest the effect may be similar for most scenario types. The results show that happiness appears to suppress moral judgements, but that this effect seems to be mostly mitigated for participants with higher interoceptive awareness.

Lastly, given scenarios in each judgement category had been pre-normed for wrongness (5.23 ± 0.03 on a 9-point scale), the appearance of within-participant

differences across judgements categories was unexpected. The comparative increase in the magnitude of these ratings may be explained by labelling one end point of the scale 'perfectly okay' (following Seidel & Prinz, 2013a) rather than 'not at all wrong' (following Landy & Bartels, 2018), but this change seems unlikely to explain the pattern of results. That violations of sanctity were rated as the more wrong than any other type of violation runs contrary to suggestion that such violations are merely weird, less severe varieties of harm (cf. Gray & Keeney, 2015a), and might be taken as supportive of a position whereby violations of sanctity may be considered at least on par with 'harm' in terms of importance (e.g., Prinz, 2009). Furthermore, that violations of 'liberty' were rated on par with 'harm', and both were rated as worse than the remaining types of moral violation, fits well with Landy and Bartels (2018) bottom-up taxonomy of moral concepts - where 'Sanctity' and 'Liberty' are discernible moral virtues, and 'Harm/Care' might be considered the most important of those factoring under the virtue of 'Propriety'.

Overall, the planned analyses show that greater self-reported sensitivity to physiological components of emotion experience was associated with harsher judgements of most kinds of violations, and that certain types of violation were rated more harshly than others. The exploratory analyses show an amplification effect of disgust on moral judgement - moderated by private body consciousness, a suppression effect of happiness on moral judgement - moderated by interoceptive awareness, and moralization effects for both anger and disgust. However, all these results can be accommodated by either Constructive Sentimentalism (Prinz, 2009), Moral Foundations Theory (Graham et al., 2013), or the Theory of Dyadic Morality (Schein & Gray, 2018), although the latter of these may have more of a challenge in accounting for the wrongness ratings given over sanctity scenarios.

The results also show that participants affective inclination towards the experimenter (if any) may vary as a result of emotion induction and/or responding to scenarios depicting moral violations, and that this affective inclination may also be

associated with self-reported sensitivity to physiological components of emotion experience. Yet participants affective inclination towards the experimenter did not appear to affect any other measures, and any effects concerned with self-reported sensitivity to physiological components of emotion experience were small at best. Further research on experimenter effects would be useful to better evaluate Landy and Goodwin's (2015a) contention that experimenter directed disapprobation may confound the results in studies such as this. Further research involving physiological components of emotion experience may also be useful for advancing debate in this area (cf. Schnall et al., 2015, Johnson et al., 2016, Landy & Goodwin, 2015b) - particularly given differing results for different emotions in relation to measures of private body consciousness and interoceptive awareness. However, given the numerous, and sometimes competing, contentions as to what may act as a confound in studies of amplification/moralization effects, methodologies which can side-step such concerns whilst successfully targeting one particular type of affect (e.g., Tracy et al., 2019) may provide a more promising means of investigation.

Chapter 9 - On Purity

Having provided an overview of theoretical positions, a review of pertinent literature, and experimental investigation of key hypotheses, discussion now turns to theoretical evaluation and the establishment of a consilient common ground. Popper (2014) provides three criteria for assessing the merits of a particular hypothesis or theoretical position. Firstly, it must be able to explain everything successfully explained by other (previous) hypotheses. Secondly, it must withstand tests which other hypotheses do not, and not fall prey to at least some of their errors. Thirdly, it should explain things which other hypotheses do not. Thus, in order for Constructive Sentimentalism (Prinz, 2009) to come out on top, it must be able to account for the findings taken in support of both the Theory of Dyadic Morality (Schein & Gray, 2018) and Moral Foundations Theory (Graham et al., 2013), whilst simultaneously avoiding their errors and charting new territory.

The explanatory power of Constructive Sentimentalism stems from its empirically informed philosophical roots, which provide a means of bridging differing theoretical positions and subsuming their findings into a common account. Given Prinz (2009) describes the approach of Moral Foundations Theory as compatible with Constructive Sentimentalism, the primary focus in what follows in on the Theory of Dyadic Morality. Indeed, if moral judgement is not reliant on perceptions of dyadic harm, as defined and defended by Schein and Gray (2018), then the Theory of Dyadic Morality is (by definition) an incomplete account of morality - even though it may provide a good account of moralization processes, and of 'harm' as defined by Moral Foundations Theory (cf. Haidt et al., 2015, although see Skitka et al., 2018). Much of the merit in the Theory of Dyadic Morality may be preserved, but its ambitions likely exceed its abilities. This is best illustrated by returning to previously advanced arguments regarding the

importance of 'purity' - a construct which Constructive Sentimentalism claims is on par with that of 'harm'.

9.1. Purity has Primacy

In discussing the ethics of Divinity, Shweder et al. (1997) state "[t]his discourse ultimately brings one full circle, back to the origins of the sacredness of the individual, human or otherwise" (p.148). They also state "[a] particular feature of the Hindu worldview is the disposition to make connections between all aspects of secular, domestic, and psychological life and *a sacred order that is the ultimate reference point for all sources of obligation*" (p. 149, *emphasis mine*). Furthermore, Shweder et al. (1997) take care to note that discourses around divinity need not be theistic in nature - "[t]he central theme is reverence for the forms of the world, the realization that pleasure and pain, right and wrong, are communicated through those forms and that the world communicates its message in accordance with the way one acts towards its forms. *Reverence motivates taking seriously the obligations inherent in autonomy and community*. It motivates as well as suspension of ultimate judgement and an antidogmatic attitude toward the 'letter of the law'. A reverential attitude places responsibility for moral discrimination with personal intentionality, intellect, and will" (p. 149, *emphasis mine*). On Shweder et al.'s (1997) approach, the ethics of Divinity - concerned with the sacred/natural order, seems to take primacy. It undergirds both autonomy and community ethics, and ensures everything is encompassed within the domain of morality. The beginning and end connect like an ouroboros. Purity is the alpha and the omega.

9.2. Purity is a Problem for Dyadic Morality

Shweder et al.'s (1997) presentation of Divinity concerns as having paramount importance, such that other moral discourses may all be related back to this ethic, can be juxtaposed with TDM's claims. Following Shweder et al., one might argue harm is wrong because it is impure. Results from both Gray and Keeney (2015a) and Franchin et al. (2019), showing violations of harm being rated as more impure than violations of purity, might be taken in support of these claims. Conversely, Schein and Gray (2018) argue impurity is wrong because it is (dyadically) harmful. However, purity seems to occupy many different roles during the course of their argument. In some cases, it appears associated with the cause of suffering - "the specific act *can cause damage through* physical destruction, mental suffering, or *spiritual defilement*." (Schein & Gray, 2018, p. 3, *emphases mine*). In other places it appears able to take the role of victim - "[b]y intermediary, we are referring to a concept (e.g., purity) that is *seen as a vulnerable entity in its own right*. For example, in explaining the immorality of a purity violation such as a widow eating 'hot' food, one of Shweder's Indian participants noted that the act not only impacts her deceased husbands' spirit (a direct harm), but *the act will lead her to 'lose her sanctity' (i.e., harm her purity), a consequence that in turn leads to her tangible suffering* (Shweder et al., 1987, p. 44). Dyadic morality therefore allows for two links to harm: the direct perception that an act is harmful, and the indirect perception *that an act destroys a value* — which then causes direct harm." (Schein & Gray, 2018, p. 16, *emphases mine*). A third usage is apparent in considering Schein and Gray equate norms and values, such that foundational values fall outside the template of dyadic harm, but within the template of immorality - moral foundations such as Sanctity relate to varieties of norms.

The plurality of roles is further apparent during arguments TDM puts forward regarding constructionism. "In morality, constructionism means that *different moralized concerns (e.g., loyalty, purity) consist of various combinations of norms, affect, and*

perceived harm. For the sake of argument, imagine that — across cultures — there were five different varieties of agents (e.g., gods, adults, groups), 20 different varieties of acts (e.g., hitting, insulting, *defiling*), and 15 different varieties of vulnerable patient (e.g., children, adults, animals, souls, *social order*). When multiplied together, this number would give the possibility for 1,500 varieties of moral judgment — and we haven't even yet considered the nuances of norms." (Schein & Gray, 2018, p. 24, *emphases mine*). Sidestepping this fourth usage of purity, where it is described as a combination of TDM's three elements of immorality, if "patients can be anything perceived to have vulnerability" (p. 3), this would seem to jeopardize one of TDM's most basic claims - that harm "involves *two* perceived and causally connected minds" (p. 1, *emphasis mine*). It is by no means clear how concepts of purity, or the social order, can be classed as a moral patient when concepts have no obvious capacity for victimhood. Thus, in claiming purity as an intermediary of harm, Schein and Gray (2018) seem to undermine an important definitional premise in their argument - "Experience — *being a vulnerable feeler* — is thus what *qualifies one as a moral patient* (who possesses moral patiency)" (p. 7, *emphasis mine*).

Indeed, factoring purity into position is highly problematic for the Theory of Dyadic Morality. If impurity relates to ways in which damage can be caused, such that sacrilege and spiritual defilement are types of action, this leaves the claim that perceiving harm in such actions also involves perceptions relating to thinking agents and vulnerable patients - it is a claim that these types of actions are perceived as instances of *dyadic* harm. The claim remains the same if impurity is considered as a type of norm, such that impure norm violations are considered wrong commensurate with the extent to which they (generate negative affect and) are perceived as instances of dyadic harm (i.e., 'intentionally caused suffering'). Indeed, "[o]ne could argue that the diversity of moralized norms boils down to this subset of five or six [moral foundations], but one could then argue that these five boil down further — perhaps to concerns about harm." (Schein & Gray, 2018, p. 15). In both cases, the key claim might be summarized as

'violations of impurity are perceived to have victims', as it does not seem possible for TDM to maintain a position whereby norm violations may be considered harmful in and of themselves (i.e., where the norm *is* the victim).

9.3. Purity and Patiency

If concepts (or norms) can be seen as vulnerable entities in their own right (i.e., as *intermediaries* of harm), such that (relatively 'foundational' concepts of) purity and the social order may be considered as varieties or examples of vulnerable patients, then this would necessitate theoretical revision. Either TDM must jettison its definition of patiency - that patients are vulnerable *feelers* capable of *experience* - which would substantially undermine the core of Schein and Gray's (2018) argument, or TDM must explain precisely how such concepts qualify as meeting TDM's own definition of patiency. Such qualification would also necessitate clarification with regard to notions of dyadic reversal, where "perceptions of mental agency and patiency are exactly opposite the physical structure of the act" (p. 8), which TDM draws on in explaining how it might account for phenomena such as victim blaming. Even if concepts could meet the definition of patiency provided, this would seem to make them a particularly special case of patient; and if concepts can be patients, then it is not clear what might prevent them from further being able to take the role of (thinking, intentional, doing) agents. Indeed, the (potential) capacity for agency seems common to all other types of vulnerable experiencing feelers - one may give (cause harm) as one has received (suffered), and vice versa. Furthermore, even if any such qualification could be provided, it would either apply to all kinds of norms - rendering the definitions of 'patient' meaningless, or TDM would need to explain why only certain norms achieve status as vulnerable entities. This seems equivalent to asking why only certain norms achieve 'sacred' (or 'foundational') status - which would appear to be what Moral Foundations Theory aims to explain.

Schein and Gray's (2018) definition of 'moral patiency' thus seems incompatible with one of their definitions of purity as an 'intermediary' of harm - both definitions may be valid individually, but they cannot be actualized simultaneously. On a weaker definition, TDM claims it "supports diverse moral concerns such as loyalty, purity, industriousness, and social order, but suggests that they are best understood as 'transformations' or 'intermediaries' of harm, values whose violation *leads to* perceptions of concrete *harm*." (p. 3, *emphases* mine). However, this sounds similar to dyadic completion, and this too would seem to put purity in prime position. Dyadic completion refers to moral condemnation (e.g., of a purity-norm violation) driving perceptions towards identifying the missing element of dyadic harm (agents, patients, or causes). The corollary process in the dyadic loop, that of dyadic comparison whereby the perception of norm violations as (dyadically) harmful leads to their moral condemnation, would either need to follow after dyadic completion (contra Schein & Gray's argument, 2018, p. 17), or would need to have pre-established the norm violation as immoral due to it having been perceived as an instance of dyadic harm. Yet even if violations of purity can lead to harm by other means, a core claim remains - 'violations of impurity are perceived to have victims' - and this claim has been founding wanting for evidence.

Results from both Royzman et al. (2009) and DeScioli et al. (2012) show that around 30% of their participants reported consensual incest as being immoral despite also reporting that no-one was harmed by the act. Of course, TDM could argue these results emerge as a result of methodology, such that people do in fact perceive consensual incest as being harmful when you ask them about it carefully enough - they are not 'dumbfounded' (see Royzman, Kim, & Leeman, 2015). Yet incorporating this finding into TDM's position would seem to come at too high a cost elsewhere - 42% of participants in Royzman et al.'s (2015) third study reported considering the violation of moral norms as inherently wrong - that is wrong irrespective of harm, potential or otherwise. Furthermore, subsequent research on moral dumbfounding (McHugh, McGann, Igou, & Kinsella, 2017) does show that people can maintain moral judgements

in spite of explicit admissions that they are unable to provide reasons in support of their judgement (i.e., with no recourse to harm) - they are dumbfounded.

Notably, McHugh et al. (2017) also report that interviews where participants could be considered dumbfounded contained more frequent instances of amusement (i.e., smiling, laughing). There may be several explanations available as to why this is the case, but recall Franchin et al. (2019) report their participants showed a tendency to smile at violations of purity. That this seemingly content-specific response appeared more frequently when participants were dumbfounded - maintaining the wrongness of an act despite an apparent inability to articulate why they perceived the act as harmful/immoral - seems entirely consistent with notions that violations of purity (i.e., non-harmful moral violations) cannot be readily subsumed within the dyadic template. These (dumbfounding) actions are not seen as harmful, but can still be seen as morally wrong, and in such instances, these seem to tend toward eliciting a response specific to moral violations that share this (harmlessly wrong) profile - those which involve impurity.

9.4. Purity provides Parsimony

TDM's claims against the existence of content specific responses associated with impurity are also weaker than they appear. Cameron et al.'s (2015) review is dismissive of studies which use ANCOVA-based analyses, arguing there is little left over once shared variance between anger and disgust is considered - yet there are at least some morally relevant differences between anger and disgust. Anger and disgust seem differ in their relationships with intentionality, sexual impropriety, moral mitigation, and moral approval (Piazza et al., 2018), as well as with regard to aggressive responses, and relative associations with self- versus other-directed actions (Molho et al., 2017; Tybur et al., 2019). These emotions have also been shown to bias inferences made about the wrongness of content-ambiguous actions in favour of their hypothesized

emotion-domain associations (Heerdink et al., 2019). Furthermore, results from Landmann and Hess (2018) show support for what Franchin et al. (2019) term 'weak MFT' - that anger tends to be elicited more often than disgust in response to harm, and disgust tends to be elicited more often than anger in response to impurity. This pattern of results is not only present in many of the (ANCOVA-analysed) studies Cameron et al. argue against, but is also apparent over both validated MFT scenarios and scenarios depicting naturalistic moral violations (as shown in Chapter 4), and seemingly remains present even when open response options for emotion are provided (as shown in the Chapter 7).

Leaving aside competing explanations for frequent emotional co-occurrence and the appearance of 'loose correspondences' between emotions and moral domains (i.e., 'weak MFT'), TDM's claims regarding the role of disgust in moral judgement, particularly in relation to immorality and perceived harm, can also be curtailed. Although TDM allows that disgust may have a causal impact on moral judgements (via negative affect), Schein and Gray (2018) argue against the 'direct disgust' hypothesis - that moral judgements can be directly caused by disgust - instead claiming that (dyadic) harm is better predictive of immorality. However, the main paper Schein and Gray rely on in making this argument may not provide support for the weight TDM might like to place on it; and a recent study demonstrating a specific and exclusive link between disgust and impurity provides evidence which is seemingly beyond the reach of TDM to explain.

Schein, Ritter and Gray (2016) claim to show that "perceived harm mediates the link between feelings of disgust and moral condemnation— even for ostensibly harmless “purity” violations" (p. 862). However, their measures of perceived harm seem to bear minimal relationship with how perceived harm is defined by TDM - as perceiving intentional agents causing damage to vulnerable patients. In two of the three studies Schein et al. (2016) conduct, perceived harm is measured as the composite of three items that ask the extent to which an action is dangerous, threatening, and harmful.

Even allowing that this is a broadly valid measure of harm, it is not a measure of *dyadic* harm - none of the elements of dyadic harm are examined. Indeed, if harm can be considered as the "most important, frequent, and universal moral consideration" (Schein & Gray, 2018, p. 21), and this may be especially so within cultures which emphasize 'individualism' - from which the research samples are drawn, then Schein et al.'s (2016) results would be entirely as expected. Ratings of immorality were best predicted by a measure of perceived (potential) *suffering, not* by a measure of perceived *dyadicness*. Additionally, the measure of perceived harm in their first study is the Belief in a Dangerous World Scale (Altemayer, 1988). Given this scale shows some relationship to the construct of Right-Wing Authoritarianism, which Schein et al. (2016) acknowledge has links with political conservatism - and purity norms (Schein & Gray, 2018), it is perhaps unsurprising that results on the scale predict condemnation of acts involving homosexuals (marriage, kissing) - controlling for political orientation using what was likely a one-item measure is unlikely to provide sufficient control in this regard. Furthermore, the Belief in a Dangerous World Scale measures "the extent to which one believes the world is a dangerous place in which one must frequently protect oneself from physical harm" (Maner et al., 2005, p. 67). It is not a measure of beliefs about perceived harm, it is a measure of beliefs regarding the perceived natural order - it is a measure of 'purity'.

Tracy et al. (2019) shows that the ingestion of a nausea-suppressing substance (ginger) results in less severe moral judgements in response to moderately severe violations of purity. Furthermore, their study design is virtually free of potential confounds which may influence amplification effects (Landy & Goodwin, 2015a) and investigations regarding the exclusivity of emotions and moral content (Cameron et al., 2015). Indeed, attempting to explain these results with reference to certain potential confounds only serves to illustrate the difficulty TDM has in accounting for them. One might argue that half of Tracy et al.'s moderately impure scenarios involve ingestion, which is conceptually related to nausea, but then the effect would also have shown in the

placebo group (i.e., via eating a pill) as there is no obvious influence from the induction methodology which may otherwise have caused the effect (e.g., induction using the sound of an emetic event). The results cannot be readily explained by potential confounds introduced via conceptual knowledge of emotions. Similarly, some of the scenarios could be argued to include 'core disgust' elicitors, such as (sanitized) faeces, or touching the eye of a corpse, and as Tracy et al. show, ginger does suppress ratings of such kind of disgust. However, this would suggest that 'core' (i.e., non-moral) disgust is directly linked to moral judgements (without recourse to harm), such that suppressing the ability to experience core disgust can suppress moral judgements. This cannot be explained by reference to core affect, as otherwise the effect would also be apparent in other experimental categories (e.g., for harm). That the effect seems specific to 'impurity' challenges claims that disgust may work on moral judgement via negative affect, rather than operating directly. Impaired negative affect would be expected to affect ratings of all moderately (and potentially highly) severe scenarios in a similar fashion via core affect, whereas the results from Tracy et al.'s research are what would be expected when impairing a specific type of emotional response hypothesized to partly constitute judgements of morally impure acts. TDM has no current means by which to explain Tracy et al.'s (2019) results.

9.5. Persevering on Dyadic Morality

All the above issues notwithstanding, there are further issues with TDM's formulation of immorality, both with regard to measurements and definitions. First, no provision is made for *dispositional* negative affect, despite (occurrent) negative affect being detailed as a component of immorality. Given a lack of reports to the contrary, it seems reasonable to assume that participants responding to moral scenarios are not erupting in a fit of full-blown rage upon reading depictions of severely immoral actions. That participants report these as severe suggests either negative affect is not

contributing much, or participants may be reporting the strength of their affective disposition rather than their current affective state. Providing for dispositional affect allows that the strength of any moral judgement can be maintained even if the elicitation potential of negative affect might be reduced - either via contextual factors, such as by responding to questions in an experimental setting, or in individual cases where affect may be flattened in general (e.g., depression). This point is relatively minor, and simple enough for TDM to concede, but leads the way to a second issue in TDM's moral maths.

According to TDM, perceptions of immorality scale with perceptions of dyadic harm, such that "[m]oral judgment is proportional to the agency of agents, the experience of patients, and the clarity of causation between them; acts with obviously intentional agents who cause obvious damage to obviously vulnerable patients should seem both most harmful and immoral." (Schein & Gray, 2018, p. 7). Yet if violations of purity, which tend to lack obvious victims, can be considered equally (or even) more severe than actions which better conform to the dyadic template (i.e., harm), then the relative boost to ratings for such violations has to come from somewhere. To compensate for the apparent lack of patiency, these acts would need to be somehow more obviously agentic, and/or more obviously capable of causing damage; and although this is possible, it is by no means clear that it is the case - nor why it might be. However, it could further be the case that impurity involves more obvious and/or more deviant violations of norms which are contributing to ratings of immorality. Yet this, once again, equivocates between norms and patients by creating a class of norm capable of compensating - violating such norms would be considered sufficiently severe to counterbalance any severity lost as a result of the victim being less than obvious. This would seem to rank the violation of purity norms as more severe than the violation of norms about harm, which may be somewhat problematic for TDM - and arguments to this effect are readily available.

A third point follows from the relative flexibility of norms against harm compared to those concerned with purity, such that rules concerning purity are seen as more immutable than those regarding harm. Suppose someone engages in an act of planned revenge. The act is obviously intentional, obviously causes suffering, and obviously has a victim, so would seem to be highly immoral on TDM's account. Even allowing for moral typecasting (Gray & Wegner, 2009), whereby the agent is seen as less responsible and the patient as suffering less, the act is still clearly structured in accordance with the template of dyadic harm - yet there are multiple ways in which planned revenge may be morally construed. One might argue the act is wrong - two wrongs do not make a right, or one could be relatively neutral towards the act - the wrongs (and roles of agent/patient) balance each other out (cf. certain folk concepts of karma; Shweder et al., 1997), or one might actively endorse the act - exacting revenge may be considered a moral obligation within certain cultures or by particular individuals. TDM may account for the first two of these possibilities, but does not readily handle the third possibility - that the intentional infliction of suffering may be precisely the point of the act. Even if any such suffering might be construed as morally deserved, such that the infliction of it may be considered entirely morally justified, and there are no reservations about its administration, it still involves obvious suffering. TDM may surmount this objection by noting a requirement that one has at least some degree of empathy for the victim, which may be absent with regard to planned revenge, yet this requirement may be problematic for TDM when addressing other morally relevant issues.

For example, consider the non-therapeutic alteration of children's genitals. Detailed discussion of this topic is beyond the scope of the current inquiry, but it may be granted that discourse in this area does indeed revolve around culturally informed notions of harm (e.g., Earp & Darby, 2017). The problem for TDM is that it would seem possible for an individual to perceive such actions as a clear instance of dyadic harm, given what is involved, whilst maintaining that such actions are (somehow) morally permissible. TDM may claim instances of morally permissible dyadic harm are rare,

and/or maintained through effortful reasoning, but it provides scant account of why any such rare exceptions may persist. TDM does not explain what might be termed justifiable harm - instances where dyadic harm is clearly perceived, but is considered to be justified in some way. One such potential justification is that the harm is subservient to some other purpose, or greater good, such that the harm is considered in some way instrumental to achieving some outcome. Instances of instrumental harm are especially problematic for TDM. This is partly because they necessitate an explanation of how TDM might resolve moral dilemmas - such as the extent to which one can do wrong (e.g., blackmail someone) in order to achieve an outcome viewed as morally positive (e.g., so that they donate to charity), but particularly because of what purposes instrumental harm can be taken to serve. In at least some cases, the justification given for the infliction of suffering can be explicitly related to notions of community and divinity.

In the first instance, consider any harm which may occur either as a side effect, or required aspect, of initiation practises within a particular community. Of course, one might argue that the 'harm' of not joining one's community outweighs consideration of any harm inherent to the initiation, but one might also argue that inflicting some kind of harm during the passage of initiation is precisely the point. A community may recognise that its initiation practises are harmful, yet members of the community may also consider such practises as valuable in some way - such as by strengthening the bonds between members through a shared experience of suffering. Initiation practises may be considered harmful without necessarily being considered immoral. In the second instance, consider practises where harm is inflicted with reference to divinity-based ethics, such as scarification, self-flagellation, or eagle-hanging (Garudan Thookkam). Once again, at least for some cases, suffering would seem to be precisely the point - such that pain provides a means of purification, or acts as a demonstration of devotion. Admittedly, these latter examples may be less dyadic, in that any suffering is generally self-inflicted, but this leads back to concerns about the verifiability of victims. One might argue that one's future self is perceived to be the victim in certain cases, although this

somewhat stretches the meaning of 'dyadic', or one might point to findings which suggest 'purity' is more concerned with self-directed actions (Dungan et al., 2017) - relational context may be an important factor in moral judgement, but not all relational contexts are dyadic. The first instance above illustrates that harm can be done in the service of other norms - concerns for purity can outweigh those of harm; the second instance further illustrates links between purity and actions which are either apparently victimless, or only involve one person.

Even if these further points might be addressed separately, it is not clear they can be addressed simultaneously. Yet theories which advocate for moral pluralism (i.e., Constructive Sentimentalism, Moral Foundations Theory) over harm pluralism (i.e., Dyadic Morality) have little problem in accounting for the issues discussed. They can also readily accept many of TDM's claims with minimal theoretical consequence - TDM may account for some morally relevant phenomena unaddressed by other theories, but these claims may simply be subsumed within other accounts. TDM may provide a more detailed account of harm, and illustrate the importance of dyadic processing in morality, despite issues which may be taken with TDM as an account of moral judgement. Even taking a highly concessive position, whereby 'foundations' of fairness, loyalty, authority, and so forth may readily be conceived of in dyadic terms, or as factoring with harm (e.g., Landy & Bartels, 2018), the issues TDM has in providing a satisfactory explanation of purity violations remain. Considering purity either as relating to a special type of norm and/or as being concerned with self-directed transgressions may allow TDM to account for a wider range of evidence, but this would come at the price of its dyadic premise - and even then, TDM may still have trouble accounting for the appearance of content specific effects.

The fourth, final, and potentially fatal flaw in the Theory of Dyadic Morality follows from a combination of the issues raised above. Suppose there is an abstract, but completely unambiguous instance of dyadic harm of which you (or someone you care

about) are the victim - such that there is no question that all the required elements of immorality are present. A norm has been violated, someone else has intentionally caused your suffering, and you are experiencing strong negative affect (e.g., anger) as a result. Surely this is immoral! Yet what makes this the case? Suppose further that your negative affect could have arisen for one of two reasons. The first possibility is the one advanced by all accounts, negative affect does indeed relate to, or follow from, the morality of the act. The second possibility is more problematic - the reason for your negative affect may be that you missed an opportunity. Sure, you suffered in this instance, but you would have had no compunction about doing the exact same thing to the other person if your roles could be reversed - if only you had thought of doing it first! Any reaction in the first case might be expressed as 'what a horrible thing to do, I would never do something like that!', whereas the second case might be expressed as 'what a horrible thing to do, but I would have done the same if given the chance!'. Yet it is unclear how one might assert or maintain that the action is immoral in this second case, rather than merely inconvenient - either there would be disagreement about the meaning of the term, or its use would be hypocritical. As such, even though an act may fully meet all the criteria specified by TDM as being necessary for the act to be judged as immoral, the second possibility shows TDM's specification may still not be sufficient for it to be judged in this manner. That the exact same act might be considered wrong if you it, but considered acceptable if I do it, would seem to violate a basic requirement for consistency in definition of terms. The Theory of Dyadic Morality would seem to have the potential for hypocrisy built into its definition, which makes its current position simply unsustainable.

9.6. Prinz provides Parsimony

Constructive Sentimentalism (Prinz, 2009) effectively forecloses this point in its definition of what counts as morally wrong. Not only must the individual be disposed to

respond to the act with negative affect when someone else does it, they must also be similarly disposed towards negative affect if they were to do the act themselves. Importantly, these instances of negative affect, or emotion, must be disapprobative in both cases - the disposition must tie to emotions of blame, such that the action may be considered blame-worthy. On this approach, the first case above counts as immoral, whereas the second case does not count because the negative affect involved is not disapprobative. Furthermore, the amended definition of morally wrong would seem to rule out notions that purity may be considered as an intermediary of (dyadic) harm - it suggests the *same act* must be reversible without the admission of hypocrisy. Certain norms (i.e., purity) may be regarded as 'vulnerable entities in their own right' (i.e., moral patients), but the ways in which damage might be done to such a norm do not seem to be the same as the ways in which certain norms may be perceived to cause damage (even provided that agentic norms might be accounted for). The direction of causation does not appear reversible in the case of norms, nor would there seem to be the potential for a stable norm to damage itself in some way. If norms cannot be victims, and violations of purity can (at least sometimes) be considered victimless but still immoral, then perceptions of patiency are not necessary for moral judgement - at least not in the sense detailed by the Theory of Dyadic Morality. In contrast, Constructive Sentimentalism argues that it is emotions of (agent-focused) blame which are necessary in this regard, such that there is no need for any victim to be perceived. Indeed, emotions hypothesized as relating to victims were elicited substantially less often than those relating to agents in a free response paradigm (see Chapter 7), and less often than might be expected if concern for the particular victim(s) were a necessary component of moral judgement.

Mechanics aside, the critical point is that the disposition has (at least) two points of reference, one other-focused and one-self-focused. TDM seems to lack a second point of reference, self-focused or otherwise, which would appear to be necessary in providing a sound definition of 'morally wrong'. Yet the Theory of Dyadic Morality cannot

incorporate such a reference point without ceding some theoretical ground. The only apparent 'outs' both lead to 'purity'. One might reformulate dyadic harm, such that the observer may be considered the victim instead of any actual victim - although this might suggest that what is being harmed is the observers' worldview (i.e., their perception of the natural order). Even granting a maximally concessive position, where one may find a creative way of reformulating TDM to address the concern, it would still require the 'self' as a reference point - yet the 'self' seems aligned to 'purity' (Dungan et al., 2017). Indeed, simply granting 'purity' as having (at least) equivalent status with 'harm', such that both may be considered as relating to *grounding norms*, may be capable of addressing all the issues raised - this is precisely the approach Constructive Sentimentalism takes.

In conclusion, Constructive Sentimentalism seems able to explain everything which the Theory of Dyadic Morality can explain. This has been partly illustrated over the course of the thesis, but including any findings specific to TDM's approach within Constructive Sentimentalism seems plausible. TDM may be considered as simply providing a more detailed account of morality as related to dyadic interactions, and 'harm' or 'autonomy' concerns - it complements, rather than competes with, Constructive Sentimentalism. Compatibility in this regard is facilitated by both theories sharing a constructionist approach, such that notions of the moral dyad as a fuzzy cognitive template (Schein & Gray, 2018) fit well with Constructive Sentimentalism's supporting theory of concepts (Prinz, 2004a). The key point of difference is 'what', rather than 'how'; although Constructive Sentimentalism suggests the Theory of Dyadic Morality errs in constructing (im)morality from dyadic harm (in combination with norms and negative affect), rather than from emotions. In doing so, Constructive Sentimentalism may more easily explain any correspondences between emotions and moral content and, as it is not reliant on the perception of dyadic harm, may more easily account for the apparent immorality of (victimless) purity violations. Importantly, Constructive Sentimentalism can also explain evidence beyond the reach of TDM, such as results showing the

suppression of the ability to experience core disgust exclusively affects moral judgements of moderately severe purity violations (Tracy et al., 2019). Constructive Sentimentalism suggests putting purity on par with harm is a more parsimonious approach, and better able to account for the evidence.

Chapter 10 - On Other Foundations of Morality

Having shown that the Theory of Dyadic Morality (Schein & Gray, 2018) must cede some ground to Constructive Sentimentalism (Prinz, 2009), the discussion now turns to Moral Foundations Theory (Graham et al., 2013). The delay in contrasting MFT with Constructive Sentimentalism stems from both securing a role for purity in morality, which both theories advocate, and its degree of overlap with TDM, such that Constructive Sentimentalism can address MFT and TDM simultaneously in places. In this regard, Constructive Sentimentalism argues against the 'innate' and 'intuitive' premises of both theories. However, these are both simple to address. The claim against the innateness of *morality* follows from Constructive Sentimentalism's definition of morally wrong, as an act can only be qualified as morally wrong once an individual possesses the relevant sentiment. The means by which a sentiment comes to be possessed may be 'organised in advance of experience', but its specific content is not. The claim against intuition follows from the inclusion of sentiments, as this approach allows that moral intuitions can be related to parts of the sentimental machinery; whereas both MFT and TDM token intuitions (regarding harm or otherwise) as being responsible for moral judgements, yet do not explain how such intuitions may come to 'suddenly appear in consciousness'. Sidestepping the defence of these arguments (for detail see Prinz, 2008; 2009), and given all three theories agree morality is subject to cultural learning, the last claim of MFT is that of moral pluralism. Here, although Constructive Sentimentalism disagrees with the Theory of Dyadic Morality regarding the status of purity as an intermediary, it would seem to provide some concession to TDM in agreeing that other foundations are derivable from more basic elements.

For brevity, given both the roots of Moral Foundation Theory and that it is stated as such by the authors, the foundations of 'loyalty/betrayal' and 'authority/subversion' may be considered as relating to Shweder et al.'s (1997) ethic of community.

Constructive Sentimentalism argues such norms relate to the 'natural order of persons', and are derived from grounding norms - those concerning the 'natural order' (i.e., divinity/purity/sanctity/degradation) and those concerning 'persons' (i.e., autonomy/harm/care). This definitional derivation aside, both foundations might also be considered in more general terms. For example, construing loyalty more in line with MFT's earlier work, where the foundation is labelled 'in-group', may relate this to the extent of moral concern - or in the dyadic sense, to whom moral concern is extended and applied (i.e., patiency). Indeed, certain foundational values may be considered more relevant when the victim is liked (Erikkson, Simpson & Strimling, 2019); and we may be inclined to protect close others from the consequences of their immoral actions (Weidman, Sowden, Berg & Kross, 2020). Similarly, construing authority in more general terms may relate this to norm enforcement as an aspect of broader concerns about the social order (i.e., the natural order of persons). Authority may also provide some form of licensing for acts which would otherwise be of moral concern - or in a more dyadic sense, may permit some individuals (e.g., police officers) to harm others (e.g., criminals) in certain circumstances (e.g., to prevent harm to the public). Similarly, authority may also place greater restraints on the actions of those who hold it, such that those in positions afforded respect may become disliked to a greater extent than others were they to behave immorally (Lu, Peng, Liao & Cui, 2019).

Positioning the 'fairness/cheating' foundation may also illustrate the importance of considering relational context in moral matters. For example, suppose you are the victim of some kind of fairness violation, such that in this case, you may construe the act as being a violation of (your) autonomy. In contrast, suppose someone else is the victim - construal could seemingly go one of two ways. You may still construe it as a violation of (someone else's) autonomy, or you may construe the act more as a violation of the perceived natural order of persons (or even both). Differences in construal may be even more apparent when substituting loyalty or authority violations for fairness in the formulation above, such that when oneself is not the target of such violations, these

seem less likely to be construed in terms of autonomy. This may help explain the results of studies where moral judgements were best predicted by care, fairness, and purity for both liberals and conservatives (Frimer, Biesanz, Walker & MacKinlay, 2013), studies showing fairness and purity both factored most prominently in concerns about inequality (Franks & Scherr, 2019) and candidate choice in the US 2012 Election (Franks & Scherr, 2015), and studies suggesting care, fairness, and purity as being more emotive than loyalty and authority (Landmann & Hess, 2018). These findings suggest that whilst 'fairness/cheating' and 'care/harm' may contribute similarly as 'individualizing foundations', the 'sanctity/degradation' foundation may contribute disproportionately in comparison to the other 'binding foundations' (loyalty and authority); such findings might also be taken as supportive of Constructive Sentimentalism's formulation whereby violations of the 'natural order of persons' are a derived class of transgression.

These brief points of critique serve to illustrate how certain moral foundations - those associated with autonomy (care and fairness) and divinity (sanctity) - may be considered as having greater moral relevance than others (i.e., loyalty and authority), but also show how each foundation may serve different moral roles and functions. These points also fit well with Constructive Sentimentalism's argument for grounding norms, given their associations with concern for 'persons' and 'the natural order' respectively. Indeed, much of the preceding material has been focused on establishing that notions of purity cannot be readily accounted for in terms of concern for persons (i.e., dyadic harm), and investigating claims of emotion-content associations argued to support the distinction of purity concerns from those of harm. Given that purity concerns cannot be easily explained away, that the Theory of Dyadic Morality must seemingly admit of some purity-aligned point of reference, and that Moral Foundations Theory and Constructive Sentimentalism may be considered compatible through their shared points of reference, it would seem reasonable to propose that Constructive Sentimentalism may be able to bridge these other theories and provide common ground between them. However, a recently developed theory, which closely follows MFT's approach, makes claims which

seem diametrically opposed to those which have been advanced thus far. Constructive Sentimentalism argues that 'harm' and 'impurity' are fundamental elements of morality, whereas this theory argues that 'harm' and 'impurity' are not coherent and distinct moral domains.

10.1. Morality as Cooperation

The Theory of Morality-as-Cooperation (MAC; Curry, Mullins & Whitehouse, 2019) starts from a similar premise to Moral Foundations Theory (Graham et al., 2013) - that morality relates to a set of 'recurrent adaptive social problems'. However, MAC takes a more systematic approach in advancing this premise, proposing that morality has the function of promoting cooperation. Curry et al. draw on mathematical models of cooperation (i.e., game theory) in identifying a range of different (non-zero-sum) cooperative behaviours, finding these behaviours seem to be regarded as uniformly positive within a range of diverse societies. "The present incarnation of the theory incorporates seven well-established types of cooperation—helping family, helping group, exchange, resolving conflicts through hawkish and dovish displays, dividing disputed resources, and respecting prior possession—and uses this framework to explain seven types of morality—obligations to family, group loyalty, reciprocity, bravery, respect, fairness, and property rights." (Curry, Mullins & Whitehouse, 2019, p. 3).

Although MAC is an anthropological theory, and still in its infancy, it provides considerable challenge to Moral Foundations Theory's account of 'foundations'. Both theories advertise as being incomplete, in that there are likely more 'foundations', or forms of cooperation, than either currently advances. However, MAC covers more ground, with greater nuance, from relatively more secure theoretical underpinnings. It provides foundations dedicated to types of altruism (e.g., kin), heroism, and property rights, distinguishes fairness from reciprocity, and allows that advances in game theory

may provide novel predictions and explanations with regard to forms of cooperation. Furthermore, these foundations seem to be more distinct than those of MFT, with measures relating to MAC showing greater internal reliability and more coherent factor structuring (Curry, Chesters & Van Lissa, 2019). In contrast, when analysing the factor structure of the Moral Foundations Questionnaire, Curry, Chesters and Van Lissa's (2019) results show that a (by now familiar) three factor structure may be a better fit, as care loaded with fairness (i.e., autonomy), and loyalty loaded with authority (i.e., community).

Merits of Curry and colleagues approach aside, the pressing issue for the current project is that MAC argues against 'harm' and 'impurity' as distinct moral domains. The argument follows from MAC's formulation, in that neither 'care' nor 'purity' seems related to any distinct type of cooperation. On this approach, the moral valence of harm is context dependent, such that un-cooperative harms are seen as immoral (e.g., battery), cooperative harms are seen as moral (e.g., punishment), and certain kinds of competitive harm (e.g., mixed martial arts) may be seen as morally neutral. Similarly, MAC contends that disgust (*viz.* pathogen avoidance) is moralized to the extent it relates to cooperative behaviours, particularly those which may be partially solved by 'avoidance'. Furthermore, MAC contends that aspects of sexual morality, which are suggested to be strongly 'purity' orientated (Schein & Gray, 2018), may be explained with reference to each of the seven moral domains they identify. Additional work is needed to confirm whether these contentions are correct, particularly with regard to purity - as MAC only addresses proxies of this, but the overall approach seems compelling. Yet despite MAC advocating for moral domains in seemingly direct opposition to those argued for under Constructive Sentimentalism, it may still be possible to reconcile these approaches.

Chapter 11 - Common Ground

Navigational turns of phrase are relatively common when it comes to morality. Morality can be mapped (Graham et al. 2011) with a compass (Curry et al., 2019), the moral landscape (Harris, 2012) can be surveyed (Janoff-Bulman & Carnes, 2013), and you can have an atlas to map out the terrain of moral psychology (Gray & Graham, 2019). You might also object to something on moral grounds, take the moral high ground in an argument, and refer to another as lacking a moral compass. Indeed, several dictionaries list 'moral compass' as a noun, with definitions emphasizing links between morals and behaviour: "the ability to judge what is right and wrong and to behave in an appropriate way" (Oxford), "a natural feeling that makes people know what is right and wrong and how they should behave" (Cambridge), "an internalized set of values and objectives that guide a person with regard to ethical behavior and decision-making" (dictionary.com), "an inner sense which distinguishes what is right from what is wrong, functioning as a guide (like the needle of a compass) or morally appropriate behavior" (wiktionary.org).

As such, it is perhaps surprising that moral psychologists do not appear to have made more functional use of such terms in approaching the topic of morality. The term 'moral compass' seems mostly used as a linguistic device, although sometimes it may be used to figuratively illustrate a constellation of values, and on rare occasions authors have drawn links with how an actual compass works (e.g., Moore & Gino, 2013). However, it does not appear that anyone has seriously pursued notions of navigation with a (moral) compass in any practical sense. The aim in what follows is to outline navigational metaphors of morality as a means of combining, locating, and orientating different moral theories.

The thesis thus far has focused on establishing the relative merits of Constructive Sentimentalism (Prinz, 2009) in comparison to both the Theory of Dyadic Morality (Schein & Gray, 2018) and Moral Foundations Theory (Graham et al., 2013). The approach in this regard has been along two fronts, experimental and theoretical. On the experimental front, it has been shown that violations of purity are not necessarily weirder or more severe than violations of harm, that participants readily distinguish these types of violation in line with a priori content classifications, and that emotions of anger and disgust were elicited with relatively greater intensity in response to violations classed as harmful or impure respectively (see Chapter 4).

This pattern of emotion selection was also found when investigating the emotion elicitation patterns proposed by Constructive Sentimentalism in an open-response paradigm (see Chapter 7). Importantly, the results of this study lend support to some of the main hypotheses regarding any associations between morality and emotions. Moral scenarios elicited more emotions, more often, and more intensely, than non-moral scenarios - providing support for the elicitation hypothesis. Similarly, participants reporting an emotional response to non-moral scenarios were more likely to construe those scenarios as having greater moral relevance - providing support for the hypothesis than emotions contribute to moralization processes. Additionally, the results provide some 'weak' (see Franchin et al., 2019) support for the specificity hypothesis with regard to immoral actions, showing emotions of anger and disgust tended to be associated with moral domains of autonomy and divinity respectively. Results also suggest a broad level of support for Constructive Sentimentalism's account of morally positive emotional responses, as well as suggesting links between 'self' and 'purity' which fits within the wider literature (notably Prinz and Nichols, 2016).

Furthermore, although the last study (see Chapter 8) had some issues examining the amplification hypothesis, it did show those reporting greater sensitivity to bodily states which relate to emotion tended to rate scenarios more harshly than others -

again providing support for links between morality and emotion. Exploratory analyses conducted in this study also provide tentative support for moralization effects specific to anger and disgust, a private body consciousness moderated amplification effect of disgust on wrongness ratings, and an interoceptive awareness moderated suppression effect of happiness on wrongness ratings. In short, the results for each study can be taken favourably in support of Constructive Sentimentalism (Prinz, 2009).

Discussion of theory thus far has primarily focused on providing an overview of key theories, how these pertain to relevant aspects of the experimental hypotheses, and areas of disagreement between the approaches. This has shown that Constructive Sentimentalism (Prinz, 2009) is at least no worse than other theories on offer, and has suggested several ways in which it may be better - particularly in comparison to the Theory of Dyadic Morality (Schein & Gray, 2018). However, the preceding discussion has focused on the merits of Constructive Sentimentalism in comparison to other theories, rather than on its own terms. Drawing links between morality and navigation allows some of these comparative points to be elucidated, but also allows the merits of Prinz's approach to be better illustrated.

11.1. Moral Navigation

Firstly, recall Constructive Sentimentalism's (Prinz, 2009) definition of morally wrong, which combines "(S1') An action has the property of being morally wrong (right) just in case there is an observer who has a sentiment of disapprobation (approbation) toward it." (p.92), and "(S2-W) The standard concept WRONG is a detector for the property of wrongness that comprises a sentiment that disposes its possessor to experience emotions in the disapprobation range." (p.94). Recall also that a disapprobative sentiment is defined as one which must dispose its possessor to experience ***both self- and other-directed*** emotions of blame (or praise). The definition

links its metaphysical thesis - what it means to be morally wrong, with its epistemic thesis - showing how the concept WRONG corresponds to the metaphysical property, in such a way that relations between the two (cf. ethics, behaviour/mechanics) depend on emotion. This proposes moral concepts are response-dependent, and relative to the speaker, providing better scope to account for morality at the individual level. It also gives emotions a central role in moral psychology.

Most importantly, though, the definition incorporates dual points of reference. Indeed, the utility of this approach remains apparent even if sentimentalist approaches might be jettisoned (cf. McAuliffe, 2019), as has been illustrated in previously advanced critique of the Theory of Dyadic Morality. Definitions that do not incorporate self-reference provide permit for terms to be used hypocritically or pathologically. Prinz's (2009) definition establishes the 'self' as an important reference point for moral judgements. However, this is not just important as a matter of definition, it also better accounts for the evidence. For example, Miller and Cushman (2013) propose that negative affect in response to moral violations may arise from an aversion to the outcome, or to the action itself, arguing that evaluative simulation plays a role in the latter case. They suggest "that the process of judging third-party harmful behavior in the context of personal moral dilemmas involves asking yourself how you would feel performing the same behavior, and part of this feeling is best characterized as an aversion to particular features of the action." (Miller & Cushman, 2013, p.714). Furthermore, in terms of navigation, if one is going to evaluate one’s position on the ground then having more than one point of reference is highly beneficial.

For present purposes, the map may be best conceptualized with reference to the 'moral matrix' - a metaphor employed by Haidt (2013). Haidt describes moral matrices at the level of culture, such that they may be construed as cultural constructions of morality arising via the ways and means in which different societies have drawn on, emphasized, and circled around different conceptions of sacred values

(i.e., moral foundations). In short, "[e]ach matrix provides a complete, unified, and emotionally compelling worldview, easily justified by observable evidence and nearly impregnable to attack by arguments from outsiders." (p. 107). Situating this matrix at the level of the individual, the map can be described in the broadest terms as created from the individuals' mental model(s) of reality, such that any navigation may be conducted from inside the map. Importantly, as navigation is being conducted with respect to individual world view, the map should be orientated, or movement conducted, in such a way that the direction of travel is away from 'bad' - which is to say that no-one is actively, deliberately, and purposelessly trying to make their own life worse. This may be taken as providing the common direction between map, compass, and reality (i.e., North, although technically ‘reality’ may be better considered as 'True North' to preserve the metaphor).

Following the argument that morality requires a dual point of reference, this provides for two axes within which morality may be located, although there are a few possible ways these lines might be labelled. Following Haidt (2006), these might be labelled 'autonomy' and 'community', as Haidt stays with a stricter interpretation of 'divinity' (cf. religion) such that this is derived from 'community'. Following Moral Foundations Theory (Graham et al., 2013), these lines might be labelled 'individualizing' and 'binding' (or collectivising) if using a broad approach. Alternatively, these might be labelled 'care/harm' - given the importance of 'harm' in moral cognition, and 'sanctity/degradation' - often considered maximally distinct from the 'care/harm' foundation. This would fit with Constructive Sentimentalism's approach (Prinz, 2009), which provides a wider interpretation of 'divinity', such that lines would be labelled as relating to grounding norms concerned with 'persons' and 'the natural order'. Supposing the Theory of Dyadic Morality (Schein & Gray, 2018) is able to incorporate self-reference, lines might be labelled as relating to 'other' and 'self' in some way, which might tie in with Constructive Sentimentalism's labels on the 'good' side. In navigational terms, these lines are best analogized with lines of latitude and longitude, or northings

and eastings, such that coordinates (i.e., a grid reference) may be provided to identify any particular location on the map. This formulation provides several points of note.

Firstly, morality is being defined here with reference to the negative end of the spectrum. This is drawn partly from Prinz (2009), in that 'ought' refers to what it would be wrong *not* to do; but also draws in part from Harris (2012), who advances a negatively framed version of utilitarianism. The general thrust of Harris's argument aims to establish "that a concern for well-being (defined as deeply and as inclusively as possible) is the only intelligible basis for morality and values" (p.44), and as such, we should act in such a way so as to avoid 'the worst possible misery for everyone'. Harris's project is of interest in that morality is described as a navigation problem, and because the argument is illustrated with reference to a moral landscape - "a space of real and potential outcomes whose peaks correspond to the heights of potential well-being and whose valleys represent the deepest possible suffering." (p.19). It is beyond the scope of the thesis to address Harris's argument in detail (for critiques see Blackford, 2010; Kaufman, 2012; Nagel, 2010; Earp, 2016), but is worth mentioning given possible similarities with the current approach. Both approaches highlight that defining with reference to the negative end makes morality substantially less onerous on the individual than positive definitions. Commands so as to act in a way that does not increase misery, or to do what it would be wrong not to do, cover a relatively narrow range of possibilities in comparison to commands to act in such a way to maximize utility, or to do what it is right to do. Indeed, the positively phrased commands would seem to encompass the negatively phrased ones by definition, but not vice versa.

Secondly, Constructive Sentimentalism provides an empirical definition of 'morally wrong' which is meaningful at the individual level. Other theories either tend to provide no such definition, or tend to focus on a specific property of the act as determining its wrongness - such as the extent to which it might be perceived as violating a value (MFT), an instance of dyadic harm (TDM), or detrimental to cooperation

(MAC). There is also some overlap here with Harris, who in less guarded sections of *The Moral Landscape* reduces the claim from "...questions about values - about meaning, morality, and life's larger purpose - are really questions about the well-being of conscious creatures" (Harris, 2012, p.11-12), to a simpler version whereby it is about the well-being of conscious creatures *like me*. This fits to some extent with Prinz, who defines emotions as "perceptions of the organism-environment relationship which bear on well-being" (Prinz, 2004b) - such that emotions may be said to represent concerns. Grounding morality in emotion would therefore seem to ground morality (to some extent) in concern for well-being, as Harris suggests. Indeed, Prinz defines grounding norms as those where reasoning bottoms out in emotion (i.e., perceived concern for well-being), arguing that this is shown by the results of moral dumbfounding studies. However, this is not to say that well-being itself is necessarily considered moral - even though well-being may become moralized by different individuals to various extents. Prinz (2009) considers well-being as an extramoral value, such that reference to well-being is one of the ways in which moral rules might be evaluated. Extramoral standards of assessment (e.g., consistency, coherence, universality) are values themselves, but are outside of questions about good and evil. Constructive Sentimentalism's formulation of 'morally wrong' thus maintains strong links with individually defined (i.e., subjective) well-being, whilst being empirically tractable.

Thirdly, distinguishing different aspects of navigation illustrates different philosophical and scientific considerations. The extent to which the map corresponds to reality may be construed normatively in epistemological terms (e.g., how maps should be made - cartographic principles), or descriptively in terms typically associated with cognitive psychology (e.g., how maps are made - cartography in practice). Similarly, the extent to which the compass corresponds to reality may be construed in terms of ethics (normative – how compasses should work) or moral psychology (descriptive – how the compass actually works). However, these considerations suggest Harris's analogy of the landscape fails as it locates well-being, defined as deeply and inclusively as possible, as

relating to peaks and valleys on the landscape. Even supposing well-being relates to such locations, navigation (i.e., morality) might tell you how to get there, but is silent on whether you should go there. Hume (1739) cautioned that one does not simply derive an 'ought' from an 'is', yet this is precisely what Harris seems to do – morality is a navigation problem, yet Harris seemingly advocates mountaineering. Following the current analogy, a compass is not an altimeter, and it would be a strange compass indeed if it could point to any and every 'real peak'. Furthermore, even if 'avoid maximizing misery' might be considered as related to 'True North', there would remain open questions as to how this aligns with regard to both 'Grid North' (map/concepts) and 'Magnetic North' (compass/emotions), as well as how the map and compass relate to each other.

Fourth, the current formulation readily illustrates both why monistic approaches to morality may seem appealing, and why they fall short. Claims that morality operates via a harm-based template (Schein & Gray, 2018), or a unified concept such as 'well-being' (Harris, 2012), lack the additional point(s) of reference necessary for marking a location on a map – which is why monistic approaches fall short. However, there are other navigational tasks (e.g., taking a bearing) which typically only refer to one axis, such that one may focus on assessing the deviation from the set axis (usually 'North') to establish one's position in a similar way to which one may focus on assessing the deviation from certain norms (e.g., regarding 'persons') when making a moral judgement. Notably, this 'set axis' may differ between cultures (and/or individuals) as a result of how they have learned to 'navigate' over the course of their history.

Fifth, although there is substantial explanatory power which follows from working in two dimensions, such that the current formulation can both incorporate additional points of reference and allow for additional dimensional considerations as needed. Non-axial moral foundations (e.g., loyalty, authority) might be analogous to prominent landmarks, such that it may be possible to resect one's position by taking bearings

without necessary recourse to a particular axis – provided the landmarks are visible. This allows that such moral foundations may be navigationally relevant, even though these might be ultimately reduced to axial references - such that concern for non-axial foundations may be expressed in axial terms (i.e., reasoning turns to 'harm' or 'degradation'). Alternatively, certain foundations (e.g., loyalty/betrayal, or liberty/oppression) might relate to broader navigational concerns within the map, such as regard for travelling companions (loyalty) or freedom to choose one's route (liberty). The analogy further permits extramoral concerns to be analogized in line with various factors relating to general navigational practice. For example, questions regarding universality might be considered as questions regarding particular aspects of cartography, and questions of virtue may be analogous to the development of navigational expertise. The current navigational formulation allows multiple moral (and extramoral) concerns to be included, but aims to position these so as to illustrate their relative importance in moral matters.

11.2. Sketching a common (theoretical) map

A better illustration of some of these points is provided by detailing the equipment under discussion. Firstly, an arrow can be placed to provide a common orientation offering a fixed point of reference between map, compass, and reality – and this orientation is negatively defined, such that it points to '*not wrong*', rather than 'right'. Extending a line parallel to the arrow and intersecting this at the mid-point also provides for a graphical construction, such that the navigational analogy may be measurable, rather than merely metaphorical. In this sense, the line incorporates notions of a moral continuum (Cameron et al., 2015; Schein & Gray, 2018), whereby judgements of (im)morality are questions about 'extent' (e.g., how wrong/right?) rather than 'category' (e.g., is wrong/right?). Morally neutral actions would be plotted near the centre, maximally immoral actions would be plotted near the base of the vertical line, and

maximally moral actions would be plotted near the top. A broad outline of this basic moral map is shown in Figure 11.0 - markings may be considered in navigational terms (i.e., as grid lines), or as part of a graph whereby line length indicates scale points.

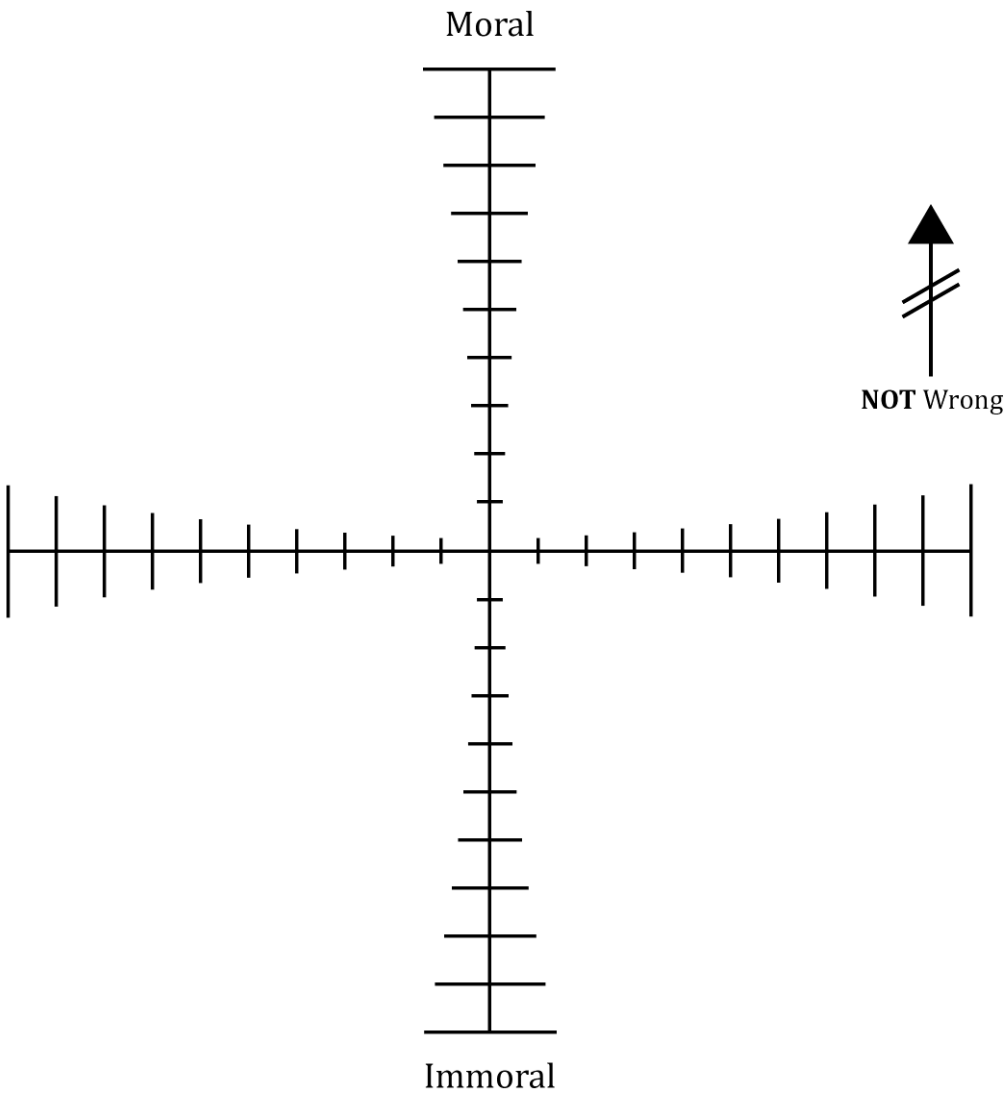


Figure 11.0. The base moral map template.

Having orientated the work surface, moral axes may now be added, although it is worth noting that the orientation of these axes may change, as may the angle between them, depending on the individual, culture, or theory in question. For example, the axes

according to the Theory of Dyadic Morality (Schein & Gray, 2018) are minimally different from those described above. The *y* axis would relate to perceived dyadic harm, whilst the *x* axis may be related to 'norms' or 'negative affect' – more likely the former, although it is not clear how this axis might be labelled, nor how norms might be placed along this axis if it were to remain in this orientation. In contrast, following Constructive Sentimentalism (Prinz, 2009), these axes would be offset from the common orientation, with one axis relating to norms about 'persons' (i.e., autonomy) and the other to norms about the (perceived) 'natural order'. Similarly, these may be considered in graphical terms, such that the extent to which an action violates grounding norms about 'persons' and/or 'the natural order' may be plotted along them. There may be alternative ways to label these axes, but these are illustrative of moral pluralism on approaches which incorporate some notion of 'purity'. However, following Prinz (2009), only these two axes may be considered 'grounded', such that they are akin to 'North' and 'East' in navigational terms – and any other type of norm violation may be derived with regard to these (e.g., northeast).

It is worth briefly noting how derivative axes may factor into the arrangement. Following Moral Foundations Theory (Graham et al., 2013), foundations which would factor under Shweder et al.'s (1997) ethic of community are more closely aligned with sanctity (i.e., divinity) than those which would factor under autonomy (i.e., harm/cheating). Also, following Haidt (2006), 'divinity' would appear to be the derived axis - one's community may be considered as the root source of such concepts. Yet whilst it may be the case that concepts of the divine/sacred may be communally learned, perceptions of 'the natural order' can encompass these considerations without being reduced to them. Indeed, Prinz (2009) defines 'the natural order' in terms closer to Shweder et al.'s (1997) description of the divine order, wherein it is the origin of the sacredness of persons and the ultimate root of all obligations. Prinz's formulation, with 'community' as derivative, also seems to fit better with evidence suggesting such violations may be more closely associated with 'conventional' transgressions than moral

violations relating to ‘autonomy’ or ‘divinity’ (see Chapter 8). Yet however any derived axis might be labelled it can be plotted with respect to the grounding axes, such that it lies somewhere between the two.

The addition of moral axes are shown in sequential Figures, added in stages so as to better illustrate theoretical points. Focussing on immoral arrangements first, Figure 11.1a is illustrative of the Theory of Dyadic Morality (Schein & Gray, 2018), Figure 11.1b depicts Haidt’s (2006) proposed arrangement of Shweder et al.’s (1997) ethical codes, Figure 11.1c shows an alternative arrangement of these codes with regard to Moral Foundations Theory (Graham et al., 2013), whilst Figure 11.1 follows from Constructive Sentimentalism (Prinz, 2009). Similar illustrations can also be drawn on the positive side of the scale, such that Figures 11.2a, 11.2b, 11.2c, and 11.2 depict the positive side of their respective theory. Taking a combination of theoretical positions, the first axis may be labelled ‘A’ for ‘autonomy’, which fits with the approaches discussed. The second axis may be labelled ‘C’ for ‘continuity’, which captures concepts of ‘the natural order’, but also captures ‘norms’, such that the label might be shared by both the Theory of Dyadic Morality and Constructive Sentimentalism. The third (derived) axis may be labelled ‘B’ for ‘balance’, as this axis would lie somewhere between ‘A’ and ‘C’. This may be considered as a balance point between the two, such that its orientation is subject to change - although it is shown in the centre for now. This common labelling is shown in Figure 11.3.

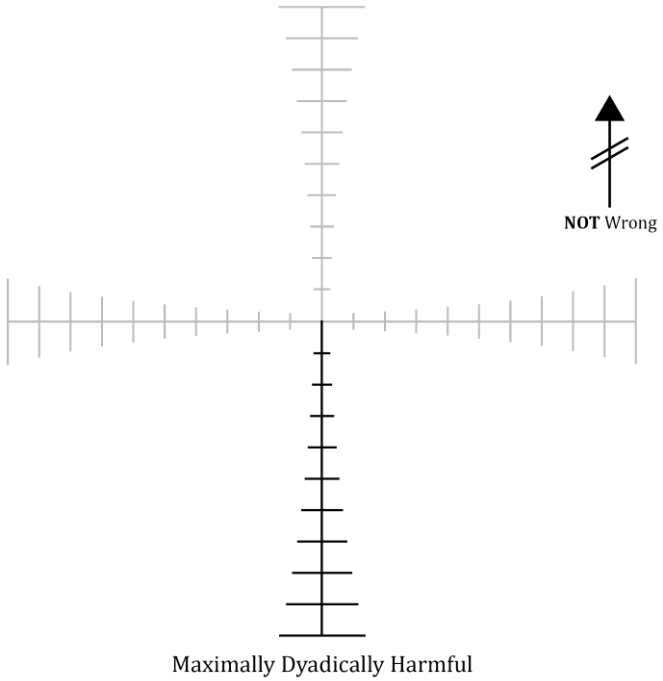


Figure 11.1a. Immorality according to the Theory of Dyadic Morality

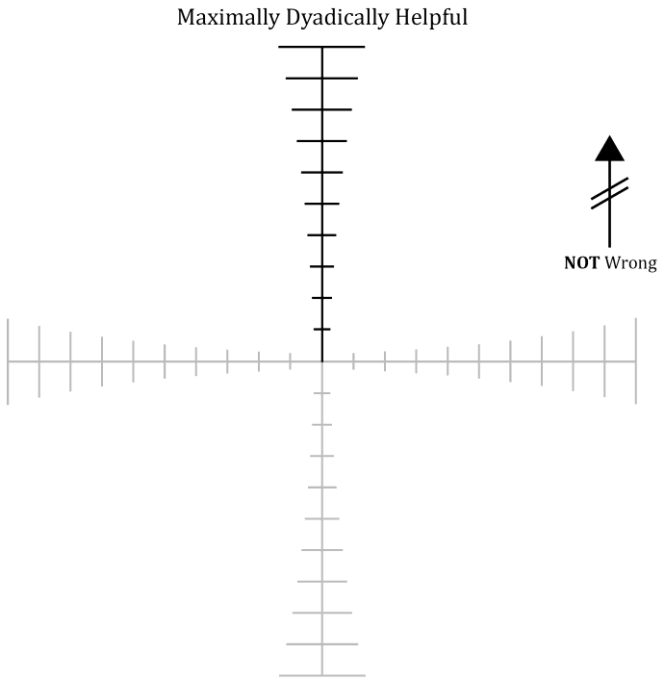


Figure 11.2a. Morally Good according to the Theory of Dyadic Morality

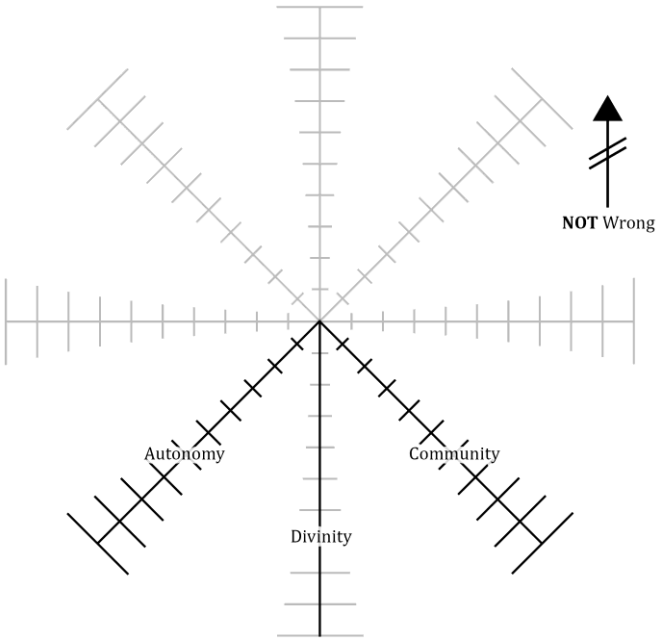


Figure 11.1b. Immorality according to Haidt (2006)

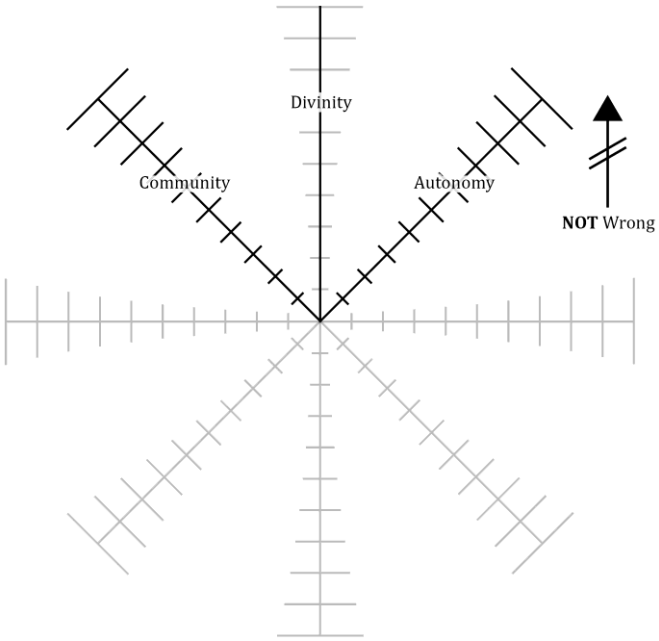


Figure 11.2b. Morally Good according to Haidt (2006)

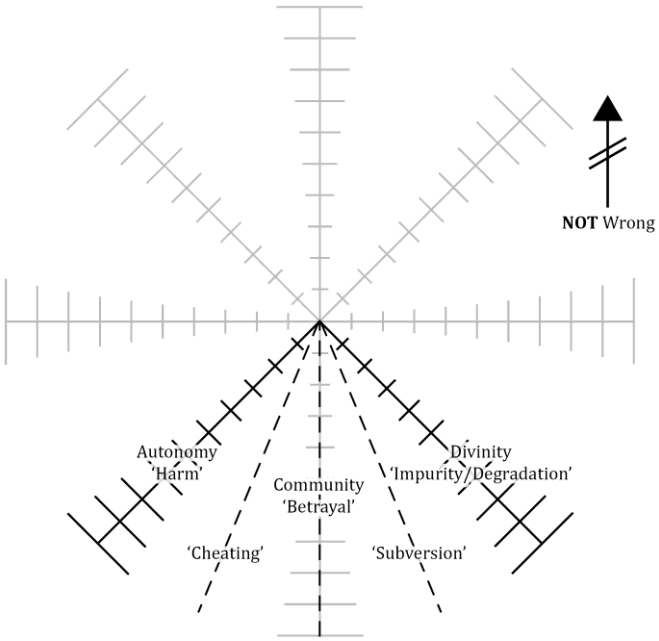


Figure 11.1c. Immorality according to Moral Foundations Theory

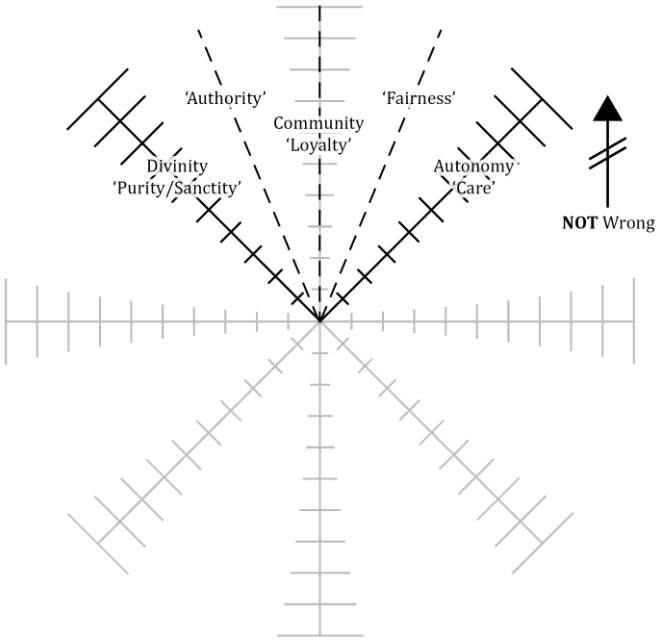


Figure 11.2c. Morally Good according to Moral Foundations Theory

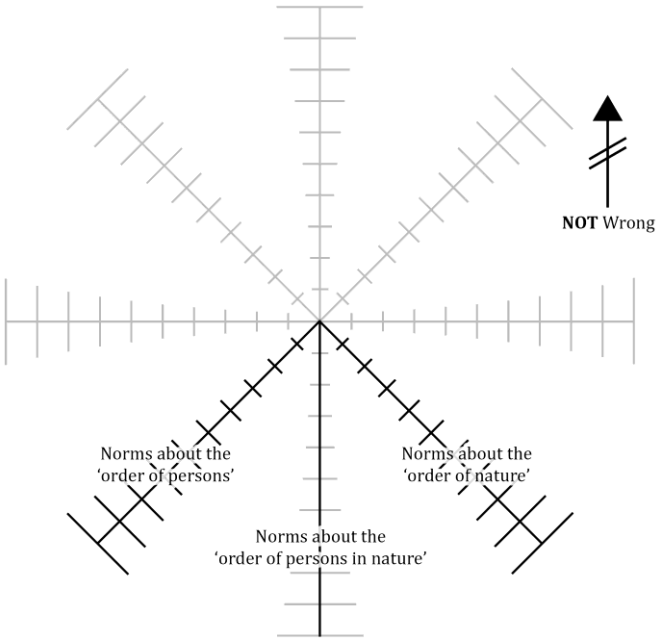


Figure 11.1. Immorality according to Constructive Sentimentalism

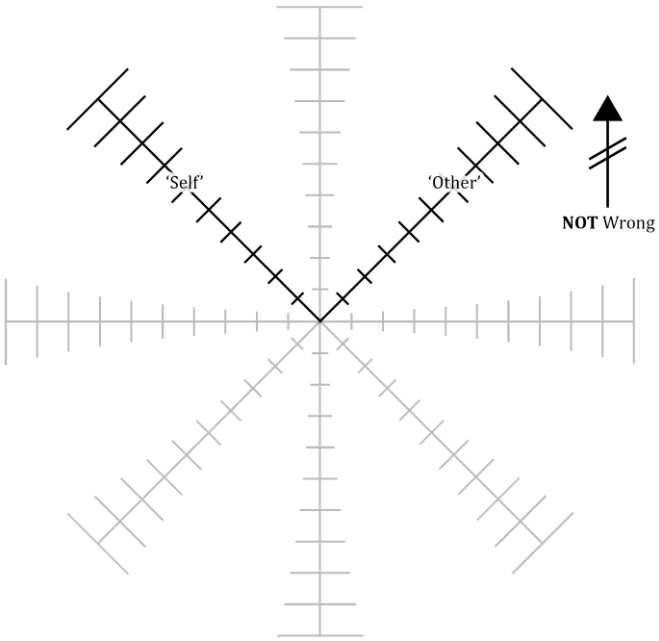


Figure 11.2. Morally Good according to Constructive Sentimentalism

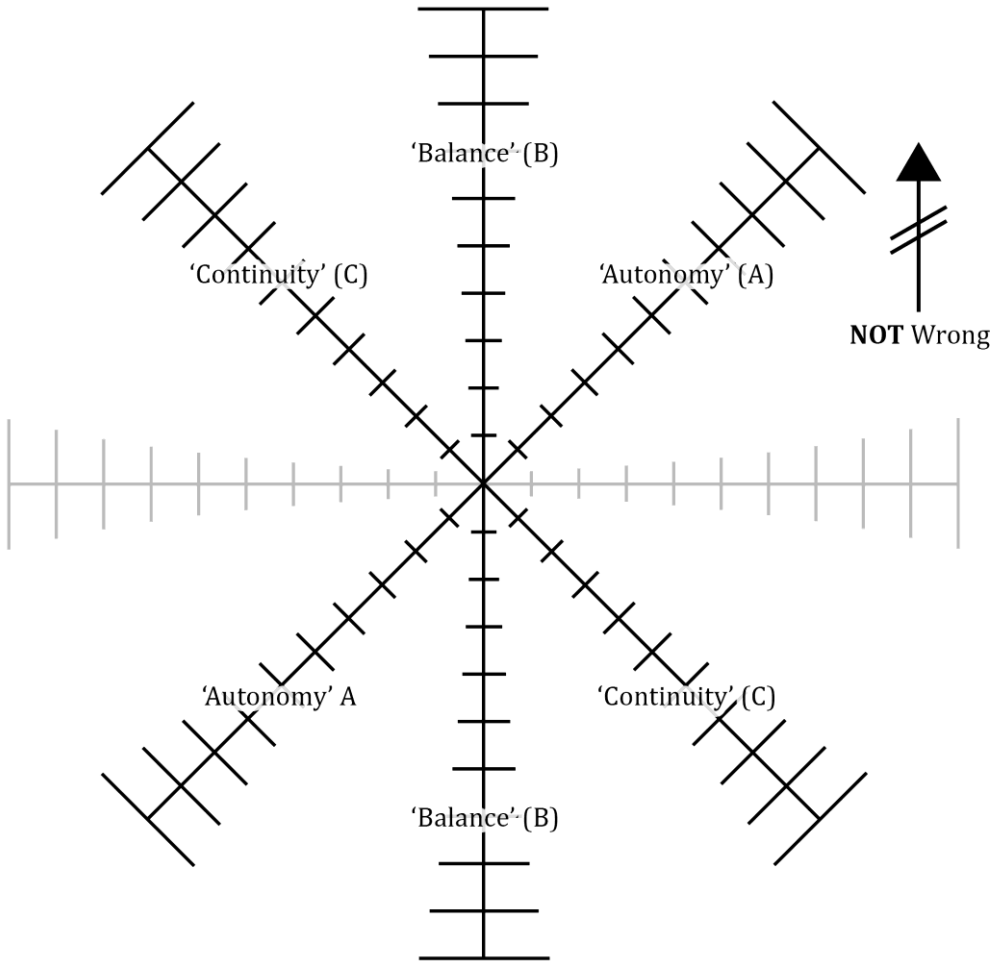


Figure 11.3. Common Theoretical Orientation

However, combining the negative and positive layouts may be problematic for two reasons. Firstly, following Prinz (2009), labels on the positive end (i.e., other and self) seem somewhat different to those on the negative end (i.e., persons and the natural order) - although this concern might be curtailed by noting that 'other' may align with 'persons', and 'self' may align with 'the natural order' (e.g., Dungan et al., 2017). Secondly, maintaining a graphical construction would appear to create issues with plotting locations for pluralist moral theories. Simply extending the axes would seem to provide for actions with highly conflicted moral content. In terms of Moral Foundations Theory, it allows for actions which might be extremely caring, but also extremely

degrading, or extremely harmful actions of extreme sanctity. Actions could score on the positive end of one axis, but the negative end of the other, and although actions fitting either of these descriptions are likely relatively rare, they seem unlikely to be construed as morally neutral – which is where such actions would be plotted if the axes were simply extended. Furthermore, mixed valence actions may be more common within moral domains than between them - an action may be bad for ones' autonomy in one regard, but good in another. Simply extending the axes would seem to prohibit plotting, as this would suggest two points on one axis, rather than one point which can be referenced from two axes.

This second issue may be addressed by *mirroring*, rather than extending the axes, and follows from employing a negative orientation. It is achieved through defining the space above the mid-line as 'not bad', rather than 'good', such that 'not bad' is the reverse image of 'bad', rather than its direct opposite. This may fit better with lay intuitions – actions which are not harmful are not necessarily caring actions, nor do non-degrading actions necessarily involve sanctity. However, interestingly, it may be plausible that mirroring the axes allows for them to be reverse-extended - providing a means to address the first issue. Actions which may be considered in terms of 'positive autonomy' may be considered as positive for the 'self', whereas actions which might factor as reflecting a 'positive natural order' might be considered as good for the 'other' (i.e., for 'persons'). Mirroring may address the potential issues with simply extending the axes, and better allows for the graphical, rather than merely figurative, formulation to be maintained. This mirror effect is shown in Figure 11.3a.

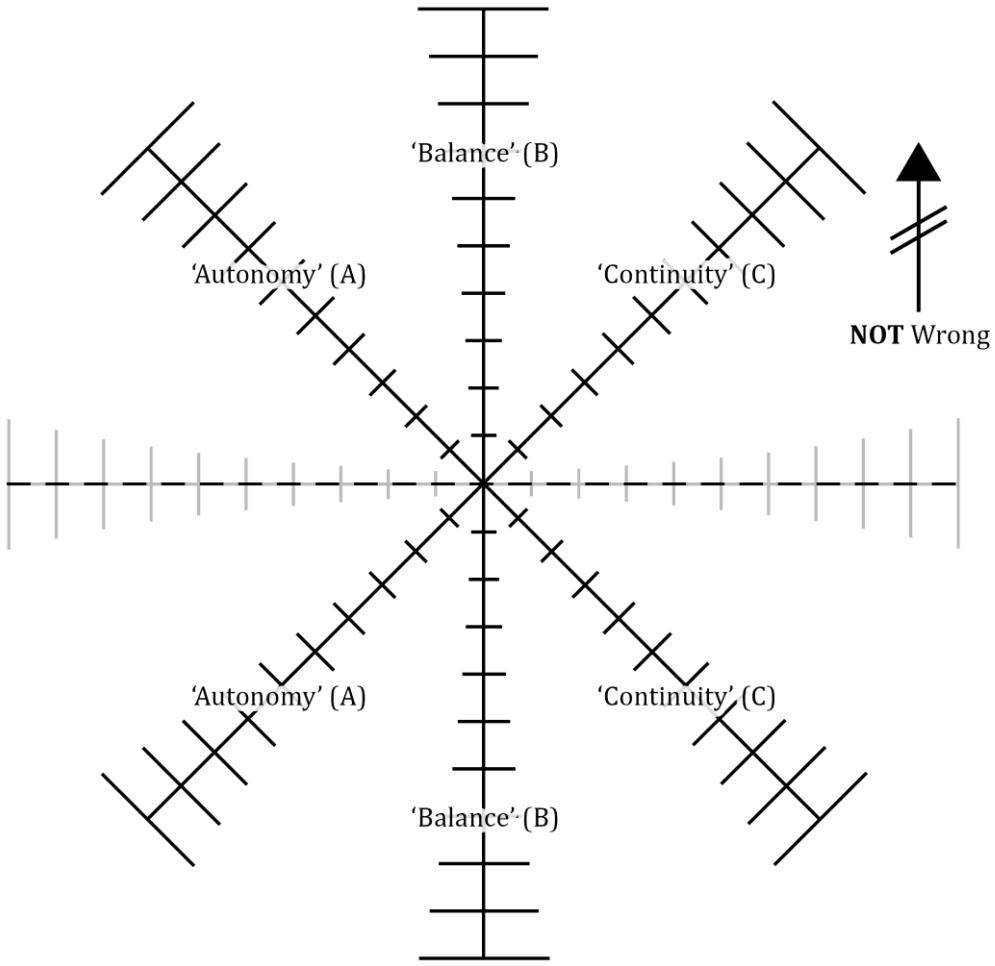


Figure 11.3a. Common Theoretical Orientation with the Mirror Effect

The combination of negative and positive layouts for Constructive Sentimentalism is shown in Figure 11.4, alongside the common labelling to maintain the effect of mirroring. This provides a sketch of a key feature of the map - lines of orientation - which would correspond to north and east in standard navigational terms. This figuratively illustrates key aspects of Constructive Sentimentalism - morality is orientated so as to correspond with 'self' and 'other', and immorality is orientated so as to correspond with norms regarding 'persons' and 'the natural order'. Thus, in both standard and moral terms, orientation corresponds with compass directions.

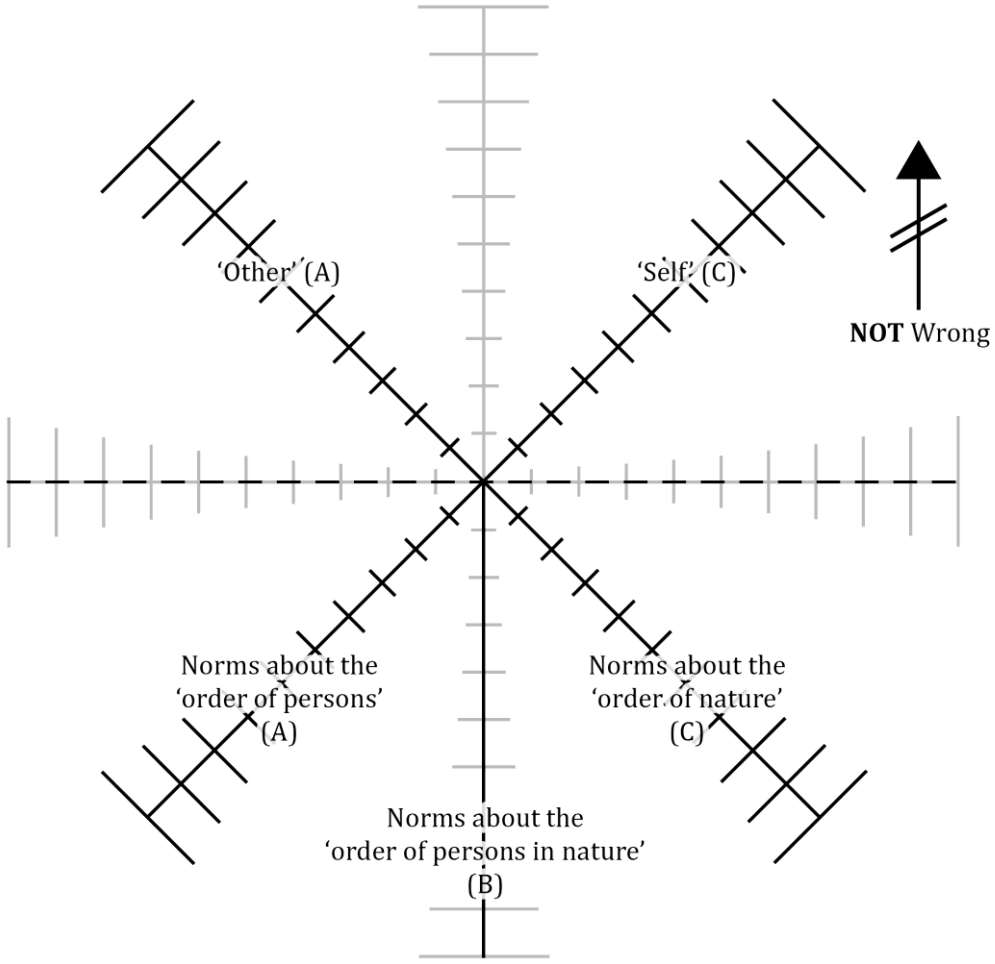


Figure 11.4. Constructive Sentimentalism with the Mirror Effect

The progression of Figures illustrates how the Theory of Dyadic Morality bears remarkable similarities to Constructive Sentimentalism, either through aligning norms to 'purity' (a maximally concessive position), or adding a point of 'self' reference. The axial components detailed also illustrate links with certain moral foundations (e.g., harm, degradation), whilst providing that other moral foundations (e.g., betrayal, subversion) may still appear on the map, or be related to navigation in some way (e.g., by ruling some terrain 'off limits'). In this manner, these may be considered as 'landmarks' rather than foundations, such that they may be marked by symbols on the map, rather than analogized to grid lines. Arguments for whether moral foundations are (massively)

modular mental systems aside, some studies using the Moral Foundations Dictionary (e.g., Graham et al., 2009) may be taken as providing evidence for 'foundational discourses', which fits well with the work of Shweder et al. (1997). "*Discourses are symbol systems for describing aspects of experience*. More than one such symbol system may be applicable to any area of experience, such as individual psychological development, ethics, health, or suffering. *There is no reason that one must select one and only one discourse to represent an area of experience*. Indeed, there may be some advantage in possessing multiple discourses for covering the complexities of such an important area of human experience as ethics." (Shweder et al., 1997, p.140, *emphases mine*).

The mapping analogy also fits with Shweder et al.'s (1997) point regarding cultural differences in the construction of morality. "Indeed, it appears that different cultural traditions try to promote human dignity by specializing in (and perhaps even exaggerating) *different ratios* of moral goods. Consequently, they moralize about the world in somewhat different ways and try to construct the social order as a moral order in somewhat different terms. Cultures *differ in the degree to which one or another of the ethics and corresponding moral 'goods' predominates* in the development of social practices and institutions and in the elaboration of a moral ideology" (p. 141-142, *emphases mine*). This is analogous to stating different cultures may place differing emphases on what they consider important for navigation, and may thus have different navigational (i.e., moral) practices as a result. It captures the notion of ‘cultural learning’ advanced by the theories under consideration, as well as that of ‘pluralism’.

However, the current formulation may also be able to take account of theories which do not (directly) include concerns of ‘harm’ and ‘purity’ (e.g., Janoff-Bulman & Carnes, 2013; Rai & Fiske, 2011), or argue against their status as moral domains (Curry et al., 2019). The duality of axial references suggest that certain moral concerns may be plotted at a single point which falls between both axes. Theories which argue morality is

about relationship regulation (Rai & Fiske, 2011), or cooperation (Curry et al., 2019), would seem to require 'self' and 'other' for relational or cooperative partners – unless one permits of ones' relationship with oneself, or cooperating with your future self, as the Theory of Dyadic Morality (Schein & Gray, 2018) might allow. It may further be possible to factor Janoff-Bulman and Carnes's (2013) Model of Moral Motivation onto the current layout (see Figure 11.5). Their approach broadly follows the help/harm distinction also found in the Theory of Dyadic Morality, but incorporates work from Brewer and Gardner (1996) on social identify and self-representation. On this approach, self/personal concerns would appear in the right quadrant of the map, other/interpersonal concerns would factor in the left quadrant, and group/collective concerns would feature in the top or bottom quadrants dependent on whether they are prescriptive or proscriptive respectively.

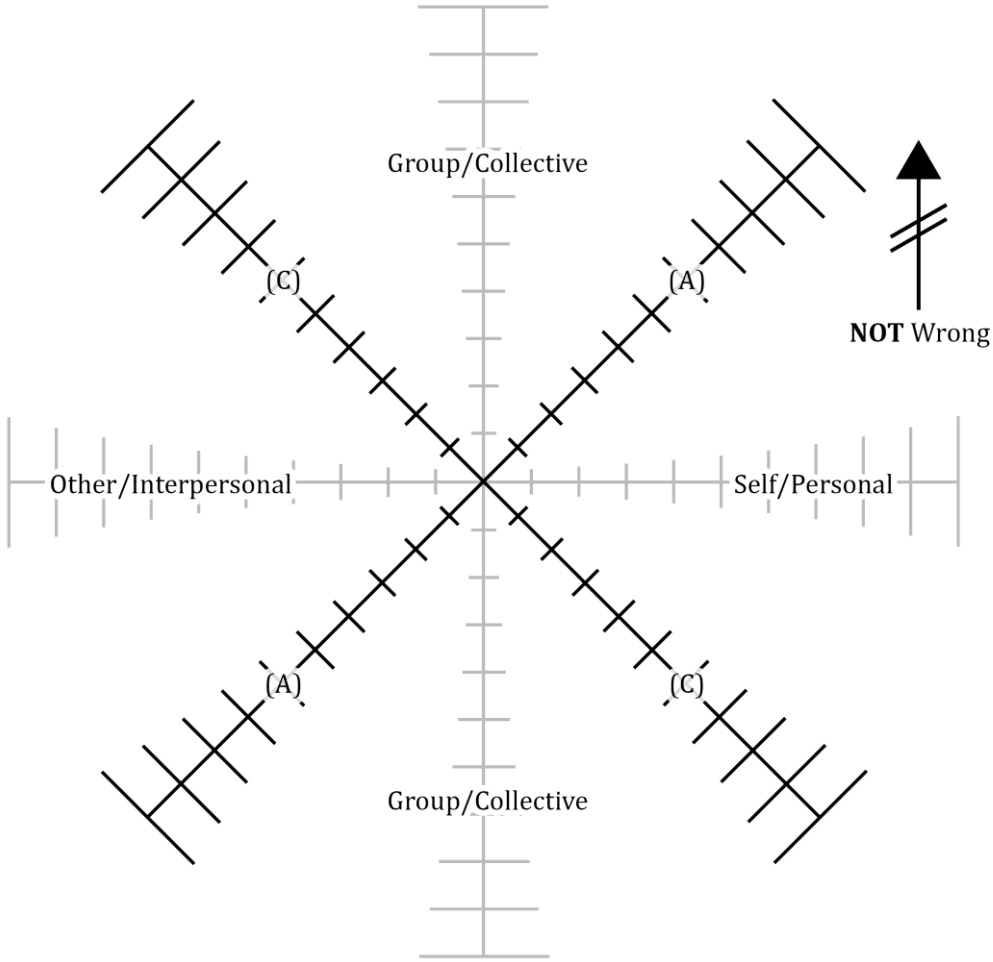


Figure 11.5. Moral Map of The Model of Moral Motives

The analogy to mapping illustrates how multiple theoretical approaches (i.e., maps) might be reconciled within the same space, such that many of their differences may be considered in terms of focus rather than content per se. However, Constructive Sentimentalism (Prinz, 2009) provides a means to get more adventurous with the navigational analogy. If morals are emotionally constructed, as Prinz (2009) argues, then this provides a means by which to orientate the map. Notably, Constructive Sentimentalism is the only approach which readily allows for such orientation through its focus on emotional dispositions, although it does allow parts of other theories to feature. For example, Moral Foundations Theory (Graham et al., 2013) argues in favour of

‘characteristic associations’ between emotions (i.e., anger and disgust) and moral content (i.e., harm and impurity), whereas Constructive Sentimentalism argues in favour of a constitutional relationship. Similarly, although the Theory of Dyadic Morality (Schein & Gray, 2018) argues for different kinds of characteristic associations (e.g., anger and disgust with ‘villains’; Gray & Wegner, 2011), it seems highly favourable towards constitutive appraisal models of emotion – linking these ‘appraisals’ with the ‘conceptual knowledge’ proposed on some constructionist accounts (see Cameron et al., 2015). Indeed, Constructive Sentimentalism draws on exactly this type of model in defining emotions, and provides that intuitions can be defined with regard to sentimental machinery, such that it may provide a better account of moral mechanics than the other theories.

11.3. Constructing the Compass

Pursuing the navigational analogy further, a moral compass may be constructed with reference to emotions. The account of emotions here follows from Embodied Appraisal Theory (Prinz, 2004b), a perceptual theory whereby emotions are defined as ‘perceptions of the organism-environment relationship which bear on well-being’. However, any reasonably similar account may suffice, as Embodied Appraisal Theory can be considered independently from Constructive Sentimentalism. The basic claim advanced by Prinz (2004b) is that cognitive theories are correct about the content of emotions (i.e., appraisals) whilst somatic theories of emotion (e.g., James-Lange) are correct about the form of emotions. Prinz's approach provides for basic emotions, or families of emotion, as there seems to be a limited number of discernible somatic states in comparison to the range and number of (human) emotions. However, it also allows for basic emotions to blend into ‘new’ emotions, and permits emotions to be distinguished with reference to their conceptual content, such that complex culturally-calibrated emotions may be constructed from innate emotional components. The perceptual

approach considers emotions as arational, such that the experiencing of an emotion is itself neither rational nor irrational – it is simply the result of perceptual machinery in operation.

The outline of the compass can be constructed to overlay the map by encircling the map from a default (or resting) position at the centre. This circle may be considered as representing the range of possible emotional experiences, such that extreme emotional experiences may be plotted near the outer ring, whilst relatively mundane everyday experiences may be plotted closely around the centre point. This circle also fits better with the graphical formulation of the map, as it provides a means by which the axes can be read by radial rather than linear measures. It permits that actions which are maximally wrong in terms of both violating norms about persons and norms about the natural order would still be plotted as maximally immoral. Furthermore, a circle has the benefit of being readily relatable to 'affective circumplex' models of emotion (e.g., Feldman-Barrett & Bliss-Moreau, 2009), and arguments that emotion and morality are better described with reference to circles rather than arrows (Gray, Schein & Cameron, 2017).

However, just as the angle of axes may vary depending on the individual, so too may the shape of the compass. For example, past experience of trauma might redraw the circle in various ways, such as making it more ovular, or reducing the curve on the upper half (i.e. reduced positive affective experience); depression or numbness might be considered as reducing the circle radius (i.e., flattened affective experience); and the 'circle' for individuals with certain traits (e.g., psychopathy) might be depicted as being closer to a pie chart or crescent moon shape (i.e., minimal potential for certain negative affective experiences such as fear). There may also be individual and cultural level variation with regard to orientation between map and compass, allowing for cross-cultural differences in emotion to be factored into account. The aim in what follows is to provide a sketch of how a moral compass might operate with regard to emotion. This

takes a broad approach to combining various theoretical elements within the same space as a means of demonstrating utility, rather than seeking to outline and defend a more technical illustration.

Continuing the progression of Figures, the approaches to emotion used by the Theory of Dyadic Morality and Constructive Sentimentalism provide different labels, and potentially different means of operation, for parts of the compass. For example, following arguments for a role of 'core affect' advanced by Cameron et al. (2015), the vertical axis might be considered in terms of 'valence', and the horizontal axis in terms of 'arousal' - although the relative direction of travel along this axis may be important for reconciling these labels with the 'affective circumplex' model. Travel towards the edge of the circle can be considered in terms of activation, and travel towards the centre in terms of deactivation (cf. Feldman-Barrett & Bliss-Moreau, 2009). Additionally, the proposed arrangement has the advantage of allowing more distance between emotions of anger and disgust, which often feature closely together in maps of affective space given they are both considered negatively valenced, high arousal emotions. It permits that important differences between these emotions, such as whether they relate to approach or avoidance, may be taken into account. For simplicity, this behavioural tendency may be plotted as a third axis, aligned with 'C', such that approach orientated emotions are those above line 'A', and avoidance orientated emotions are those below 'A'. This labelling is shown in Figure 11.6.

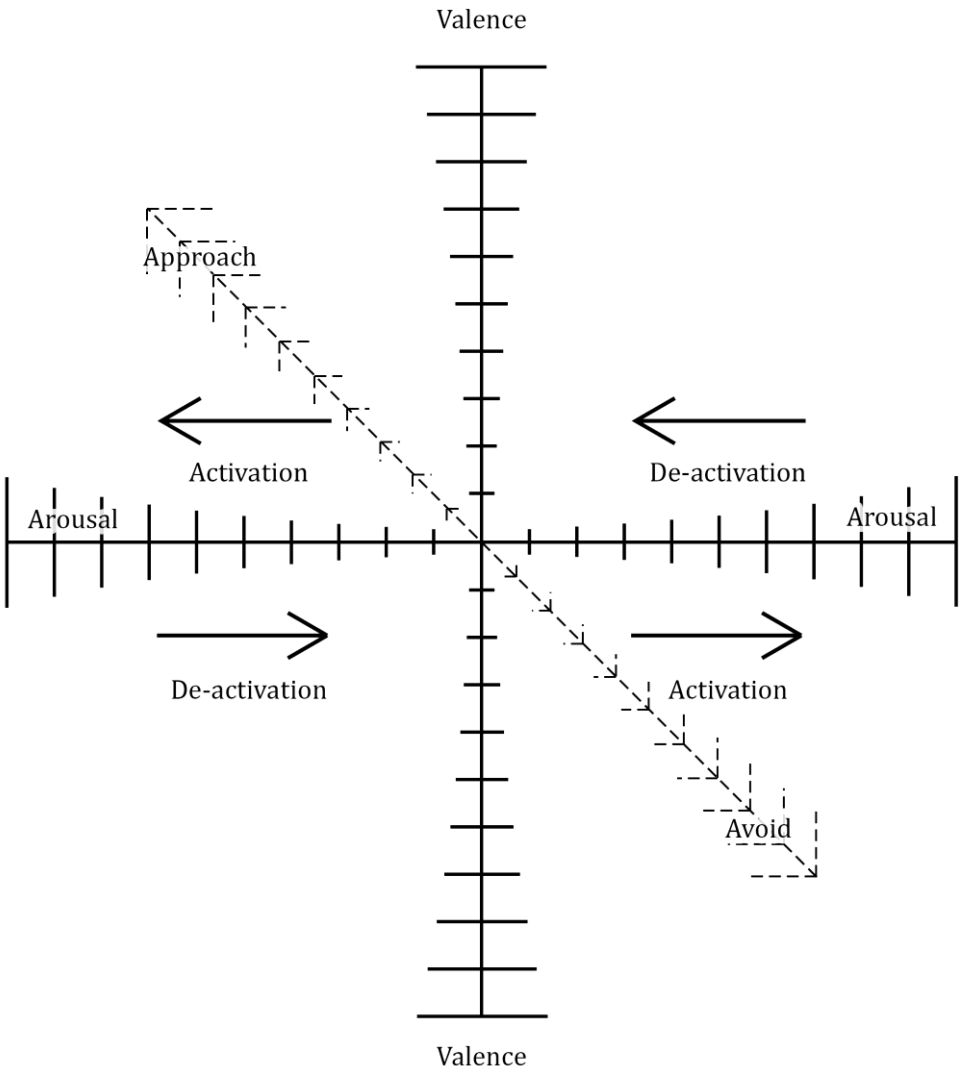


Figure 11.6. A compass from basic emotion dimensions

This latter point also functions to reconcile the approach from 'core affect' with the sketch that arises from following Embodied Appraisal Theory (Prinz, 2004b). Taking a basic familial approach to emotions, each emotion family may be considered as filling in part of the compass. For the sake of argument, suppose anger, disgust, fear, happiness, sadness, and surprise may be used to fill in the compass. Happiness and sadness would provide an opposing pair in terms of valence, anger and fear may provide another in terms of opposing arousal dimensions (i.e., fight or flight), leaving disgust to pair with surprise - and in need of reformulation as the valence of surprise may vary (e.g., one may be amazed or shocked). Focusing on positively valenced states

so as to maintain balance, surprise may instead be replaced by 'awe' or ‘wonder’ - a familial category which would include 'elevation' as an opposite to disgust. This provides two emotions with uniform positive valence (i.e., happiness and awe), two with uniformly negative valence (i.e., sadness and disgust), and two emotions which may have variable valence potential (i.e., anger and fear). Allowing equal space for each of these emotion families, and maintaining the concept of mirroring used when combining positive and negative axes on the map, these emotions would factor onto the compass as shown in Figure 11.6a - with the map re-orientated underneath so as to align anger with autonomy violations and disgust with violations of the perceived natural order in Figure 11.7.

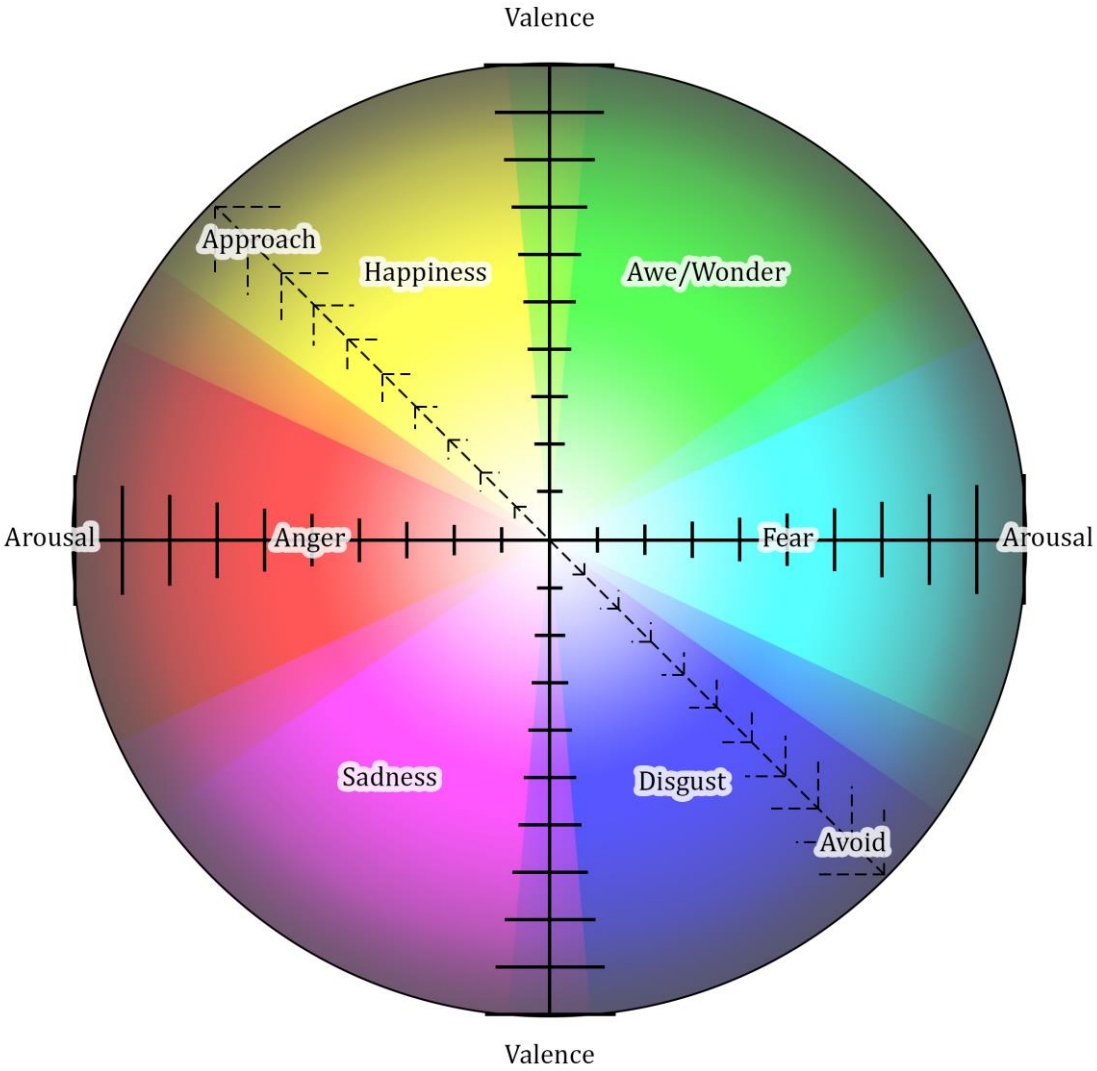


Figure 11.6a. A compass of the 'basic' emotion families

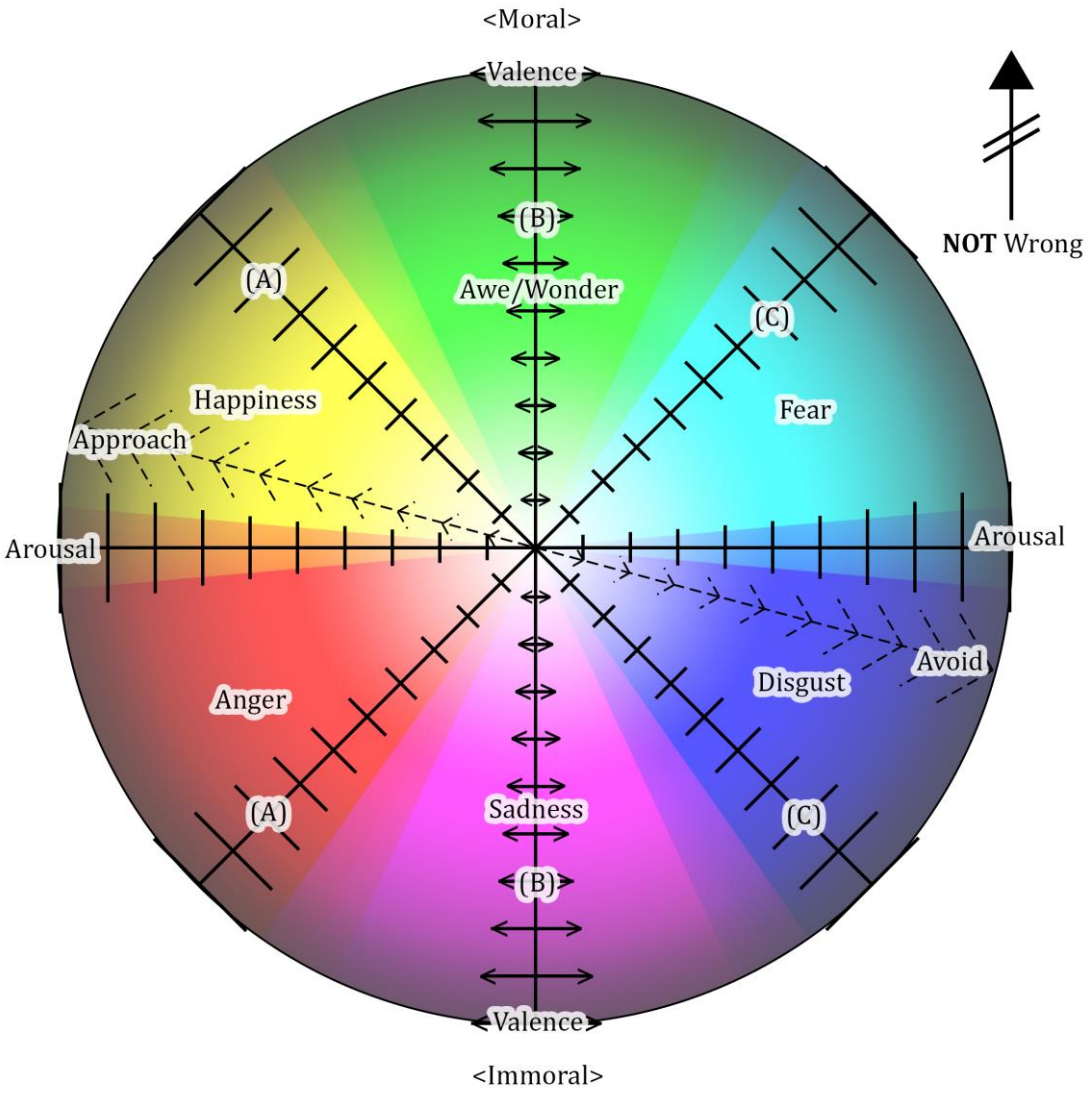


Figure 11.7. A common moral map and emotion-based compass

Having orientated both map and compass with regard to morality and emotion, the resultant image seems remarkably in keeping with tests of Constructive Sentimentalism's proposed patterns of emotion elicitation. The frequent elicitation of disappointment in response to moral transgressions might be explained by this emotion being a member of the sadness family located between both moral axes - its presence would almost seem expected on this formulation. Similarly, emotions of admiration, gratitude, and other such positive moral emotions may be considered as members of the awe family, which is also located between both axes when morally orientated. Areas between axes relating to emotion families of happiness and fear may also contribute in

this regard. For example, pride may feature as part of the happiness family on the autonomy side, whereas some kind of threat-based awe (Gordon et al., 2017), certain notions of 'god-fearing', or potentially ‘bravery’ (in the sense of overcoming fear), may feature as being related to the natural order. This layout also serves to capture Constructive Sentimentalism's argument that both the moral domain and the actor-observer relationship matter for immorality, whereas only the latter matters for morality. This is illustrated via colours, such that anger appears red, disgust appears blue, and wonder appears green. The frequent co-occurrence of anger and disgust (i.e., various shades of purple) in response to moral wrongs, compared to the relatively uniform emotion elicitation patterns found in response to positive moral actions (i.e., green), fits well with this formulation.

11.4. Future research

This is as far as the navigational analogy might be safely advanced based on the material covered. However, there are a few further points of relevance worth noting, as these suggest ways in which the analogy may be extended further. For example, Constructive Sentimentalism also provides means by which the compass may self-calibrate via meta-sentiments. If sentiments govern norms relating to behaviour, then meta-sentiments govern norms relating to emotion. The former relates to whether an observer has possession of the requisite emotional disposition to (dis)approbative emotions, the latter as to whether the observer deems it appropriate to experience emotions in the (dis)approbation range. One may feel anger at injustice (a sentiment), and also feel that experiencing anger in response to injustice is warranted (a meta-sentiment), or one may feel angry (or disgusted) towards individuals’ who do not appear to be in possession of this sentiment (i.e., feeling disapprobative emotions towards someone who does not respond to injustice with disapprobation). Meta-sentiments may also impinge on suggestions that ‘harm’ (and anger) is act directed whereas ‘purity’ (and

disgust) is person-directed (Dungan et al., 2017) – the former may be drawing on sentiments, and the latter on meta-sentiments, although these 'levels' are rarely distinguished methodologically.

However, further research is required to establish how the compass may work in this regard. For example, any emotion-domain associations may reverse at the level of meta-sentiments, such that (meta-sentimental) disgust acts to reinforce (sentimental) anger and (meta-sentimental) anger reinforces (sentimental) disgust – which may also provide some explanation of frequent anger-disgust co-occurrence. Alternatively, there may be no reversal, such that each emotion reinforces itself at the meta-level, although this may make it harder to investigate experimentally if the same response is expected at both levels. In this regard, future research may find it useful to examine the 'directedness' of emotional responses – what the emotion is primarily aimed towards/what is considered to be the source of the elicited emotion – alongside any other measure of emotion, and more nuanced questioning methods which allow for sentiments and meta-sentiments to be discernible from responses given to moral scenarios.

Future research on emotions and morality with regard to moral dumbfounding may also benefit from development. Previous research has worked either with pre-set scenarios which have been finessed to make them 'harmless' (i.e., they are carefully worded impurity scenarios), or with moral dilemma scenarios phrased to have enacted the less popular choice option (see McHugh et al., 2017). There appears to be a gap in the literature for studies which better pursue iterative questioning in response to simpler, or self-generated, examples of morally loaded events. Several additional measures may also be of interest in such a paradigm, such as whether the type of reasoning changes between iterations (e.g., harm based or norm based), and how many iterations are needed for any reasoning to 'bottom out'. It would also be of interest to examine whether potentially purity-specific emotional responses, such as smiling and laughing (cf.

Franchin et al., 2019), are common to dumbfounded responses (as suggested by McHugh et al., 2017), or whether this follows from the majority of example scenarios in dumbfounding research being focused on violations of purity.

Continuing research on disgust may also assist in teasing apart what this emotion actually contributes to morality. Piazza et al.'s (2018) review of 'what disgust does and does not do for moral cognition' highlights many concerns regarding research in this area, and suggests there is little evidence for a unique role of 'core' (physical/pathogen) disgust in morality. Yet Piazza et al. are not prepared to jettison a role for disgust, suggesting that the concept of 'disgustingness' - being aware that others may construe an act as disgusting - may be an 'organizing property of wrongdoing'. This fits reasonably well with the navigation analogy, as it maintains a role for disgust (in addition to anger) as being axially aligned. However, the analogy also seems to fit in addressing concerns regarding disgust sensitivity advanced by Landy and Piazza (2019). That general emotion sensitivity seems to correlate with more extreme judgements is merely a feature of the design, but that disgust sensitivity appears to impact the extremity of aesthetic (i.e., non-normative) judgements may appear to be (loosely) predicted on this formulation. Taking the orientation of map and compass in their default (i.e., non-normative) arrangements, the 'wonder' family of emotions – associated with aesthetics (Prinz, 2007b), rests on the mirrored axis of 'the natural order' – associated with disgust. Disgust sensitivity may be related to aesthetic judgements via the associations between emotions and moral axes.

Further research which tests claims of a constituent role of emotions in moral judgements would also help inform debate. Tracy et al.'s (2019) research showing that inhibiting the 'machinery' of disgust via the use of an anti-emetic resulted in less severe moral judgements poses problems for Piazza et al.'s (2018) argument for 'disgustingness'. Replication of Tracy et al.'s research to include a wider range of scenarios, which would include scenarios varying in terms of the type and 'strength' of

disgust they elicit (e.g., pathogen, sexual, moral), may be informative in this regard. Pairing such replication with the approach of Oaten et al. (2018), which suggests moral disgust may be closer to anger at the neurological level, may also help further distinguish which facets of disgust might be responsible for the suppression effect shown by Tracy et al. (2019). Pharmacological suppression of other emotional components, particularly those related to anger or fear, may also be informative with regard to constitution approaches to morality - although such studies are likely more challenging to design and administer, as the pharmacological intervention would need to inhibit specific components of certain emotions rather than providing for inhibition of affective experience more generally.

Future research may also benefit from more detailed consideration of the variety of actors and roles available with regard to moral scenarios. In particular, the role of the participant tends to be observational, such that they are generally responding to scenarios depicting an agent performing some misdeed or other action, which may (or may not) cause *someone else* (or something else) to suffer (in some way). In what seems to be the majority of cases, neither the agent nor victim of these misdeeds is the participant - responses are based on 'observations' of the moral actions of others. This is not necessarily an issue for most ratings, and may even help curtail ceiling effects in wrongness ratings, although it is worth noting that such responses may be markedly different across contexts. For example, phrasing the materials in a way whereby the participant occupies the role of victim (or agent) may increase (or decrease) ratings on a variety of dimensions (e.g., wrongness), and would likely also affect any emotional response elicited via this change in context. Moral violations directed towards oneself (hypothetically or otherwise) may elicit a stronger emotional response than those directed towards hypothetical others and, unlike in the observational cases, there is no additional (often implicit) requirement that one must have some kind of empathy/sympathy/concern for the victim. Moral violations directed towards oneself also appear less open to possible concerns regarding socially desirable responding than

when merely 'observing' others' misdeeds. Following Constructive Sentimentalism, research which examines misdeeds committed by 'loved ones' or 'sacred others' (e.g., ones' children, parents, or spouse) may also be of interest, as would research which adds a responding-observer into scenarios in order to better assess meta-sentiments. For example, 'observer' does not think 'agent' did anything wrong to 'victim' - how do you feel towards 'observer'?

More adventurous research paradigms might examine the extent to which the navigational analogy is defensible in graphical (i.e., measurable) terms, rather than merely illustrative terms. The metaphor may be mechanised. All axial components mentioned for both morality (e.g., whichever theoretical version of 'autonomy' and 'continuity' axes is preferred) and emotion (e.g., arousal, valence) may be considered as graphical axes. Moral graph points may be plotted by 'wrongness', and emotional graph points may be plotted with regard to intensity/salience, in a radar-style manner. However, the arousal axis would necessarily lack a (central) zero point to reflect a constant stream of affective experience. This might be taken into account via arguments that arousal and valence are related in a dynamic non-linear fashion (Noel, Fevrier & Deflandre, 2018), such that the zero point is avoided naturally; although it may be better to consider the arousal axis as relating to relative (rather than absolute) change in arousal levels, such that the zero point may be construed as reflecting a point of homeostatic equilibrium. This latter suggestion has the benefit of being able to capture activation (away from equilibrium) and de-activation (towards equilibrium) whilst providing room in the 'affective circumplex' (Feldman-Barrett & Bliss-Moreau, 2009) to allow for emotions to be distinguished in (behavioural) terms of approach/avoidance. Most importantly, further research establishing the degree of common orientation between the axes of the map and those of the compass (i.e., the extent to which moral and emotional 'plot points' coincide) may provide a means of connecting with broader themes within moral psychology.

11.5. The potential of Constructive Sentimentalism

At the metaphorical level, in addition to illustrating the emphases of different moral theories, the navigational analogy also provides a means of organizing the main research themes in moral psychology identified by Ellemers et al. (2019). Moral reasoning, moral judgements, and moral emotions may be organized as relating to aspects of map and compass. In broad terms, emotions relate to the construction of the compass, judgements relate to the marking of locations on the map - which makes use of the compass, and reasoning relates to justifying why a given location (i.e., judgement) has been plotted at a particular point. Emotions also provide a means of connecting with moral behaviour given links with affordances. Furthermore, as the map has been defined with reference to 'worldview', the theme of moral self-views may be considered in terms relating to the view from within the map, how one is positioned with regard to important features of the map, or in more general terms as relating to the kind of movements available, or previously conducted, within this space. In this regard, it is worth briefly noting that morality and 'the self' may be linked through concepts of identity.

Prinz and Nichols (2016) "defend the thesis that moral continuity (i.e., retaining the same moral values over time) is central to ordinary beliefs about what makes someone qualify as the same person as they advance through life. In fact, moral continuity is more important, according to our ordinary understanding, than memory, narrative, or agency. Each contributes to our sense of identity over time, but moral continuity contributes appreciably more." (p.449). Empirical support for this claim is available via Strohminger and Nichols (2014), with further support also suggested via recent research. Heiphetz et al. (2018) report similar links between morality and identity, showing that changes to widely shared moral beliefs are associated with greater changes in perceived identity than are changes to controversial moral beliefs; and that adults report identity changes as greater for negative shifts in belief (i.e., good to bad) than positive shifts (i.e., bad to good). Similar support for links between morality and

identity is available via Lefebvre and Krettenauer (2020), who also report changes to moral beliefs (especially positive to negative shifts) are most associated with changes to identity - and that moral beliefs tend to have a closer relationship with identity in adulthood. Additionally, Han et al. (2019) report that the emergence of 'moral identity predicts the development of presence of meaning during emerging adulthood', suggesting that the developmental establishment of associations between morality and identity is predictive of 'beliefs that ones' existence has meaning, value, and purpose'. Furthermore, Chen et al. (2018) suggests that making decisions which are reflective of ones' 'true' self may be associated with greater decision satisfaction in response to moral dilemmas.

Consideration of links between emotions and behaviour (i.e., action) brings the thesis full circle to return to the connective potential of Constructive Sentimentalism (Prinz, 2009). Part of this promise, whereby considering emotions as a form of perception (Prinz, 2004b) would seem to beg for connections with affordances, has been fulfilled during the production of this thesis. Prinz now endorses an enactivist theory of emotional content (Shargel & Prinz, 2018), whereby the affordances associated with emotion are state-dependent and imperatival. This updates embodied appraisal theory (Prinz, 2004b) such that these affordances are what emotions 'represent' - although the update permits that state-dependent imperatival affordances may qualify as 'core relational themes' if these themes are broadly defined. The enactivist update also fits well within the navigational analogy developed in this thesis. It is suggestively supportive of distinguishing between emotional vectors (i.e., approach versus avoidance, activation versus deactivation), provides a 'to be doneness' with regard to emotional content, and allows for greater flexibility with regard to relations between map (i.e., concepts) and compass (i.e., emotions). Most importantly, the connection drawn between emotions and ontology may provide the origin of the 'moral matrix', and the means by which this 'map' may overlay reality. Emotions provide a means to "furnish the world with normative properties" (Shargel & Prinz, 2018, p. 129).

However, the connective potential on offer from Prinzean approaches may be illustrated further by outlining potential dimensions into which the analogy may expand. Thus far, the formulation of the analogy has remained within two-dimensional space, such that navigation relies on five points of reference - two points connecting the *x*-axis (i.e., arousal), and two points connecting the *y*-axis (i.e., valence), from which one can locate one's target destination (i.e., intensity) with regard to ones' point of origin. However, a fully developed analogy would need to detail navigation in a three-dimensional space - requiring two further points of reference to connect the *z*-axis, whilst operating in four dimensions so as to account for emotional dynamics and change over time. Following Constructive Sentimentalism (Prinz, 2009), the most likely dynamic for movement along the *z*-axis relates to whether the process is reactive (bottom-up movement) or reflective (top-down movement). This would allow reflective emotions of guilt and shame to feature in their hypothesized positions in line with anger and disgust respectively. The *z*-axis itself, as implied during the formulation of map and compass, may be considered as relating to levels of perceptual processes in line with Prinz's theory of consciousness (Prinz, 2012). This allows that the 'plot point' can be analogized to conscious experience, such that it provides a navigator within the analogy. The addition of a *z*-axis might be taken as transforming the two-dimensional circle into a three-dimensional sphere, whereby adding rotation (to account for dynamics) changes this into a torus shape which may be taken as depicting the individuals' sphere of consciousness. It is notable that this shape fits well with Prinz's contention that both higher and lower perceptual processes - the areas to the top and bottom of the torus with regard to the *z*-axis - are inaccessible to conscious experience.

Research on the neuroevolutionary origins of human emotions also seems relatively compatible with both the formulation of the compass, and the navigational analogy more generally. However, the connection here is highly speculative, so the illustration is accordingly brief and without background - the aim is to show that similar

approaches can be analogized in a similar manner. Panksepp and Biven (2012) argue in favour of seven emotional systems, and hypothesize a 'core self'. Six of the emotional systems they describe may be taken as providing axial reference points. The 'care', 'lust' and 'play' systems may map to the (respective) positive axes of A, B, and C; and the 'rage', 'grief' and 'fear' systems may map to the (respective) negative axes of A, B, and C (cf. Figure 11.4.) - although whether (and how) these axial references may be orientated with regard to the *x, y,* and *z* dimensions proposed is an open question. The seventh emotional system, 'seeking', provides for a direction of travel, and the 'core self' may act as ones point of origin, such that Panksepp and Biven's (2012) approach may be taken as providing all the relevant points of reference for three-dimensional navigation in a manner compatible with the formulation advanced here. Furthermore, in considering that different emotional systems may offer different contributions to experience, be sensitized differently through experience, and may usefully inform therapeutic practices, Panksepp and Biven (2012)'s approach may allow for further development of the navigational analogy. For example, variations in the sensitivity of different emotional systems may make it relatively easier for individuals to 'move' in some directions than others; and such variations may also lead to some areas of the map being better charted, or more frequently traversed, than other regions. Fully exploring this avenue may likely require a further thesis, but there appear to be some commonalities between arguments advanced by Panksepp and Biven (2012) and Prinz's (2012) account of consciousness which suggest such an undertaking may be worthwhile.

The navigational analogy derived from Constructive Sentimentalism may fit particularly well with therapeutic approaches given its person-centric formulation and professed ability to derive 'oughts' and obligations in an empirical manner. If one ought to do what it would be wrong not to do, then this may provide a means of determining ones' general direction of travel by exploring ones' sentimental commitments. Similarly, reversing terms, such that one ought not do what it would be wrong to do, is analogous to the statement “not that way!” - and it may be easier to work out which way(s) not to

go. The connection with therapeutic approaches may also be related back to Shweder et al.'s (1997) research, which suggests explanations of suffering tend to be morally imbued. Following their analogy, whereby the development of moral virtue is considered akin to a form of preventative medicine, establishing what one ought and ought not do (in accordance with ones' own values) and acting in line with ones (self-imposed) obligations may serve to minimize perceptions that oneself is the cause of ones' (future) suffering. Linking to the existing literature, such an approach may be considered as a form of inoculation for 'characteristic self-blame' (as distinguished from behavioural self-blame by Janoff-Bulman, 1979) - measures of which converge with measures of shame (Tilghman-Obsorne et al., 2008). Similarly, Prentice et al. (2019) suggests that being able to perceive oneself as morally good (i.e., not 'sick' or 'ill') fits the criteria for being considered as a basic psychological need, providing connections with well-being, flourishing, and positive psychology more generally.

In similar (brief) regard, the navigational analogy may be readily connected to literature on personal change and post-traumatic growth. Janoff-Bulman and Schwartzberg's (1991) general model of personal change outlines four processes in common to the experience. Confrontation - one recognises a feature of the map is not in quite the right place. Resistance - one does not wish to change the map. Validation - accepting the feature is in fact in place. Integration - redrawing the map to incorporate the new feature. Similarly, Janoff-Bulman (2004) proposes three models of post-traumatic growth. Strength through suffering - growth resulting from having survived the 'psychological earthquake' of trauma. Psychological preparedness - having coped with such events, one may have acquired various 'equipment' which facilitates such navigation, it provides a 'packing list' for future travel and incorporates adverse conditions into ones 'assumptive world'. Existential re-evaluation - a newfound appreciation of ones' landscape following the 'psychological earthquake'. Connections may be further established with Structural Existential Analysis (van Deurzen et al., 2014), which also draws on notions of an emotional compass and on many concepts

relevant to the navigational analogy developed here. In particular, analogies linking 'mood' with 'weather', ideas of there being tension between values at many levels, and the emphasis on emotions in orientating to values, provide a means of linking the navigational analogy with existential approaches to therapy.

11.6. Concluding Summary

Constructive Sentimentalism (Prinz, 2009) draws on independently motivated accounts of concepts (Prinz, 2004a), and emotions (Prinz, 2004b; Shargel & Prinz, 2018), in support of an account of moral judgement that has strong parallels with Hume's (1751) approach. This alone may be sufficient to recommend it to many moral psychologists, especially as Prinz goes further than Hume by detailing how a normative conclusion (an 'ought') might be derived from descriptive premises (an 'is') by reference to emotion. However, there are further reasons to advocate for greater attention to be given to both Constructive Sentimentalism and Prinzean philosophy more broadly - including that one may accept only certain parts of Prinz's approach, and still use it to improve on other theories.

Constructive Sentimentalism has been used to show that the Theory of Dyadic Morality (Schein & Gray, 2018) cannot survive in its current formulation, as the most relevant dyad for moral judgement is ‘self-other’ rather than ‘agent-patient’. However, Constructive Sentimentalism otherwise grants or supports many of the claims made by the Theory of Dyadic Morality; and these theories may be readily reconciled should the latter come to accommodate ‘purity’, which the former argues is a more parsimonious position. Constructive Sentimentalism also fits well with Moral Foundations Theory (Graham et al., 2013), although it suggests some 'foundations' (i.e., loyalty, authority) may not be as 'foundational' (or relevant) as others, such that these may be derived from more fundamental 'grounding norms' concerning persons or the natural order.

Additionally, Constructive Sentimentalism's focus on both moral domains and the actor-observer relationship allows it to incorporate approaches focusing on relational context (e.g., Rai & Fiske, 2011, Janoff-Bulman and Carnes, 2013), giving it an edge over Moral Foundations Theory. Furthermore, it may be taken to compliment arguments for Morality-as-Cooperation (Curry et al., 2019), such that Constructive Sentimentalism provides the psychological components (i.e., 'in here') from which anthropologically established 'cooperative landmarks' might be located (i.e., 'out there'). Importantly, all these claims may be advanced without recourse to emotion, such that Constructive Sentimentalism may offer a better account of moral content regardless of associations with emotion. However, in linking intuition to sentimental (i.e., emotional) machinery, Constructive Sentimentalism offers an account of the genesis of moral intuitions which neither the Theory of Dyadic Morality nor Moral Foundations Theory provide.

Constructive Sentimentalism further offers a more parsimonious account of moral emotions which fits better with many of the findings in literature, and which is able to explain results seemingly beyond the reach of the Theory of Dyadic Morality - such as why ingesting an anti-emetic appears to reduce the severity of only moderately severe violations of purity (Tracy et al., 2019). Constructive Sentimentalism accounts for the absence of 'characteristic associations' between emotions and moral domains for morally positive events, and its predictions for emotions elicited in morally positive contexts found support despite using an open-response paradigm (see Chapter 7). It also offers several possible explanations for the appearance of associations between moral domains and emotions for immoral actions, whereby 'harm' typically elicits more anger than disgust, and 'impurity' tends to elicit more disgust than anger (e.g., Chapters 4 and 7; also, Franchin et al., 2019), as well as explanations for the frequent co-occurrence of anger and disgust in response to immoral actions. For example, Constructive Sentimentalism suggests moral content may be mixed *at source*, such that immoral acts may be construed as *both* harmful *and* impure to varying extents, whereby the co-elicitation of *both* anger *and* disgust (to varying extents) would be expected.

Another possibility is that anger and disgust may (respectively) relate to the consequences of an action (i.e., harm) and the action itself (i.e., purity), providing co-occurring yet discernible sources of affect (Miller & Cushman, 2013). Alternatively, anger and disgust may be operating at different levels, such that one relates to the act and is elicited via a sentiment, whereas the other relates to the actors' character (e.g., Giner-Sorolla and Chapman, 2017) and is elicited via a meta-sentiment. Likewise, one might claim that 'impurity' relates more to the effect perpetrators have on themselves (i.e., their character), whereas 'harm' relates to the effect had on others (i.e., the act; following Chakroff et al., 2013; Dungan et al., 2017). If meta-sentiments serve to reinforce sentiments, then this may provide some explanation for the frequent co-occurrence of such emotions given the former typically accompanies (and would co-occur with) the latter as a result of moral development. Alternatively, assessments of character might be considered as informing perceptions of one's natural order (e.g., predicting how specific others might behave). These explanations allow for emotional associations with moral content to be maintained, but are not reliant on them.

Additionally, anger and disgust may arise via different motivations, such that imperatives to "confront wrong*doing*" or "avoid wrong*doers*" may be underwritten by anger and disgust respectively. Molho et al.'s (2017) results which show anger and disgust relate to different kinds of aggressive responses, and vary depending on whether the target of the action is oneself (comparatively more anger - confront) or another (comparatively more disgust - avoid), are consistent with this possibility. Furthermore, even if one were to deny 'harm' and 'purity' form coherent moral domains, such that these concern other-blame and self-blame respectively, there remains room for emotions to correspond with moral reference points. If anger and disgust are each associated with a particular reference point, and one needs two points of reference to establish a sentiment as one of disapprobation, then the co-occurrence of these emotions would remain entirely expected, and even necessary. Indeed, Salerno and

Peter-Hegene's (2013) results showing that moral outrage was only predicted by the *co-occurrence* of *both* anger *and* disgust could be considered strongly supportive of this explanation. Importantly, each of these explanations may be advanced without recourse to moral domains, such that Constructive Sentimentalism may offer a better account of moral emotions regardless of any associations with moral content. However, in linking emotions with moral content, Constructive Sentimentalism offers much stronger connections with motivation and behaviour than those available via intuitionist theories.

Yet Constructive Sentimentalism's greatest strength is its definition of what counts as being morally wrong as dependent on having two points of reference. Even if one rejects claims of there being different moral domains and any contributory role for emotions in morality, having the definition formulated in this way provides a standard which the other theories have yet to meet. The navigational analogy, which readily follows from Constructive Sentimentalism's formulation, offers the potential reconciliation of different approaches within moral psychology through identifying common reference points for orientation - it offers a common (theoretical) map. Similarly, although the compass design advanced here is made with reference to an Enactivised Embodied Appraisal Theory (Prinz, 2004b; Shargel & Prinz, 2018), it shares common ground with several approaches to emotion advanced in literature (e.g., Clore & Ortony, 2008; Feldman-Barrett & Bliss-Moreau, 2009; Panksepp & Biven, 2012; Lindquist, 2013) - suggesting the design is not unreasonable. Indeed, the navigational analogy derived from Constructive Sentimentalism (Prinz, 2009) is valuable in its own right. It provides an empirically testable illustration of the connections and common ground between different moral theories, offers a framework within which the main research themes of moral psychology may be organized, and identifies several connecting pathways into a range of related literature. The navigational analogy demonstrates both the explanatory power, and vast connective potential, on offer from Prinzean approaches - to the extent that moral psychologists would seem obliged to pursue these further.

References

Algoe, S. B., & Haidt, J. (2009). Witnessing excellence in action: The 'other-praising' emotions of elevation, gratitude, and admiration. *The journal of positive psychology*, *4*(2), 105-127.

Alicke, M. D. (1992). Culpable causation. *Journal of personality and social psychology*, *63*(3), 368.

Altemeyer, B. (1988). *Enemies of freedom: Understanding right-wing authoritarianism*. Jossey-Bass.

Batson, C. D. (2011). What's wrong with morality? *Emotion Review, 3(3),* 230-236.

Blackford, R. (2010). Book Review: Sam Harris' The Moral Landscape. *Journal of Evolution and Technology*, *21*(2), 53-62.

Brauer, M., & Chekroun, P. (2005). The relationship between perceived violation of social norms and social control: Situational factors influencing the reaction to deviance. *Journal of Applied Social Psychology*, *35*(7), 1519-1539.

Brewer, M. B., & Gardner, W. (1996). Who is this" We"? Levels of collective identity and self representations. *Journal of personality and social psychology*, *71*(1), 83.

Cameron, C. D., Lindquist, K. A., & Gray, K. (2015). A constructionist review of morality and emotions: No evidence for specific links between moral content and discrete emotions. *Personality and Social Psychology Review*, *19*(4), 371-394.

Chakroff, A., Dungan, J., & Young, L. (2013). Harming ourselves and defiling others: What determines a moral domain? *PloS one*, *8*(9), e74434.

Chakroff, A., & Young, L. (2015). Harmful situations, impure people: An attribution asymmetry across moral domains. *Cognition*, *136*, 30-37.

Chen, K., Zhang, H., Friedman, M., & Schlegel, R. (2018). The Authentic Catch-22: The Effect of Following True Self on Decision Satisfaction in Moral Dilemmas.

Cheng, J. S., Ottati, V. C., & Price, E. D. (2013). The arousal model of moral condemnation. *Journal of Experimental Social Psychology*, *49*(6), 1012-1018.

Clifford, S., Iyengar, V., Cabeza, R., & Sinnott-Armstrong, W. (2015). Moral foundations vignettes: A standardized stimulus database of scenarios based on moral foundations theory. *Behavior research methods*, *47*(4), 1178-1198.

Clifford, S., & Jerit, J. (2013). How words do the work of politics: Moral Foundations Theory and the debate over stem cell research. *The Journal of Politics, 75(3),* 659-671.

Clore, G. L., & Ortony, A. (2008). *Appraisal theories: How cognition shapes affect into emotion.* In M. Lewis, J. M. Haviland-Jones, & L. F. Barrett (Eds.), *Handbook of emotions* (p. 628–642). The Guilford Press.

Curry, O. S., Chesters, M. J., & Van Lissa, C. J. (2019). Mapping morality with a compass: Testing the theory of 'morality-as-cooperation' with a new questionnaire. *Journal of Research in Personality*, *78*, 106-124.

Curry, O., Whitehouse, H., & Mullins, D. (2019). Is it good to cooperate? Testing the theory of morality-as-cooperation in 60 societies. *Current Anthropology*, *60*(1).

Davis, D. E., Rice, K., Van Tongeren, D. R., Hook, J. N., DeBlaere, C., Worthington Jr, E. L., & Choe, E. (2016). The moral foundations hypothesis does not replicate well in Black samples. *Journal of Personality and Social Psychology*, *110*(4), e23.

Day, M. V., Fiske, S. T., Downing, E. L., & Trail, T. E. (2014). Shifting liberal and conservative attitudes using moral foundations theory. *Personality and Social Psychology Bulletin*, *40*(12), 1559-1573.

Dennett, D. C. (2014). *Intuition Pumps and Other Tools for Thinking.* Penguin

Dehghani, M., Johnson, K. M., Hoover, J., Sagi, E., Garten, J., Parmar, N. J., Vaisey, S., Iliev, R., & Graham, J. (2016). Purity homophily in social networks. *Journal of Experimental Psychology: General, 145(3*), 366.

DeScioli, P., Gilbert, S. S., & Kurzban, R. (2012). Indelible victims and persistent punishers in moral cognition. *Psychological Inquiry*, *23*(2), 143-149.

Dungan, J. A., Chakroff, A., & Young, L. (2017). The relevance of moral norms in distinct relational contexts: Purity versus harm norms regulate self-directed actions. *PLoS one*, *12*(3), e0173405.

Earp, B. D. (2016). Science cannot determine human values. *Think, 15(43),* 17-23.

Earp, B., & Darby, R. (2017). Circumcision, sexual experience, and harm. *University of Pennsylvania Journal of International Law*, *37*(2).

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in psychology*, *6*, 621.

Ellemers, N., van der Toorn, J., Paunov, Y., & van Leeuwen, T. (2019). The psychology of morality: A review and analysis of empirical studies published from 1940 through 2017. *Personality and Social Psychology Review*, *23*(4), 332-366.

Eriksson, K., Simpson, B., & Strimling, P. (2019). Political double standards in reliance on moral foundations. *Judgment and Decision making*, *14*(4), 440.

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.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175-191.

Feinberg, M., & Willer, R. (2013). The moral roots of environmental attitudes. *Psychological Science, 24*, 56-62.

Feinberg, M., & Willer, R. (2015). From Gulf to Bridge: When do Moral Arguments Facilitate Political Influence?. *Personality and Social Psychology Bulletin, 41(12),* 1665-1681.

Feldman-Barrett, L., & Bliss-Moreau, E. (2009). Affect as a psychological primitive. *Advances in experimental social psychology*, *41*, 167-218.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. Sage.

Franchin, L., Geipel, J., Hadjichristidis, C., & Surian, L. (2019). Many moral buttons or just one? Evidence from emotional facial expressions. *Cognition and Emotion*, *33*(5), 943-958.

Franks, A. S., & Scherr, K. C. (2015). Using moral foundations to predict voting behavior: Regression models from the 2012 US presidential election. *Analyses of Social Issues and Public Policy*, *15*(1), 213-232.

Franks, A. S., & Scherr, K. C. (2019). Economic issues are moral issues: The moral underpinnings of the desire to reduce wealth inequality. *Social Psychological and Personality Science*, *10*(4), 553-562.

Frimer, J. A., Biesanz, J. C., Walker, L. J., & MacKinlay, C. W. (2013). Liberals and conservatives rely on common moral foundations when making moral judgments about influential people. *Journal of personality and social psychology*, *104*(6), 1040.

Ghelfi, E., Christopherson, C. D., Urry, H. L., Lenne, R. L., Legate, N., Ann Fischer, M., ... & de Haan, B. (2020). Reexamining the Effect of Gustatory Disgust on Moral Judgment: A Multilab Direct Replication of Eskine, Kacinik, and Prinz (2011). *Advances in Methods and Practices in Psychological Science*, *3*(1), 3-23.

Giner-Sorolla, R., & Chapman, H. A. (2017). Beyond purity: Moral disgust toward bad character. *Psychological science*, *28*(1), 80-91.

Gordon, A. M., Stellar, J. E., Anderson, C. L., McNeil, G. D., Loew, D., & Keltner, D. (2017). The dark side of the sublime: Distinguishing a threat-based variant of awe. *Journal of Personality and Social Psychology*, *113*(2), 310.

Graham, J. (2015). Explaining away differences in moral judgment: Comment on Gray and Keeney (2015). *Social Psychological and Personality Science*, *6*(8), 869-873.

Graham, J., & Haidt, J. (2012). *Sacred values and evil adversaries: A moral foundations approach.* In M. Mikulincer & P. R. Shaver (Eds.), *Herzliya series on personality and social psychology. The social psychology of morality: Exploring the causes of good and evil* (p. 11–31). American Psychological Association.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology* (Vol. 47, pp. 55-130). Academic Press.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, *96*(5), 1029.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of personality and social psychology*, *101*(2), 366.

Gray, K., & Graham, J. (Eds.). (2019). *Atlas of Moral Psychology*. Guilford Publications.

Gray, K., & Keeney, J. E. (2015a). Impure or just weird? Scenario sampling bias raises questions about the foundation of morality. *Social Psychological and Personality Science*, *6*(8), 859-868.

Gray, K., & Keeney, J. E. (2015b). Disconfirming moral foundations theory on its own terms: Reply to Graham (2015). *Social Psychological and Personality Science*, *6*(8), 874-877.

Gray, K., Schein, C., & Cameron, C. D. (2017). How to think about emotion and morality: Circles, not arrows. *Current opinion in psychology*, *17*, 41-46.

Gray, K., 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.

Gray, K., & Wegner, D. M. (2009). Moral typecasting: divergent perceptions of moral agents and moral patients. *Journal of personality and social psychology*, *96*(3), 505.

Gray, K., & Wegner, D. M. (2011). Dimensions of moral emotions. *Emotion Review*, *3*(3), 258-260.

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological inquiry*, *23*(2), 101-124.

Greene, J. (2003). From neural 'is' to moral 'ought': what are the moral implications of neuroscientific moral psychology? *Nature reviews neuroscience*, *4*(10), 846-850.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review*, *108*(4), 814.

Haidt, J. (2006). *The happiness hypothesis: Finding modern truth in ancient wisdom*. Basic books.

Haidt, J. (2013). *The Righteous Mind: Why Good People are Divided by Religion and Politics.* Penguin.

Haidt, J., Bjorklund, F., & Murphy, S. (2000). Moral dumbfounding: When intuition finds no reason. *Unpublished manuscript, University of Virginia*, 191-221.

Haidt, J., Graham, J., & Ditto, P. H. (2015) A Straw Man can never beat a Shapeshifter. http://www.yourmorals.org/blog/2015/10/a-straw-man-can-never-beat-a-shapeshifter/

Haidt, J., Graham, J., & Joseph, C. (2009). Above and below left–right: Ideological narratives and moral foundations. *Psychological Inquiry*, *20*(2-3), 110-119.

Haidt, J., & Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, *133*(4), 55-66.

Haidt, J., & Joseph, C. (2007). The moral mind: How five sets of innate intuitions guide the development of many culture-specific virtues, and perhaps even modules. *The innate mind*, *3*, 367-391.

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.

Han, H., Liauw, I., & Kuntz, A. F. (2019). Moral identity predicts the development of presence of meaning during emerging adulthood. *Emerging Adulthood*, *7*(3), 230-237.

Harper, C. A., & Hogue, T. E. (2019). The role of intuitive moral foundations in Britain's vote on EU membership. *Journal of Community & Applied Social Psychology*, *29*(2), 90-103.

Harris, S. (2012). *The moral landscape: How science can determine human values*. Black Swan

Heerdink, M. W., Koning, L. F., Van Doorn, E. A., & Van Kleef, G. A. (2019). Emotions as guardians of group norms: Expressions of anger and disgust drive inferences about autonomy and purity violations. *Cognition and emotion*, *33*(3), 563-578.

Heiphetz, L., Strohminger, N., Gelman, S. A., & Young, L. L. (2018). Who am I? The role of moral beliefs in children's and adults' understanding of identity. *Journal of Experimental Social Psychology*, *78*, 210-219.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and brain sciences*, *33*(2-3), 61-83.

Horberg, E. J., Oveis, C., Keltner, D., & Cohen, A. B. (2009). Disgust and the moralization of purity. *Journal of personality and social psychology*, *97*(6), 963.

Hume, D. (1739). *A Treatise of Human Nature.* London: John Noon.

Hume, D. (1751). *An Enquiry Concerning the Principles of Morals.* London: A. Millar.

Iyer, R., Koleva, S., Graham, J., Ditto, P., & Haidt, J. (2012). Understanding libertarian morality: The psychological dispositions of self-identified libertarians. *PloS one*, *7*(8), e42366.

Janoff-Bulman, R. (1979). Characterological versus behavioral self-blame: Inquiries into depression and rape. *Journal of personality and social psychology*, *37*(10), 1798.

Janoff-Bulman, R. (2004). Posttraumatic growth: Three explanatory models. *Psychological inquiry*, *15*(1), 30-34.

Janoff-Bulman, R., & Carnes, N. C. (2013). Surveying the moral landscape: Moral motives and group-based moralities. *Personality and Social Psychology Review*, *17*(3), 219-236.

Janoff-Bulman, R., & Carnes, N. C. (2016). Social justice and social order: Binding moralities across the political spectrum. *PloS one*, *11*(3), e0152479.

Janoff-Bulman, R., & Schwartzberg, S. S. (1991). Toward a general model of personal change. *Handbook of social and clinical psychology: The health perspective*, 488-508.

Johnson, D. J., Cheung, F., & Donnellan, M. B. (2014). Does cleanliness influence moral judgments? *Social Psychology*. *45,* 209-215.

Johnson, D. J., Wortman, J., Cheung, F., Hein, M., Lucas, R. E., Donnellan, M. B., ... & Narr, R. K. (2016). The effects of disgust on moral judgments: Testing moderators. *Social Psychological and Personality Science*, *7*(7), 640-647.

Kaufman, W. R. (2012). Can science determine moral values? A reply to Sam Harris. *Neuroethics*, *5*(1), 55-65.

Kim, K. R., Kang, J. S., & Yun, S. (2012). Moral intuitions and political orientation: Similarities and differences between South Korea and the United States. *Psychological Reports*, *111*(1), 173-185.

Klein, S. B. (2014). What can recent replication failures tell us about the theoretical commitments of psychology? *Theory & Psychology*, *24*(3), 326-338.

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, *63*(3), 190-194.

Kohlberg, L. (1994). *Moral Development: Kohlberg's original study of moral development* (No. 3). Taylor & Francis.

Koleva, S., Graham, J., Haidt, J., Iyer, R., & Ditto, P. H. (2012). Tracing the threads: How five moral concerns (especially Purity) help explain culture war attitudes. *Journal of Research in Personality, 46,* 184-194.

Kollareth, D., Kikutani, M., Shirai, M., & Russell, J. A. (2019). Do community and autonomy moral violations elicit different emotions? *International Journal of Psychology*, *54*(5), 612-620.

Kollareth, D., & Russell, J. A. (2017). On the emotions associated with violations of three moral codes (community, autonomy, divinity). *Motivation and Emotion*, *41*(3), 322-342.

Kollareth, D., & Russell, J. A. (2018). On an Observer's Reaction to Hearing of Someone Harming Him or Herself. *Psychological Studies*, *63*(3), 298-314.

Kollareth, D., & Russell, J. A. (2019). Disgust and the sacred: Do people react to violations of the sacred with the same emotion they react to something putrid? *Emotion*, *19*(1), 37.

Kreutz, G., Ott, U., Teichmann, D., Osawa, P., & Vaitl, D. (2008). Using music to induce emotions: Influences of musical preference and absorption. *Psychology of music*, *36*(1), 101-126.

Kugler, M., Jost, J. T., & Noorbaloochi, S. (2014). Another look at moral foundations theory: Do authoritarianism and social dominance orientation explain liberal conservative differences in “moral” intuitions? *Social Justice Research*, *27*(4), 413-431.

Landmann, H., & Hess, U. (2018). What elicits third-party anger? The effects of moral violation and others’ outcome on anger and compassion. *Cognition and emotion*, *31*(6), 1097-1111.

Landy, J. F., & Bartels, D. M. (2018). An empirically-derived taxonomy of moral concepts. *Journal of Experimental Psychology: General*, *147*(11), 1748.

Landy, J. F., & Goodwin, G. P. (2015a). Does incidental disgust amplify moral judgment? A meta-analytic review of experimental evidence. *Perspectives on Psychological Science*, *10*(4), 518-536.

Landy, J. F., & Goodwin, G. P. (2015b). Our conclusions were tentative, but appropriate: A reply to Schnall et al. (2015). *Perspectives on Psychological Science*, *10*(4), 539-540.

Landy, J. F., & Piazza, J. (2019). Reevaluating moral disgust: sensitivity to many affective states predicts extremity in many evaluative judgments. *Social Psychological and Personality Science*, *10*(2), 211-219.

Lee, S. W., & Ellsworth, P. C. (2013). Maggots and morals: Physical disgust is to fear as moral disgust is to anger. *Components of emotional meaning: A sourcebook*, 271-280.

Lefebvre, J. P., & Krettenauer, T. (2020). Is the true self truly moral? Identity intuitions across domains of sociomoral reasoning and age. *Journal of Experimental Child Psychology*, *192*, 104769.

Lindquist, K. A. (2013). Emotions emerge from more basic psychological ingredients: A modern psychological constructionist model. *Emotion Review*, *5*(4), 356-368.

Liuzza, M. T., Lindholm, T., Hawley, C. B., Gustafsson Sendén, M., Ekström, I., Olsson, M. J., & Olofsson, J. K. (2018). Body odour disgust sensitivity predicts authoritarian attitudes. *Royal Society open science*, *5*(2), 171091.

Lu, J., Peng, X., Liao, C., & Cui, F. (2019). The stereotype of professional roles influences neural responses to moral transgressions: ERP evidence. *Biological psychology*, *145*, 55-61.

Maner, J. K., Kenrick, D. T., Becker, D. V., Robertson, T. E., Hofer, B., Neuberg, S. L., ... & Schaller, M. (2005). Functional projection: How fundamental social motives can bias interpersonal perception. *Journal of personality and social psychology*, *88*(1), 63.

Matsumoto, D., & Ekman, P. (2004). The relationship among expressions, labels, and descriptions of contempt. *Journal of personality and social psychology*, *87*(4), 529.

McAdams, D. P., Albaugh, M., Farber, E., Daniels, J., Logan, R. L., & Olson, B. (2008). Family metaphors and moral intuitions: How conservatives and liberals narrate their lives. *Journal of personality and social psychology*, *95*(4), 978.

McAuliffe, W. H. (2019). Do emotions play an essential role in moral judgments? *Thinking & Reasoning*, *25*(2), 207-230.

McHugh, C., McGann, M., Igou, E. R., & Kinsella, E. L. (2017). Searching for moral dumbfounding: Identifying measurable indicators of moral dumbfounding. *Collabra: Psychology*, *3*(1).

Mehling, W. E., Price, C., Daubenmier, J. J., Acree, M., Bartmess, E., & Stewart, A. (2012). The multidimensional assessment of interoceptive awareness (MAIA). *PloS one*, *7*(11), e48230.

Miller, R., & Cushman, F. (2013). Aversive for me, wrong for you: First-person behavioral aversions underlie the moral condemnation of harm. *Social and Personality Psychology Compass*, *7*(10), 707-718.

Miller, L. C., Murphy, R., & Buss, A. H. (1981). Consciousness of body: Private and public. *Journal of personality and social psychology*, *41*(2), 397.

Molho, C., Tybur, J. M., Güler, E., Balliet, D., & Hofmann, W. (2017). Disgust and anger relate to different aggressive responses to moral violations. *Psychological science*, *28*(5), 609-619.

Moore, C., & Gino, F. (2013). Ethically adrift: How others pull our moral compass from true North, and how we can fix it. *Research in organizational behavior*, *33*, 53-77.

Nagel, T. (2010). The facts fetish. *New Republic, 241(18),* 30-33.

Nichols, S. (2002). Norms with feeling: Towards a psychological account of moral judgment. *Cognition*, *84*(2), 221-236.

Nichols, S. (2004). *Sentimental rules: On the natural foundations of moral judgment*. Oxford University Press.

Noel, Y., Fevrier, F., & Deflandre, A. (2018). Two factors but one dimension: An alternative view at the structure of mood and emotion. https://psyarxiv.com/tv9ys/

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615-631.

Oaten, M., Stevenson, R. J., Williams, M. A., Rich, A. N., Butko, M., & Case, T. I. (2018). Moral violations and the experience of disgust and anger. *Frontiers in Behavioral Neuroscience*, *12*, 179.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251).

Panksepp, J. & Biven, L. (2012). *The Archaeology of Mind: Neuroevolutionary Origins of Human Emotions.* W. W. Norton.

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528-530.

Piazza, J., Landy, J. F., Chakroff, A., Young, L., & Wasserman, E. (2018). What disgust does and does not do for moral cognition. *The moral psychology of disgust*, 53-81.

Popper, K. (2014). *The myth of the framework: In defence of science and rationality*. Routledge.

Prentice, M., Jayawickreme, E., Hawkins, A., Hartley, A., Furr, R. M., & Fleeson, W. (2019). Morality as a basic psychological need. *Social Psychological and Personality Science*, *10*(4), 449-460.

Prinz, J. J. (2004a). *Furnishing the mind: Concepts and their perceptual basis*. MIT press.

Prinz, J. J. (2004b). *Gut reactions: A perceptual theory of emotion.* Oxford University Press.

Prinz, J. J. (2006a). Is the mind really modular? *Contemporary debates in cognitive science, 7*.

Prinz, J. J. (2006b). The emotional basis of moral judgments. *Philosophical explorations*, *9*(1), 29-43.

Prinz, J. J. (2007a). Can moral obligations be empirically discovered? *Midwest Studies in Philosophy*, *31*, 271-291.

Prinz, J. J. (2007b). Emotion and Aesthetic Value. http://subcortex.com/EmotionAndAestheticValuePrinz.pdf

Prinz, J. J. (2008). Is morality innate. *Moral psychology, 1,* 367-406.

Prinz, J. J. (2009). *The emotional construction of morals*. Oxford University Press.

Prinz, J. J. (2011). Is empathy necessary for morality. *Empathy: Philosophical and psychological perspectives*, *1*, 211-229.

Prinz, J. (2012). *The conscious brain*. Oxford University Press.

Prinz, J. J. (2013). Constructive sentimentalism: Legal and political implications. *Nomos*, *53*, 3-18.

Prinz, J. J., & Nichols, S. B. (2016). Diachronic identity and the moral self. In *The Routledge handbook of philosophy of the social mind* (pp. 449-464). Taylor and Francis.

Prinz, J., & Seidel, A. (2012). Alligator or squirrel: Musically induced fear reveals threat in ambiguous figures. *Perception*, *41*(12), 1535-1539.

Pronin, E., Wegner, D. M., McCarthy, K., & Rodriguez, S. (2006). Everyday magical powers: The role of apparent mental causation in the overestimation of personal influence. *Journal of personality and social psychology*, *91*(2), 218.

Rai, T. S., & Fiske, A. P. (2011). Moral psychology is relationship regulation: moral motives for unity, hierarchy, equality, and proportionality. *Psychological review*, *118*(1), 57.

Rempala, D. M., Okdie, B. M., & Garvey, K. J. (2016). Articulating ideology: How liberals and conservatives justify political affiliations using morality-based explanations. *Motivation and Emotion*, *40*(5), 703-719.

Rottman, J., Kelemen, D., & Young, L. (2014). Tainting the Soul: Purity concerns predict moral judgments of suicide. *Cognition, 130(2),* 217-226.

Rottman, J., Young, L., & Kelemen, D. (2017). The impact of testimony on children's moralization of novel actions. *Emotion*, *17*(5), 811.

Royzman, E. B., Leeman, R. F., & Baron, J. (2009). Unsentimental ethics: Towards a content-specific account of the moral–conventional distinction. *Cognition*, *112*(1), 159-174.

Royzman, E., Atanasov, P., Landy, J. F., 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.

Royzman, E. B., Kim, K., & Leeman, R. F. (2015). The curious tale of Julie and Mark: unravelling the moral dumbfounding effect. *Judgment & Decision Making*, *10*(4).

Royzman, E., Cusimano, C., & Leeman, R. F. (2017). What lies beneath? Fear vs. disgust as affective predictors of absolutist opposition to genetically modified food and other new technologies. *Judgment and Decision Making*, *12*(5), 466.

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.

Russell, P. S., & Giner-Sorolla, R. (2011a). Moral anger is more flexible than moral disgust. *Social Psychological and Personality Science, 2(4),* 360-364.

Russell, P. S., & Giner-Sorolla, R. (2011b). Moral anger, but not moral disgust, responds to intentionality. *Emotion, 11(2),* 233.

Salerno, J. M., & Peter-Hagene, L. C. (2013). The interactive effect of anger and disgust on moral outrage and judgments. *Psychological Science*, *24*(10), 2069-2078.

Santee, R. T., & Jackson, J. (1977). Cultural values as a source of normative sanctions. *Pacific Sociological Review*, *20*(3), 439-454.

Schein, C., & Gray, K. (2014). The prototype model of blame: Freeing moral cognition from linearity and little boxes. *Psychological Inquiry*, *25*(2), 236-240.

Schein, C., & Gray, K. (2015). The unifying moral dyad: Liberals and conservatives share the same harm-based moral template. *Personality and Social Psychology Bulletin*, *41*(8), 1147-1163.

Schein, C., & Gray, K. (2016). Moralization and harmification: The dyadic loop explains how the innocuous becomes harmful and wrong. *Psychological Inquiry*, *27*(1), 62-65.

Schein, C., & Gray, K. (2018). The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review*, *22*(1), 32-70.

Schein, C., Ritter, R. S., & Gray, K. (2016). Harm mediates the disgust-immorality link. *Emotion*, *16*(6), 862.

Schnall, S. (2017). Disgust as embodied loss aversion. *European Review of Social Psychology*, *28*(1), 50-94.

Schnall, S., Benton, J., & Harvey, S. (2008). With a clean conscience: Cleanliness reduces the severity of moral judgments. *Psychological science*, *19*(12), 1219-1222.

Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H. (2008). Disgust as embodied moral judgment. *Personality and social psychology bulletin*, *34*(8), 1096-1109.

Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H. (2015). Landy and Goodwin confirmed most of our findings then drew the wrong conclusions.

Scott, S. E., Inbar, Y., & Rozin, P. (2016). Evidence for absolute moral opposition to genetically modified food in the United States. *Perspectives on Psychological Science*, *11*(3), 315-324.

Seidel, A., & Prinz, J. (2013a). Sound morality: Irritating and icky noises amplify judgments in divergent moral domains. *Cognition*, *127*(1), 1-5.

Seidel, A., & Prinz, J. (2013b). Mad and glad: Musically induced emotions have divergent impact on morals. *Motivation and Emotion*, *37*(3), 629-637.

Shargel, D., & Prinz, J. J. (2018). An enactivist theory of emotional content. *The ontology of emotion*, 110-129.

Sheikh, S., & Janoff-Bulman, R. (2010). The “shoulds” and “should nots” of moral emotions: A self-regulatory perspective on shame and guilt. *Personality and Social Psychology Bulletin*, *36*(2), 213-224.

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. *Morality and health*, *119*, 119-169.

Skitka, L. J., Wisneski, D. C., & Brandt, M. J. (2018). Attitude moralization: Probably not intuitive or rooted in perceptions of harm. *Current Directions in Psychological Science*, *27*(1), 9-13.

Stolerman, D., & Lagnado, D. (2020). The Moral Foundations of Human Rights Attitudes. *Political Psychology*, *41*(3), 439-459.

Strohminger, N., & Nichols, S. (2014). The essential moral self. *Cognition, 131(1),* 159-171.

Suhler, C. L., & Churchland, P. (2011). Can innate, modular “foundations” explain morality? Challenges for Haidt's moral foundations theory. *Journal of cognitive neuroscience*, *23*(9), 2103-2116.

Tilghman-Osborne, C., Cole, D. A., Felton, J. W., & Ciesla, J. A. (2008). Relation of guilt, shame, behavioral and characterological self-blame to depressive symptoms in adolescents over time. *Journal of Social and Clinical Psychology*, *27*(8), 809 842.

Tracy, J. L., Steckler, C. M., & Heltzel, G. (2019). The physiological basis of psychological disgust and moral judgments. *Journal of Personality and Social Psychology*, *116*(1), 15.

Trafimow, D., & Earp, B. D. (2016). Badly specified theories are not responsible for the replication crisis in social psychology: Comment on Klein. *Theory & Psychology*, *26*(4), 540-548.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly review of biology*, *46*(1), 35-57.

Tybur, J. M., Molho, C., Cruz, T. D. D., Cakmak, B., Singh, G. D., & Zwicker, M. (2019). Tybur Et Al. Disgust Anger Aggression. https://psyarxiv.com/wuspt/

Ugazio, G., Lamm, C., & Singer, T. (2012). The role of emotions for moral judgments depends on the type of emotion and moral scenario. *Emotion*, *12*(3), 579.

Uhlmann, E. L., & Zhu, L. (2013). Acts, persons, and intuitions: Person-centered cues and gut reactions to harmless transgressions. *Social Psychological and Personality Science, 5*, 279-285.

van Deurzen, E. (2014). Structural Existential Analysis (SEA): A phenomenological research method for counselling psychology. *Counselling Psychology Review*, *29*(2), 70-83.

van Leeuwen, F., Dukes, A., Tybur, J. M., & Park, J. H. (2017). Disgust sensitivity relates to moral foundations independent of political ideology. *Evolutionary Behavioral Sciences*, *11*(1), 92.

van 't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—A discussion and suggested template. *Journal of Experimental Social Psychology, 67,* 2–12

Wagemans, F., Brandt, M. J., & Zeelenberg, M. (2018). Disgust sensitivity is primarily associated with purity-based moral judgments. *Emotion*, *18*(2), 277.

Wasserman, E. A., Chakroff, A., Saxe, R., & Young, L. (2017). Illuminating the conceptual structure of the space of moral violations with searchlight representational similarity analysis. *NeuroImage*, *159*, 371-387.

Weidman, A. C., Sowden, W. J., Berg, M. K., & Kross, E. (2020). Punish or protect? How close relationships shape responses to moral violations. *Personality and Social Psychology Bulletin*, *46*(5), 693-708.

Wheatley, T., & Haidt, J. (2005). Hypnotic disgust makes moral judgments more severe. *Psychological science*, *16*(10), 780-784.

Wisneski, D. C., & Skitka, L. J. (2017). Moralization through moral shock: Exploring emotional antecedents to moral conviction. *Personality and Social Psychology Bulletin*, *43*(2), 139-150.

Young, L. & Saxe, R. (2011). When ignorance is no excuse: Different roles for intent across moral domains. *Cognition, 120,* 202-214.

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.